İZMİR
KÂTİP ÇELEBİ
ÜNİVERSİTESİ
2010
GRADUATE SCHOOL OF NATURAL
AND APPLIED SCIENCES

Automated Captioning of Image and Audio for Visually and Hearing Impaired

Submitted for the Degree of Master of Science from the İzmir Kâtip Çelebi University in Electrical Electronics Engineering

by

Özkan Çaylı

ORCID 0000-0002-3389-3867

Advisor: Assoc. Prof. Dr. Volkan Kılıç

January, 2024

This is to certify that we have read the thesis **Automated Captioning of Image and Audio for Visually and Hearing Impaired** submitted by **Özkan Çaylı**, and it has been judged to be successful, in scope and in quality, at the defense exam and accepted by our jury as a MASTER'S THESIS.

APPROVED BY:

Advisor:	**Assoc. Prof. Dr. Volkan Kılıç** İzmir Kâtip Çelebi University
Co-advisor:	**Assoc. Prof. Dr. Aytuğ Onan** İzmir Kâtip Çelebi University

Committee Members:

Assoc. Prof. Dr. Volkan Kılıç
İzmir Kâtip Çelebi University

Assoc. Prof. Dr. Aytuğ Onan
İzmir Kâtip Çelebi University

Assist. Prof. Dr. Emin Borandağ
Manisa Celal Bayar University

Date of Defense: January 18, 2024

Declaration of Authorship

I, **Özkan Çaylı**, declare that this thesis titled **Automated Captioning of Image and Audio for Visually and Hearing Impaired** and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for the Master's / Doctoral degree at this university.

- Where any part of this thesis has previously been submitted for a degree or any other qualification at this university or any other institution, this has been clearly stated.

- Where I have consulted the published work of others, this is always clearly attributed.

- Where I have quoted from the work of others, the source is always given. This thesis is entirely my own work, with the exception of such quotations.

- I have acknowledged all major sources of assistance.

- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: 18.01.2024

Automated Captioning of Image and Audio for Visually and Hearing Impaired

Abstract

Generating captions and text descriptions of images will enable visually and hearing impaired extended accessibility to the real-world, thus reducing their social isolation, and improving their well-being, employability, and education experience. This thesis presents significant advancements in algorithmic approaches for generating captions and text descriptions. These enhancements are pivotal in processing and interpreting both image and audio data. The focus on algorithmic innovation ensures that the platform is not only efficient but also adaptable to various types of visual and auditory information, making it a versatile tool for aiding those with visual impairments. The thesis has addressed this aim in three main contribution chapters, image captioning, video captioning, and audio-visual video captioning approaches. The progression of this research is methodically structured, starting with image captioning. This initial phase concentrates on developing sophisticated algorithms capable of accurately interpreting and describing still images. This foundational work sets the stage for the subsequent phase, video captioning. Here, the complexity increases as the algorithms are adapted to handle dynamic visual content, providing contextual and temporal descriptions of video sequences. The culmination of this research is in the integration of audio-visual video captioning. This final phase synergizes the advances from the previous stages, incorporating audio analysis to enhance the depth and accuracy of captions. This comprehensive approach ensures a robust and inclusive system, capable of providing detailed descriptions for a wide range of visual and auditory inputs, thus offering a more complete understanding of the environment for users with visual and hearing impairments.

Keywords: Image Captioning, Video Captioning, Audio-Visual Video Captioning, Computer Vision, Natural Language Processing, Audio Processing.

Görme ve İşitme Engelliler için Otomatik Görüntü ve Ses Altyazılama

Öz

Görüntülerin ve ses verilerinin işlenmesi ve yorumlanmasında önemli ilerlemeler sunan bu tez, görme ve işitme engelli bireylerin gerçek dünyaya olan erişimlerini genişleterek sosyal izolasyonlarını azaltacak, refahlarını, istihdam olanaklarını ve eğitim deneyimlerini iyileştirecek görüntü ve ses betimlemeleri üretme üzerine algoritmik yaklaşımlarda önemli gelişmeler sunmaktadır. Algoritmik yeniliklere odaklanmak, platformun sadece verimli değil, aynı zamanda çeşitli görsel ve işitsel bilgi türlerine uyum sağlayabilecek şekilde esnek olmasını sağlar. Bu, görme engellilere yardım etmek için çok yönlü bir araç haline gelir. Tez, üç ana katkı bölümünde bu amacı ele almıştır: görüntü altyazılama, video altyazılama ve sesli-görselli video altyazılama yaklaşımları. Bu araştırmanın ilerleyişi ilk olarak görüntü altyazılama ile başlar. Bu ilk aşama, durağan resimleri doğru bir şekilde yorumlayıp tanımlayabilen sofistike algoritmaların geliştirilmesine odaklanır. Bu temel çalışma, ardından gelen video altyazılama aşaması için zemin hazırlar. Burada, algoritmalar dinamik görsel içeriği ele alacak şekilde uyarlanır, video dizilerinin bağlamsal ve zamansal betimlemelerini sağlar. Bu araştırmanın son noktası, sesli-görselli video altyazılama entegrasyonudur. Bu son aşama, önceki aşamalardan elde edilen ilerlemeleri senkronize eder, altyazıların derinliğini ve doğruluğunu artırmak için ses analizini dahil eder. Bu kapsamlı yaklaşım, geniş bir görsel ve işitsel girdi yelpazesi için detaylı açıklamalar sağlayabilen sağlam ve kapsayıcı bir sistem sağlar, böylece görme ve işitme engelli kullanıcılara çevrelerini daha iyi bir şekilde anlama imkanı sunar.

Anahtar Kelimeler: Görüntü Altyazılama, Video Altyazılama, İşitsel-Görsel Video Altyazılama, Bilgisayar Görüsü, Doğal Dil İşleme, Ses İşleme.

To my supervisors, colleagues, and family members

Acknowledgment

First and foremost, I would like to extend my deepest gratitude to my supervisor, Assoc. Prof. Dr. Volkan Kılıç, for his invaluable guidance and expertise throughout my Master's studies at Izmir Katip Celebi University. His support and insightful contributions have been crucial in my academic development and in achieving my goals.

I am also grateful to my co-supervisor, Assoc. Prof. Dr. Aytuğ Onan, for his support and guidance throughout this study. His encouragement and advice have significantly enriched my educational experience.

I would like to express my sincere gratitude to my research project collaborators at the University of Surrey, Xubo Liu, Xinhao Mei, Dr. Jianyuan Sun, and Prof. Dr. Wenwu Wang, for their invaluable assistance and insightful contributions to my research. Their expertise and collaboration have been pivotal to the success of my thesis.

My thanks also go to my colleagues and friends in the IKCU AI Lab at Izmir Katip Celebi University for their valuable comments and advice, which have greatly aided me on this challenging path. Additionally, I am thankful to all my friends and professors who supported me throughout my graduate education.

I must also extend a special thank you to my family for their support and love. Their encouragement and belief in me have been the backbone of my journey. Their selfless support and unconditional love have given me the strength and courage to pursue my academic aspirations and overcome the challenges along the way.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
BLEU-n	Bilingual Evaluation Understudy
CIDEr	Consensus-based Image Description Evaluation
CNNs	Convolutional Neural Networks
GRUs	Gated Recurrent Units
LSTMs	Long Short-Term Memory networks
METEOR	Metric for Evaluation of Translation with Explicit Ordering
MSR-VTT	Microsoft Research Video-to-Text
MSVD	Microsoft Video Description Dataset
MSCOCO	Microsoft Common Objects in COntext
MVIT	Multi-Scale Vision Transformers
NAS	Neural Architecture Search
NASNet-Large	Neural Architecture Search Network (Large variant)
PANNs	Pretrained Audio Neural Networks
P3D	Pseudo-3D Convolutional Network
R3D	3D ResNet
ResNet	Residual Network
RNNs	Recurrent Neural Networks
ROUGE-L	Recall-Oriented Understudy for Gisting Evaluation
S3D	Separated 3D Convolutional Network
3D-CNNs	Three-Dimensional Convolutional Neural Networks

List of Symbols

r_t	Reset gate vector
z_t	Update gate vector
u_t	Candidate hidden vector
W_r	Weights of the reset gate
W_z	Weights of the update gate
σ	The sigmoid hyperbolic activation function
$\tanh$	The tangent hyperbolic activation function
Θ	The element-wise multiplication operator
h_{t-1}	The previous hidden state of GRU
y_t	The output of GRU
x_t	The input of GRU
b	The bias of the GRU
$k = 1, \dots, K$	The layer index
$h_t^{(k)}$	The hidden vector for the kth GRU layer
$x_t^{(k)}$	The input vector for the kth GRU layer
d_t	The output of the dense layer
T	The number of word predictions for caption generation
N^a	The number of audio frames
N^v	The number of video frames
N^c	The number of words in the caption
$N_a - 1$	The last index of audio frames
$N_v - 1$	The last index of video frames
$\boldsymbol{F}^a = \left(F_0^a, F_1^a, \dots F_{N_a-1}^a\right)$	Audio frames
$\boldsymbol{F}^v = \left(F_0^v, F_1^v, \dots F_{N_v-1}^v\right)$	Video frames
$\boldsymbol{V}$	Video
n	The neural network
L_{CE}	Cross-entropy loss
L_{rep}	L1 loss
S	The dimension of the spectral feature
q	Compression rate
$\boldsymbol{F}^a$	Compressed audio
$m = 0, 1, 2, \dots, N_v - 1$	Video frame index

Chapter 1

Introduction

This chapter introduces the foundation of a comprehensive study in the field of artificial intelligence, specifically focusing on automated natural language descriptions of images, audio, and video. The chapter outlines the motivation, objectives, and significant contributions of the study, providing a clear trajectory of the research.

1.1. Motivation

Automatically generating a natural language description of an image, audio, and video has recently received increasing attention from computer vision, machine listening, and natural language processing which are major fields in artificial intelligence (AI). Especially in image captioning, numerous datasets such as MSCOCO, ImageNet, and Flickr, were released to test the performance of the proposed methods in various conditions under the performance metrics BLEU-n, ROUGE-L, METEOR, and CIDEr. The benchmarks with these performance metrics on the datasets show that the image captioning area is open to more investigations and new advanced methodologies which are targeted in the first objective of this study. The second objective of this study shifts attention to video captioning. Video content offers a richer and more complex set of data for analysis compared to static images. This complexity arises from the temporal dynamics and the motion between successive frames, presenting unique challenges for automatic caption generation. In response, this thesis explores video captioning using advanced neural network architectures and feature extraction techniques. The aim is to develop methods that not only capture the essence of visual content over time but also translate it into coherent and contextually relevant natural language descriptions. Datasets such as MSR-VTT and MSVD are utilized for benchmarking, employing similar performance metrics as in image captioning, but with additional considerations for the temporal aspects of video data. The third and final objective covers audio-visual captioning. This approach recognizes the synergy between audio and visual elements in multimedia content. By integrating auditory information with visual data, the research aims to create a more comprehensive captioning system that leverages the strengths of both modalities. This multimodal approach is especially relevant in scenarios where audio provides contextual clues that are not visually evident, thereby enhancing the overall accuracy and richness of the generated captions. In pursuing this objective, the thesis explores various methods of audio-visual data fusion and feature representation, striving to set a new benchmark in multimodal captioning. The complementary nature of image and audio will be integrated under a new framework, allowing natural language processing to use image and audio processing results for more robust caption generation. This will be, hopefully, a milestone in automated caption generation, leading to more interest by researchers in the community in multimodal captioning.

1.2. Contributions

In this thesis, we have explored advanced methodologies in image and video captioning, making significant contributions across various facets of this rapidly evolving field. Our work primarily focused on enhancing the accuracy and computational efficiency of captioning systems using novel neural network architectures and efficient processing techniques.

Multi-layer GRU for Image Captioning: We introduced an innovative image captioning approach using the NASNet-Large encoder and a multi-layer GRU based decoder. This approach significantly improved the ability to modulate relevant information flow, addressing long-term complex dependencies in RNN decoders. Our method demonstrated enhanced performance in generating semantically consistent captions, as validated on the MSCOCO dataset.

Leveraging Pre-trained 3D-CNNs for Video Captioning: We developed a video captioning method integrating 2D and 3D-CNN architectures with a multi-layer GRU. This novel integration effectively enhanced the accuracy of caption generation from video data, as evidenced by our evaluations on the MSVD dataset.

Efficient Audio-Visual Video Captioning via Knowledge Distillation: Addressing the challenge of deploying captioning systems on low-power devices, we proposed a method that utilized simple pooling front-ends, down-sampling algorithms, and knowledge distillation for efficient audio-visual processing. This approach significantly reduced inference time with minimal accuracy loss, offering a practical solution for resource-constrained environments.

These contributions represent significant advancements in the fields of image and video captioning. We have not only addressed key challenges in these areas but also laid the groundwork for future research, particularly in the integration of transformer models, neural architecture search, and the development of real-world applications. Our research holds the potential to greatly enhance the accessibility and effectiveness of captioning systems across various applications.

1.3. Outline of the Thesis

This thesis presents an in-depth exploration of advanced techniques in image and video captioning, leveraging neural networks and efficient processing methods to enhance captioning performance. The thesis is structured as follows:

Introduction: Provides an overview of the thesis, outlining the significance of image and video captioning in the context of artificial intelligence and machine learning. It lays the foundation for the subsequent chapters by discussing the motivation, objectives, and scope of the research.

Methods and Datasets: Describes the fundamental tools and methodologies used in the thesis, including deep learning frameworks and datasets like MSCOCO and MSR-VTT. It covers various neural network layers, attention mechanisms, and feature extraction techniques relevant to captioning tasks.

Multi-layer Gated Recurrent Unit based Recurrent Neural Network for Image Captioning: Introduces a novel image captioning approach using NASNet-Large and multi-layer GRU, emphasizing its effectiveness in generating contextually accurate captions, validated on the MSCOCO dataset.

Leveraging Pre-trained 3D-CNNs for Video Captioning: Discusses the development of a video captioning method that integrates 2D and 3D-CNN architectures with a multi-layer GRU, demonstrating improved captioning accuracy on the MSVD dataset.

Knowledge Distillation for Efficient Audio-Visual Video Captioning: Presents an approach combining simple pooling, down-sampling, and knowledge distillation for efficient audio-visual video captioning, highlighting its reduced inference time and maintained accuracy on the MSR-VTT dataset.

Conclusions and Future Research: Summarizes the key findings and contributions of the thesis, reflecting on the advancements made in image and video captioning. It also outlines potential future research directions, including the exploration of transformer models, neural architecture search, and real-world applications.

This thesis structure comprehensively covers the research from foundational methods to innovative applications in image and video captioning, concluding with a synthesis of findings and future prospects.

Chapter 2

Methods and Datasets

Chapter 2 serves as the technical foundation of this thesis, outlining essential deep learning tools and datasets. It explores various neural network components, feature extraction techniques, and datasets used for benchmarking, along with the programming languages and frameworks employed, thereby setting a comprehensive foundation for the methodologies used in this research.

2.1. Deep Learning Tools

2.1.1. Linear Layer

The Linear Layer [1], often a fundamental component in neural networks, plays a critical role in text processing. It functions by applying a linear transformation to the input data, essentially mapping the input features to a higher or lower-dimensional space. In the context of NLP, Linear Layers are used to transform word embeddings or feature vectors into a format suitable for further processing or classification tasks. This layer is crucial for creating models that can understand and classify textual information accurately.

2.1.2. Embedding Layer

The Embedding Layer [2] is vital in text processing, especially in handling large vocabularies. It converts categorical data, typically words or phrases, into dense vectors of fixed-size. This dense representation is more efficient and meaningful than traditional one-hot encoded vectors. Embedding layers are extensively used in NLP models to capture semantic information about words, allowing the model to understand word similarities and relationships based on their usage in the training data. This layer is essential in tasks like word similarity, text classification, and sequence modeling, forming the basis for more advanced language models.

2.1.3. Residual Connections

Residual Connections [3] are a prominent feature in modern neural network architectures, particularly in deep convolutional networks. Introduced in the context of ResNet (Residual Network), these connections address the vanishing gradient issue common in deep networks by allowing shortcuts for the gradients to flow through. A residual connection skips one or more layers and adds the output from a previous layer to the output of a stacked layer. This technique effectively allows the model to learn an identity function, ensuring that the deeper layers can at least perform as well as the shallower ones. By facilitating the training of much deeper networks, residual connections lead to significant improvements in tasks like image classification and object detection, contributing to the overall robustness and performance of the model.

2.1.4. Dropouts

Dropouts [4] are a regularization technique used in neural networks to prevent overfitting. This technique involves randomly 'dropping out' or deactivating a subset of neurons during training. By doing so, dropout forces the network to not rely excessively on any single neuron, thereby promoting redundancy and robustness within the network architecture. During training, different subsets of neurons are dropped out randomly, ensuring that the network learns more generalized features. At test time, all neurons are used, but their outputs are scaled down to account for the reduced number of active neurons during training. Dropout is widely used in various neural network architectures, including fully connected layers and convolutional layers, enhancing the generalization ability of the models in tasks like image classification, natural language processing, and more.

2.1.5. Bahdanau Attention Mechanism

The Bahdanau Attention Mechanism [5], introduced by Dzmitry Bahdanau and his colleagues, represents a significant advancement in the field of neural network-based sequence modeling, particularly in tasks involving natural language processing. This mechanism addresses a critical limitation in traditional sequence-to-sequence models: the inability to focus on specific parts of the input sequence when generating each word in the output sequence. The Bahdanau Attention essentially allows the model to learn to assign varying degrees of importance, or 'attention', to different parts of the input sequence, creating a context vector for each output step. This context vector is then used in conjunction with the decoder's state to generate the output sequence, thus enabling the model to produce more accurate and contextually relevant outputs. This attention mechanism has been instrumental in improving the performance of various applications, including machine translation, speech recognition, and text summarization, by providing a more dynamic and adaptive approach to handling sequential data.

2.1.6. Convolutional Neural Networks (CNNs)

CNNs [6] are a class of deep neural networks highly effective in processing data with a grid-like topology, such as images. A CNN typically consists of various layers that automatically and adaptively learn spatial hierarchies of features from input images. These layers include convolutional layers, pooling layers, fully connected layers, and normalization layers.

Convolutional Layers: The cornerstone of a CNN, these layers perform a convolution operation, applying filters to the input to create feature maps. This helps the network learn image features such as edges, textures, and complex patterns.

Pooling Layers: Following convolutional layers, pooling layers (such as max pooling or average pooling) reduce the spatial dimensions (width and height) of the input volume. This operation is crucial for reducing the number of parameters and computational complexity, and also helps in achieving translational invariance in the network.

Fully Connected Layers: Towards the end of a CNN, fully connected layers integrate learned features from previous layers to determine the class of the input image. These layers are similar to those in a traditional neural network and are used for high-level reasoning in the network.

Normalization Layers: Layers like Batch Normalization or Local Response Normalization are used within CNNs to stabilize and accelerate the training process.

CNNs have revolutionized the field of computer vision, achieving remarkable success in tasks such as image classification, object detection, semantic segmentation, and more. Their ability to extract and learn feature representations makes them a powerful tool in many applications beyond vision, including audio processing and natural language processing.

2.2. Visual Feature Extraction

2.2.1. Inception-v3

Inception-v3 [7] is a deep CNN architecture renowned for its efficiency in visual feature extraction, particularly in image classification tasks. This 48-layered network, pre-trained on the ImageNet dataset, stands out for its use of an asymmetric approach, which breaks down large-scale convolution kernels into smaller, more manageable ones. The Inception-v3 architecture is characterized by its combination of convolution, pooling, and fully connected (FC) layers, which work together to process input images. Specifically, it accepts input images of size 3×299×299, efficiently resizing and handling them through its complex layer structure. A notable aspect of Inception-v3 is its global average pooling layer, which plays a critical role in the functionality of the network. The output from this layer is a feature vector of 2048 units, which effectively captures the essential characteristics of the input image. This feature vector then serves as a latent vector, crucial for various applications like feature injection in decoders or further image processing tasks. The ability of the architecture to extract high-level visual features efficiently makes it an invaluable tool in the field of computer vision, particularly for tasks involving feature extraction from visual frames. Overall, the combination of depth, efficiency, and versatility in Inception-v3 makes it a standout choice in the field of deep learning for image analysis.

2.2.2. ResNet152 v2

ResNet152 v2 [8], an enhancement of the original ResNet152, is a prominent model in deep learning, specifically in computer vision. It belongs to the Residual Network

(ResNet) family, known for enabling the training of extremely deep neural networks. The 'v2' in ResNet152 v2 indicates refinements over its predecessor, primarily in how batch normalization and activation functions are applied, leading to improved performance. This model contains 152 layers, facilitating a deeper and more complex neural network architecture. Its primary innovation lies in the use of residual connections, which address the vanishing gradient problem commonly encountered in deep networks. These connections allow the network to learn identity functions effectively, ensuring that adding more layers doesn't lead to performance degradation. ResNet152 v2 has shown significant success in image classification tasks and is widely used in applications requiring detailed feature extraction from images.

2.2.3. Xception

Xception [9], short for “Extreme Inception,” is an advanced deep learning model that rethinks the Inception architecture. It introduces the concept of depthwise separable convolutions, which makes it unique and powerful. The model separates the learning of spatial features (through depthwise convolutions) and the learning of cross-channel correlations (through pointwise convolutions), leading to a more efficient learning process. This architecture consists of 36 convolutional layers, forming the Xception’s base. The layers are structured into 14 modules, all of which have linear residual connections around them, except for the first and last modules. Xception's efficiency and effectiveness lie in its ability to handle a large number of parameters more efficiently than traditional convolutional networks, making it particularly suitable for tasks with high computational demands, such as large-scale image recognition.

2.2.4. NASNet-Large

NASNet-Large stands as a testament to the advancements in neural architecture search (NAS). It is a product of automated machine learning, where a controller neural network generates architectures to be evaluated. NASNet-Large is a convolutional neural network architecture that is specifically optimized for high performance image recognition tasks. This model is characterized by its scalability – it can be efficiently scaled up for greater accuracy. The ‘Large’ variant signifies its configuration for larger-scale and more complex tasks, featuring a higher number of layers and parameters than its smaller counterparts. The architecture of NASNet-Large is notable for its repeated cell structures, which are optimized for both convolutional and reduction operations. These cells are identified through a search process on a smaller dataset and then scaled up, allowing NASNet-Large to achieve exceptional accuracy in image classification challenges.

2.2.5. S3D (Separated 3D Convolutional Network)

S3D, or Separated 3D Convolutional Network, is an innovative approach in video processing that aims to enhance the efficiency and effectiveness of 3D convolutional networks. The core idea behind S3D is the separation of spatial and temporal components within the 3D convolution process. Unlike traditional 3D CNNs that combine spatial and temporal features in a single convolution step, S3D performs spatial convolutions and temporal convolutions separately. This separation allows S3D to capture complex spatial details through dedicated spatial convolutions, while temporal convolutions focus on the dynamics and movements across video frames. The result is a more computationally efficient model that retains the depthwise understanding of videos. S3D has shown promising results in tasks like action recognition and video classification, offering a balance between computational load and the ability to capture intricate features of video data.

2.2.6. R3D (3D ResNet)

R3D, or 3D ResNet, extends the principles of the well-known ResNet architecture into the realm of video processing. It employs 3D convolutional layers to capture both spatial and temporal information present in video sequences. The R3D model integrates the concept of residual connections, which are crucial in enabling the training of deep networks by allowing the flow of gradients through the network without significant loss. These residual connections also help in alleviating the vanishing gradient problem often encountered in deep neural networks. By adapting the ResNet architecture to 3D convolutions, R3D effectively learns representative features from video data, making it suitable for tasks like video classification, action recognition, and video captioning. The strength of R3D lies in its ability to deeply understand the temporal dynamics without compromising the spatial feature extraction.

2.2.7. P3D (Pseudo-3D Convolutional Network)

P3D, or Pseudo-3D Convolutional Network, is a unique approach that aims to bridge the gap between 2D and 3D convolutional networks for video processing. P3D adopts a series of 2D spatial convolutions followed by 1D temporal convolutions, rather than applying 3D convolutions directly. This method effectively decomposes the 3D convolution into two separate operations, enabling the model to learn spatial and temporal features in a more disentangled and efficient manner. The pseudo-3D approach of P3D allows for reduced computational complexity compared to traditional 3D CNNs while still capturing the essential aspects of both spatial and temporal information in videos. P3D networks have been successfully applied in various video understanding tasks, offering an innovative way to handle the challenges of video data analysis without the intensive computational demands of full 3D convolutions.

2.2.8. MVIT (Multi-Scale Vision Transformers)

MVIT, or Multi-Scale Vision Transformers, represent a novel approach in video processing, leveraging the power of transformer architectures. Unlike convolutional networks, MVIT uses self-attention mechanisms to process video data, allowing it to capture long-range dependencies and intricate patterns within the video frames. The multi-scale aspect of MVIT enables it to attend to features at various resolutions, providing a comprehensive understanding of both fine details and global context. This capability makes MVIT particularly adept at handling complex video tasks that require a nuanced understanding of spatial and temporal dynamics. MVIT's transformer-based approach offers a distinct alternative to traditional CNN based models, emphasizing the importance of global context and long-range interactions in video understanding. Its application spans across various tasks in video analysis, including classification, captioning, and enhanced representation learning.

2.3. Audio Feature Extraction

Audio preprocessing [10] is a crucial step in preparing audio data for analysis, particularly in machine learning applications.

2.3.1. Log-Mel Spectrogram

A log-Mel spectrogram is another powerful feature for audio analysis, providing a time-frequency representation of the sound. It involves computing the spectrogram (a visual representation of the spectrum of frequencies in a sound) and then mapping the frequencies to the Mel scale. Taking the logarithm of this Mel-scaled spectrogram enhances certain signal characteristics, making this representation particularly useful for tasks like environmental sound classification or music genre classification.

Log-Mel spectrograms are critical in transforming raw audio data into a more abstract and informative representation, enabling more accurate and efficient pattern recognition by machine learning models in audio analysis tasks.

2.3.2. Pretrained Audio Neural Networks (PANNs)

PANNs [11] are a groundbreaking development in the field of audio processing and analysis. These deep learning models are trained on large and diverse datasets of sounds, allowing them to learn a wide range of audio features and characteristics. PANNs are designed to be adaptable, and capable of performing various audio-related tasks such as sound classification, detection, and event recognition. Their pre-trained nature means they can be used effectively without the need for extensive training on specific tasks, making them highly efficient and versatile. PANNs are instrumental in applications like automated audio tagging, environmental sound analysis, and enhancing audio content in multimedia. PANNs represent a significant advancement in audio processing and analysis. Trained on vast and diverse audio datasets, these deep learning models efficiently learn a broad range of audio features and characteristics. PANNs are designed for efficient sound classification, detection, and event recognition. Their architecture typically combines a Log-mel Spectrogram for transforming raw audio into a time-frequency representation with about 10 CNN layers for feature extraction. This setup enables PANNs to handle complex audio processing tasks, extracting detailed audio features essential for applications such as automated audio tagging, environmental sound analysis, and multimedia enhancement.

2.4. Sequence Processing

2.4.1. Recurrent Neural Networks (RNNs)

RNN, a type of deep neural network, is able to model long-term dependencies in sequential data and suitable for NLP tasks such as speech recognition, machine translation, and image captioning [12, 13]. Each output is calculated by repeatedly processing the same function over each instance of the sequence in RNN.

RNN computes the hidden vector sequence $h = (h_1, \ldots, h_T)$ and output vector sequence $y = (y_1, \ldots, y_T)$ using the input sequence $x = (x_1, \ldots, x_T)$ with the variable length for $t = 1, \ldots, T$.

The hidden vector h_t at time step *t* is computed with the input vector x_t as $h_t = f(Wh_{t-1} + Ux_t)$ where W and U denote the weight matrices, and f denotes a nonlinear activation function such as *tanh*, *ReLU*, and *sigmoid*.

The output vector is computed as $y_t = f(Vh_t)$, where V is a matrix that connects the current hidden layer with the current output layer [12].

RNNs employ the information in arbitrarily long sequences in theory but suffer from vanishing and exploding gradients in practice and cannot capture long-term dependencies.

Despite the fact that a variety of RNN based architectures could be used, such as LSTM, as a proof of concept, the GRU is used here, which is more feasible in handling vanishing and exploding gradients problems, employed for processing sequential data to generate captions in our experiments.

2.4.2. Gated Recurrent Unit (GRU)

GRU, which is a type of RNN with a gating mechanism, has been implemented to address the aforementioned issues. GRU consists of a hidden state and two gates: update and reset [14]. In GRU, the transition has been carried on based on the following equations [14]:

$$
\begin{aligned}
\mathrm{r_t} &= \sigma(\mathrm{W_r x_t} + \mathrm{U_r h_{t-1}}) \\
z_t &= \sigma(W_z x_t + U_z h_{t-1}) \\
u_t &= \tanh\left(W x_t + U\left(r_{t \odot h_{t-1}}\right)\right) \\
h_t &= (1 - z_t) h_{t-1} + z_t u_t
\end{aligned}
\tag{2.1}
$$

where $\mathrm{r_t}$, $\mathrm{z_t}$, and $\mathrm{u_t}$ denote reset gate vector, update gate vector, and candidate hidden vector, respectively. The subscripts r and z in W_r and W_z denote the weights of the reset and update gates. σ and *tanh* are the sigmoid and tangent hyperbolic activation functions, respectively. ʘ denotes the element-wise multiplication operator. h_{t-1} is taken from the previous GRU as input, and the output of GRU, y_t is calculated with the sigmoid function as

$$y_t = \sigma(W_o h_t + b) \tag{2.2}$$

where the subscript o denotes the weight of the output vector, and b is the bias. This makes it easier to configure stacked or multi-layer GRU architectures with two or more layers that outperform the conventional RNN based architectures on many NLP tasks, including language modeling [15].

2.5. Feature Injection Architectures

Images can be incorporated into the decoder with feature injection architectures in two different ways (i.e., inject-based and merge) using a fixed-length image feature vector and linguistic feature vector (embedding vector) from the encoder and embedding layer, respectively. The inject-based architecture is designed based to utilize both image feature and linguistic feature vector to the decoder, such as init-inject, pre-inject and par-inject.

Init-inject Architecture

The hidden state vector of GRU is initialized with the same-sized image feature vector [16], and the embedding vector is fed to the GRU as an input vector.

Pre-inject Architecture

The image feature vector is utilized as the first input vector of GRU at $t = -1$, whereas the embedding vectors are fed to the GRU for the next time step [16]. The image feature vector can be considered as the first word of the sequence.

Par-inject Architecture

The image and embedding vector are concatenated as a single input before incorporating into the GRU [16].

Merge Architecture

GRU takes only the embedding vector that handles linguistic features in this architecture, whereas the image feature vector is fed into the architecture after the GRU processes the linguistic features [16]. The image feature vector and the output vector of the GRU are concatenated into a single vector to calculate the probability.

2.6. Performance Metrics

Performance metrics such as BLEU [17], METEOR [18], ROUGE-L [19], SPICE [20], and CIDEr [21] are crucial for evaluating the effectiveness of machine-generated captions, whether in image or video captioning systems. Each metric offers a unique perspective on assessing the quality of captions.

2.6.1. BLEU (Bilingual Evaluation Understudy)

This metric was originally developed for evaluating machine translation systems. BLEU measures the overlap of n-grams (a contiguous sequence of n items from a given sample of text or speech) between the machine-generated caption and one or more reference captions. It uses different versions, such as BLEU-n, where "n" denotes the n-gram size, allowing the evaluation of word pairings, triplets, etc. BLEU is known for its simplicity and speed, but it often fails to capture the semantic accuracy of the captions.

2.6.2. METEOR (Metric for Evaluation of Translation with Explicit Ordering)

Developed as an alternative to BLEU, METEOR assesses the quality of translations by considering not only exact word matches but also stem and synonym matching. It calculates the harmonic mean of unigram precision and recall, with more emphasis on recall. This metric is known for its better correlation with human judgment, especially at the sentence level.

2.6.3. ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence)

ROUGE-L is used primarily for evaluating text summarization and machine translation. It measures the longest common subsequence between the generated caption and the reference captions, focusing on the sequence rather than individual words. This metric is effective in evaluating the fluency and order of the generated text.

2.6.4. SPICE (Semantic Propositional Image Caption Evaluation)

SPICE is specifically designed for image captioning tasks. Unlike the other metrics that focus on syntactic or surface-level matching, SPICE evaluates semantic content. It parses both the generated and reference captions to assess the presence and accuracy of objects, attributes, and relationships, providing a deeper semantic evaluation of the caption quality.

2.6.5. CIDEr (Consensus-based Image Description Evaluation)

CIDEr focuses on evaluating the consensus between a generated caption and a set of reference captions. It utilizes sentence similarity measures to capture notions of grammaticality, saliency, and accuracy. CIDEr is particularly useful for captioning tasks as it evaluates the relevance and informativeness of captions in relation to the image or video content.

These metrics collectively provide a comprehensive evaluation of captioning systems. BLEU and METEOR are more suited for syntactic and surface-level evaluations, while ROUGE-L assesses fluency and sequence structure. SPICE and CIDEr delve deeper into semantic accuracy and contextual relevance, respectively. The choice of metric often depends on the specific aspects of caption quality that need to be assessed.

2.7. Relevant Datasets

Image captioning datasets are essential for training and evaluating algorithms designed for automatic caption generation. This task involves creating descriptive text for images, and such datasets are crucial in advancing this field.

2.7.1. MSCOCO Captions Dataset

One prominent dataset in this area is the Microsoft COCO (Common Objects in Context) [22] dataset. It is a large-scale resource widely utilized in the image captioning domain. The MSCOCO dataset encompasses a total of 123,287 images, which are split into two primary sets: 118,287 images in the training set and 5,000 in the validation set. Each image in this dataset is annotated with at least five reference captions. In summary, the MSCOCO dataset's extensive collection of images and its diverse set of annotations make it an ideal choice for a wide array of image captioning applications, catering to tasks that demand a general understanding of everyday contexts and objects.

Video captioning datasets are essential for training and evaluating algorithms that generate descriptions for video content. These datasets vary in size, content type, and the number of reference captions provided for each video.

2.7.2. MSR-VTT Dataset

The Microsoft Research Video-to-Text [23] dataset consists of 10,000 videos encompassing a variety of content, including news, sports, and other categories. Each video in MSR-VTT is described with 20 reference captions, providing a diverse set of descriptions for the same video content.

2.7.3. MSVD Dataset

The Microsoft Video Description dataset, also known as Youtube2Text, includes a total of 1,970 short video clips collected from YouTube. This dataset is particularly noted for its extensive set of reference captions, with each video having an average of 40 English descriptions. The MSVD is split into 1,200 training videos, 100 validation videos, and 670 test videos.

Each of these datasets offers unique characteristics. MSR-VTT and MSVD offer multiple captions per video, which is beneficial for models that need to understand and generate a variety of descriptions. The MSVD, with its large number of captions per video, is especially valuable for training robust and versatile video captioning systems. The choice of dataset typically depends on the specific requirements of the research or application, such as the type of content, the desired level of detail in the captions, and the computational resources available for training models.

2.8. Programming Languages and Tools

2.8.1. Python

Python is a high-level, interpreted programming language known for its simplicity and readability. Its straightforward syntax allows programmers to express concepts in fewer lines of code compared to languages like C++ or Java. Python supports multiple programming paradigms, including object-oriented, procedural, and functional programming. It comes with an extensive standard library and has a large and active community, making it versatile for a wide range of applications, from web development to data analysis, machine learning, and scientific computing.

2.8.2. PyTorch Framework

PyTorch is an open-source machine learning framework widely used in the field of artificial intelligence and deep learning. Developed by AI Research lab of Facebook, it is known for its flexibility, ease of use, and dynamic computational graphing, enabling more intuitive coding of complex AI models compared to static-graph frameworks. PyTorch provides a rich set of tools and libraries for deep learning, and it is particularly favored for its efficient memory usage and optimization, especially when working with large neural networks. Its compatibility with Python and seamless integration with other Python-based scientific computing libraries make PyTorch a preferred choice for researchers and developers in developing, training, and deploying AI models.

torchvision is a specialized library within the PyTorch ecosystem, focused on computer vision applications. It provides a rich collection of pre-trained models, datasets, and image transformations, facilitating the development and training of deep learning models in the field of image processing and analysis. torchvision includes implementations of advanced CNNs such as ResNet50, Inception-v3, and DeepLabv3, which can be used for tasks like image classification, object detection, and segmentation. The library also offers utilities for loading and normalizing various standard vision datasets, making it a comprehensive resource for researchers and developers working on computer vision projects.

torchaudio complements PyTorch by providing tools specifically designed for audio processing. This library extends the PyTorch framework to the auditory domain, enabling the creation, manipulation, and analysis of sound signals using deep learning. It includes a variety of datasets, pre-trained models, and transformations that are essential for tasks such as speech recognition, audio classification, and sound generation. torchaudio also offers functionalities for loading and preprocessing audio data, making it an invaluable tool for researchers and practitioners in fields like music technology, linguistics, and auditory scene analysis.

torchtext is a PyTorch domain library aimed at simplifying the preprocessing of textual data and making it more accessible for NLP applications. This library provides easy-to-use abstractions and interfaces for handling text data, including utilities for loading, tokenizing, and batching datasets. torchtext also supports building custom datasets and iterators, which are essential for training language models. With built-in support for common public NLP datasets and word embeddings, torchtext allows researchers and developers to focus more on model design and less on data preprocessing, streamlining the development of NLP applications such as text classification, machine translation, and sentiment analysis.

Chapter 3
Multi-layer Gated Recurrent Unit based Recurrent Neural Network for Image Captioning

Generating natural language descriptions of an image, namely image captioning, has received much attention in computer vision and natural language processing. Recent image captioning models are mainly based on the encoder-decoder framework in which visual information is extracted by an encoder, e.g. using CNN, and captions are generated by a decoder, e.g. using RNN. Although this framework is promising for image captioning, there are still issues in the RNN decoder for exploiting the visual information to generate grammatically and semantically correct captions. More specifically, the RNN decoder has limited ability in dealing with long-term complex dependencies, leading to ineffective use of contextual information from the encoded data. To address this issue, in this paper, we introduce a multi-layer gated recurrent unit (ML-GRU) within the conventional RNN decoder, which enables the modulation of the relevant information flow inside the unit, and thus leads to the generation of semantically coherent captions. The proposed ML-GRU based RNN decoder has been extensively evaluated on the MSCOCO dataset, and experimental results demonstrate the advantage of our proposed approach over the state-of-the-art approaches across multiple performance metrics.

3.1. Introduction

Image captioning aims to generate grammatically correct and human-readable descriptions of an image using techniques from computer vision (CV) and natural language processing (NLP). This task leverages the connection between CV and NLP and has attracted increasing interest, due to its potential applications such as image indexing or retrieval and virtual assistants for visually impaired people [24-26] [27-29]. Image captioning is a challenging task because it requires an advanced level of understanding of an image, including the recognition of the objects and actions in the image, in order to generate meaningful captions with proper linguistic properties. Therefore, it goes beyond the conventional CV tasks such as image classification and object detection. Early efforts to address this problem in the literature have considered the use of either retrieval-based or template-based models before using deep neural networks. Recently, the encoder-decoder [30] based neural structure has emerged, which is promising and has become a popular model for image captioning. This model is composed of two sub-networks, where the encoder aims to generate a feature representation of an image using methods such as CNN, while the decoder translates this representation into natural language descriptions using methods such as RNN.

For the encoders of the captioning systems, the CNN architectures like Inception-v3 [31], NASNet-Large [32] (neural architecture search network), Xception [9], and ResNet152 v2 [3] are popular choices. Inception-v3 [31] is a 42-layered deep CNN architecture that uses the asymmetric approach to decompose a kernel of large-scale convolution into a small-scale kernel of convolution. NASNet [31] is designed using reinforcement learning and contains two types of cells, namely, the normal cell, which keeps the width and height of the feature map, and the reduction cell, which reduces the width and height of the feature map by half. Xception [9] is a deep CNN consisting of 36 convolutional layers with 14 modules that have linear residual connections around them and a logistic regression layer for feature extraction. This architecture is obtained by modifying Inception-v3 with depthwise separable convolutional layers. ResNet152 v2 [9] is a deep CNN, which is composed of residual nets with 152 layers. Unlike the ResNetV1, this architecture uses the normalization of the stack before each weight layer. The ResNet152 v2 architecture with the removed classification layer extracts the high-level image feature vector of the input image using convolution and pooling layers. The visual information of images extracted by the encoders is then utilized in language decoders to convert this information word-by-word into natural language captions. The conventional RNN based decoders, however, have vanishing and exploding gradient problems. As a result, they are not effective in exploiting long-term temporal dependencies [1]. Long short-term memory (LSTM) [33] and GRU [14] networks are proposed to address these problems. LSTM uses memory cells to retain information for long periods, while GRU does not use additional memory cells to maintain the flow of information.

When the RNN based language decoders are used for caption generation, the visual information can be fed either directly into the RNN or in a layer preceding the RNN [16, 25]. Several RNN based architectures have been proposed, which can be categorized into the following four: init-inject [34], pre-inject [35], par-inject [36], and merge [37] . The visual information can be fed as a latent vector to the initial hidden state of the RNN in init-inject [2, 38], while the latent vector is used as the initial input of the RNN in the pre-inject architecture [39]. The latent vector is used with the word vectors of the caption prefix in parallel as an input to the RNN in the par-inject architecture [40]. Different from the above architectures, the latent vector is not fed to the RNN directly in the merge architecture as the image is presented to the language model after the caption prefix is generated by the RNN [16, 41].

Although the current encoder-decoder framework improves captioning accuracy compared to its counterparts, effectively extracting and employing contextual information from encoded data remains a challenge that results in insufficient performance in captioning. This paper introduces a novel image captioning model that utilizes NASNet-Large for image encoding and a multi-layer GRU based decoder under the init-inject architecture, thereby enhancing the use of visual information for accurate caption generation. Based on extensive experimental studies, NASNet-Large is found to be adequate for encoding visual information. The motivation behind using GRU is two-fold. First, GRU needs fewer parameters and is computationally cheaper than LSTM as GRU has one hidden state vector while LSTM has two state vectors, namely, hidden and cell states [42]. In addition, GRU has two gates, i.e. the update and reset gates, while LSTM has three gates, i.e. the input, forget, and output gates. Second, the GRU with one hidden state vector offers an excellent fit for the requirement of the init-inject architecture in terms of computational efficiency in practical implementation [16]. The number of layers in GRU is incremented to ensure the modulation of the most relevant information flow inside the unit. A higher number of upper layers deployed in the multi-layer GRU can provide detailed contextual information from the data, thereby providing an enhanced prediction model [43, 44].

Table 3.1 Comparison of different CNN encoders with single-layer GRU.

CNN	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	SPICE	CIDEr
ResNet152 v2	0.686	0.503	0.359	0.258	0.497	0.221	0.148	0.801
Inception-v3	0.693	0.513	0.368	0.264	0.506	0.230	0.161	0.851
Xception	0.702	0.520	0.373	0.265	0.508	0.230	0.162	0.859
NASNet-Large	0.707	0.524	0.376	0.270	0.510	0.231	0.161	0.876

As GRUs are operated on sequence data, adding layers will increase the level of abstraction over time for input observations. In turn, this can provide chunking of observations over time or represent the data at various time scales.

Integrating an ML-GRU into RNN enhances the ability of the decoder to retain important semantic image information, thereby improving caption generation. Furthermore, to achieve high-quality image features, we implemented NASNet-Large. This integration enriches the quality of the encoded data, thus elevating the coherence of the generated captions. Although ML-GRUs are widely used in various applications, our study presents an implementation within the field of image captioning for the first time. Our approach combines a structured integration of GRU layers under the init-inject architecture, with each GRU layer fine-tuned.

This approach ensures accurate recognition of visual elements and their meaningful linguistic translation. By adopting such a precise configuration, the approach aims to advance the state-of-the-art image captioning with enhanced accuracy and contextual relevance in the generated captions. Experimental results on the MSCOCO dataset show the advantage of our proposed approach over the state-of-the-art approaches for caption generation with a higher performance metric score.

The major contributions of this study can be summarized as follows.

- We propose a new approach to the neural encoder-decoder framework of image captioning by introducing multi-layer GRU based RNN, which refines the decoder to evaluate the image attributes extracted in the encoder for enhanced image captioning. This approach was designed under the init-inject architecture, and to the best of our knowledge, this is the first time that the

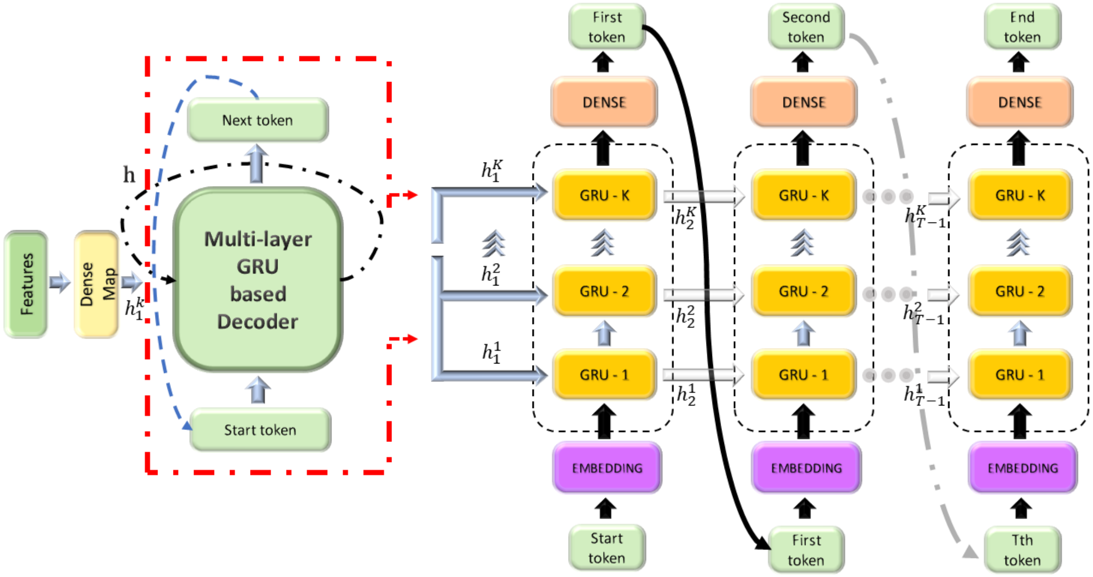


Figure 3.1 The proposed multi-layer GRU based decoder (inside the red dashed line) is given on the left side while unfolded on the right side.

multi-layer GRU is exploited in the encoder-decoder based image captioning models.

- We perform comprehensive experimental results on the MSCOCO dataset and show that the proposed model significantly outperforms the state-of-the-art approaches consistently across different performance metrics. We also investigate the optimal number of GRU layers to be used for image captioning.

3.2. Proposed Multi-layer GRU based Image Captioning Approach

This section presents a new approach to enhance the image captions by introducing multi-layer GRU to the image decoder. The proposed image captioning approach consists of two steps: image encoder and text decoder. First, the image encoder is utilized to extract features from an image. Then, these features are fed into the text decoder that processes the features to generate a caption word-by-word. CNN based encoder employed here is a recently emerged framework that has been found to be promising for feature extraction of an image. The NASNet-Large model is utilized as a CNN architecture where all image features are obtained after the average pooling layer, which returns a 4032-element vector.

The methodology proposed in this study involves a multi-layer GRU based decoder, as depicted in Figure 3.1. This decoder provides a novel solution to several limitations in the current literature, such as efficient visual attributes injection and modulation of the relevant information flow. The decoder architecture comprises an embedding layer, multiple GRU layers, and a dense layer, which are utilized under the init-inject architecture. This architecture facilitates the parallel processing of image features obtained from a dense map and linguistic features, derived from the embedding layer.

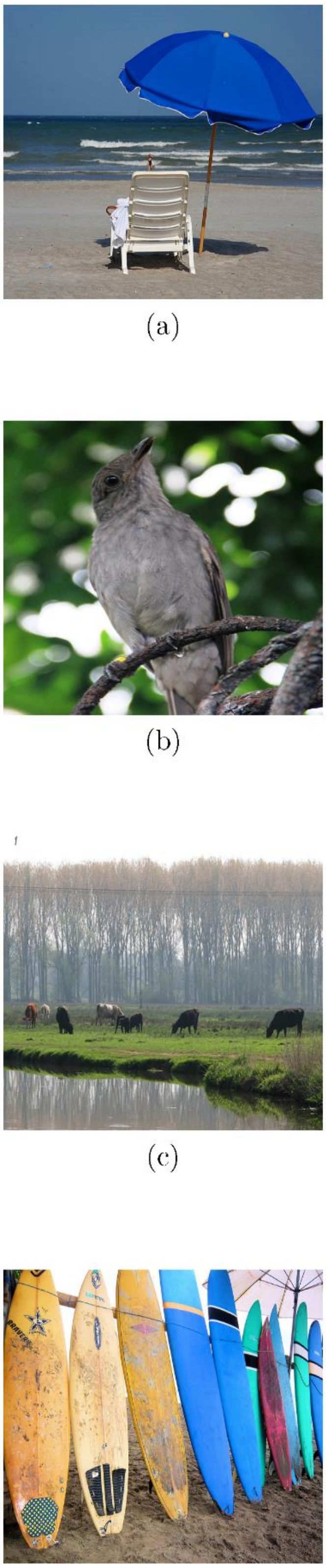
The multi-layer GRU is a combination of K-GRU for $k = 1, \dots, K$, while $h_t^{(k)}$ and $x_t^{(k)}$ are defined as the hidden and input vector for the kth GRU layer. Each initial hidden vector ($h_1^{(k)}$) contains image features as a separate vector with reduced size from 4032 to the 512-element vector by the dense map to feed the multi-layer GRU at $t = 1$. For the subsequent iterations, multi-layer GRU is fed by the updated hidden vector from the previous iteration ($h_{t-1}^{(k)}$) rather than the dense map. The first GRU layer is located after the embedding layer, which generates the predefined size of a meaningful embedding vector, namely the linguistic features, using the start token. The embedding vector is processed at the first GRU layer, leading to the first output vector ($y_1^{(1)}$), which is the input of the next GRU layer ($x_1^{(2)}$) . The same procedure is repeated K times until ($y_1^{(K)}$) is generated, which is the input for the dense layer as:

$$y_1^{(K)} = \sigma\left(W^{(K)} h_1^{(K)} + b^{(K)}\right) \tag{3.1}$$

To generate the first token, the *argmax* function has been employed on the output of the dense layer d_1, which is computed as:

$$d_1 = f\left(W y_1^{(K)} + b\right) \tag{3.2}$$

then the output and hidden state are carried to the RNN as input and hidden state, respectively, to generate the next token. This process is continued until an end-of-caption token is generated. In the end, the generated tokens are converted into their corresponding words employing a vocabulary that is created from the reference captions of the training set.



(a)

Ground Truth Captions

A woman walks out of the ocean towards a beach chair and umbrella

This is a chair and umbrella that is sitting near an ocean

A beach chair and umbrella in the sand on the beach

A chair and umbrella sitting on a beach near a person

a chair and a umbrella that is on a beach

Generated Caption

a chair with a blue umbrella sitting on the sand

(b)

Ground Truth Captions

There is a bird sitting on a tree branch

The bird is sitting on the small branch of the tree

A bird is perched on a twig in the trees

There is a bird perched on the tree branch

A gray bird is standing on small brown branch

Generated Caption

a bird is sitting on a branch of a tree

(c)

Ground Truth Captions

A group of cows grazing in a field near a body of water

Several animals standing in the grass near a lake

Several cows grazing on grass near water with trees in the background

a herd of cows graze lazily by the pond

A herd of cattle grazing on top of a grass covered field

Generated Caption

a herd of cattle grazing on a lush green field

(d)

Ground Truth Captions

A row of surfboards sticking out of the sand sitting next to each other

a row of surf boards placed in the sand

Several surfboards standing in a row on the beach

A row of surfboards leaned up against a wood rail in the sand

Many surfboards are propped against a rail on the beach

Generated Caption

a bunch of surf boards lined up in a row

Figure 3.2 The generated captions by our proposed approach for four different images from the MSCOCO dataset.

3.3. Experimental Evaluations

This section evaluates the proposed captioning approach on the MSCOCO dataset [22], and a performance comparison with state-of-the-art approaches is presented. The MSCOCO dataset contains 118287 training, 41000 test, and 5000 validation images [22] and each image is described with five reference captions. MSCOCO is the most suitable dataset for the evaluation of our proposed image captioning approach due to its various images with semantically rich reference captions. To analyze the performance of the compared captioning approaches, several metrics are employed, including BLEU [17], CIDEr [21], METEOR [18], ROUGE-L [19] and SPICE [20]. Our results are sorted based on CIDEr and SPICE metrics due to their better correlation with human assessment compared to BLEU-n, METEOR, and ROUGE-L.

Baselines

- [45] proposes a framework named StyleNet, which utilizes a factored LSTM to extract the style factors in captions.
- [46] proposes a captioning system which can tailor the captions for application specific studies.
- [34] proposes a captioning system named SemStyle, which generates a semantically accurate styled caption.
- [39] proposes a deep neural image captioner based on an encoder-decoder framework where Inception-v3 as the encoder and LSTM based sentence generator as the decoder.
- [47] proposes a gLSTM which is specifically developed to employ visual attributes to the LSTM for image captioning.
- [48] proposes a phi-LSTM to describe visual contents as a sentence that employs phrases rather than words.
- [6] proposes a CNN+CNN framework that processes natural language attributes with a CNN.
- [49] proposes a Mixture of Recurrent Exports for image captioning, which captures the semantics and generates a styled caption.
- [50] proposes an attention based image captioning where the visual attributes are weighted with *soft* and *hard* attention techniques.

Table 3.2 Comparison of different CNN encoders with multi-layer GRU.

CNN	# of Layers	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	SPICE	CIDEr
ResNet152 v2	3	0.679	0.494	0.347	0.244	0.488	0.219	**0.150**	0.782
	6	0.675	0.492	0.349	0.248	0.490	**0.221**	0.150	**0.786**
	9	**0.683**	**0.498**	**0.352**	**0.249**	**0.493**	0.219	0.148	0.778
	12	0.546	0.326	0.184	0.105	0.414	0.152	0.090	0.420
	15	0.544	0.325	0.182	0.104	0.411	0.151	0.093	0.421
Inception-v3	3	0.678	0.499	0.356	0.254	0.496	**0.229**	**0.158**	**0.821**
	6	0.680	0.500	0.357	0.254	0.496	0.227	0.157	0.818
	9	**0.689**	**0.506**	**0.362**	**0.258**	**0.497**	0.225	0.154	**0.821**
	12	0.547	0.334	0.192	0.112	0.420	0.154	0.091	0.452
	15	0.555	0.335	0.191	0.110	0.417	0.157	0.093	0.453
Xception	3	0.698	0.513	0.366	0.261	0.501	0.228	0.160	0.846
	6	0.694	0.509	0.363	0.259	0.499	0.227	0.158	0.844
	9	**0.702**	**0.519**	**0.371**	**0.263**	**0.505**	**0.229**	**0.162**	**0.850**
	12	0.692	0.507	0.358	0.251	0.496	0.221	0.155	0.792
	15	0.559	0.343	0.195	0.115	0.420	0.158	0.093	0.467
NASNet-Large	3	0.690	0.513	0.369	0.266	0.506	**0.237**	**0.169**	**0.884**
	6	0.695	0.514	0.370	0.265	0.505	0.236	0.168	0.884
	9	**0.705**	**0.522**	**0.374**	**0.268**	**0.507**	0.235	0.168	0.878
	12	0.559	0.343	0.197	0.115	0.427	0.161	0.097	0.488
	15	0.561	0.343	0.195	0.113	0.422	0.161	0.100	0.484

To construct an image captioning system with high performance, we have analyzed four different CNN architectures as an encoder in conjunction with a multi-layer GRU based decoder. In this regard, the Inception-v3, Xception, ResNet152 v2, and NASNet-Large with five different layer-sized GRU were evaluated in terms of BLEU-n, CIDEr, METEOR, SPICE, and ROUGE-L metrics. All these configurations have been evaluated based on hyper-parameter optimization.

Our proposed multi-layer GRU based decoder takes linguistic features from the embedding layer. Two critical parameters based on linguistic features for the performance of image caption generation are the embedding vector size and the vocabulary size. The size of the embedding vector is typically set between 50 and 300 [51].

The embedding vector with a small size does not capture the word relations completely, whereas the large embedding vectors cause overfitting. The size of the embedding vector affects the training time, computational costs, and the performance of embedding.

Table 3.3 Comparison of our proposed Multi-layer GRU based approach with some state-of-the-art architectures on MSCOCO dataset.

	BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**	**ROUGE-L**	**METEOR**	**SPICE**	**CIDEr**
[45]	0.625	-	-	0.212	-	0.218	0.135	0.664
[46]	-	-	-	**0.270**	0.500	**0.240**	0.009	0.680
[34]	0.653	0.478	0.337	0.238	0.482	0.219	0.157	0.769
[39]	0.667	-	-	0.238	-	0.224	0.154	0.772
[47]	0.663	0.485	0.354	0.262	-	0.230	-	0.813
[48]	0.666	0.489	0.355	0.258	0.497	0.231	0.165	0.821
[6]	0.688	0.513	0.370	0.265	0.507	0.234	-	0.839
[49]	0.679	0.501	0.356	0.252	0.501	0.226	0.166	0.844
[50]	-	-	-	0.250	**0.516**	0.230	-	0.865
Our proposed 9-layer GRU	**0.705**	**0.522**	**0.374**	0.268	0.507	0.235	**0.168**	**0.878**

The vocabulary size, which has a critical role in the image captioning tasks, is determined based on the number of common words in all reference captions and usually varies from 10000 to 40000 words [52]. To optimize the embedding vector and vocabulary values, our proposed multi-layer GRU based decoder is tested under ten different vocabulary sizes, including 250, 500, 750, 1000, 2000, 3000, 5000, 10000, 20000, and 40000, and eight different embedding vector sizes (namely, vector sizes with 25, 50, 75, 100, 125, 150, 200, and 250). The optimization tests were carried out by keeping one of two parameters fixed due to the high training time and computational cost. In the encoder side, the Inception-v3 is employed as a reference CNN architecture, while a single-layer GRU based decoder is used in the decoder.

First, this reference system was evaluated under different performance metrics with ten different vocabularies and the embedding vector of fixed-size, as 100. The best CIDEr metric was observed when the vocabulary size was 750. Then, the reference system was evaluated under the same performance metrics with eight different embedding vectors and the vocabulary of fixed-size, as 750. The best CIDEr metric was observed when the embedding vector size was 100. Hence, the embedding vector size and vocabulary size have been determined based on the empirical analysis of the aforementioned configurations.

Three different CNN based encoders (i.e., Xception, ResNet152 v2, and NASNet-Large) are employed to observe the best CNN architecture compatible with these embedding vector and vocabulary sizes, 100 and 750, respectively. The evaluation results are given in Table 3.1. The NASNet-Large based encoder outperforms the other three CNNs. The experiments were employed on NASNet-Large CNN architecture as a reference due to its promising results. To find the optimum parameters according to layer size, NASNet-Large with three-layer GRU was evaluated under the same performance metrics with ten different vocabularies and the embedding vector of fixed-size, as 100. The best CIDEr metric was observed when the vocabulary size was 10000. Then, the best CIDEr metric was observed when the embedding vector size was 125. Using these parameters (10000 for vocabulary and 125 for embedding vector), three CNN (Inception-v3, ResNet152 v2, and Xception) based encoder was employed to observe the best result for 3-layer GRU. Applying the same strategy to the 6, 9, 12, and 15 layer GRU, the optimum parameters were determined as 150, 75, 250, and 200 for the embedding vector size; 20000, 2000, 2000, and 40000 for the vocabulary size, respectively.

The empirical analysis with different vocabulary sizes with a fixed-size embedding vector indicates that the CIDEr metric gradually increases until the 9-layer GRU, where the maximum level has been reached. The performances of multi-layer GRU with four different CNN configurations have been listed in Table 3.2. The empirical results listed in Table 3.2 indicate that increasing the number of layers until 12-layer can enhance the predictive performance in the proposed image captioning system. Among all the configurations, 9-layer GRU architecture outperforms the other compared schemes in terms of BLEU-n and ROUGE-L metrics, and 3-layer GRU architecture outperforms the other schemes in terms of METEOR, SPICE, and CIDEr.

Table 3.3 presents a comprehensive performance evaluation of the proposed 9-layer GRU against various contemporary image captioning architectures utilizing the MSCOCO dataset. Evaluation employs metrics such as BLEU-1 to BLEU-4, ROUGE-L, METEOR, SPICE, and CIDEr. The proposed model consistently surpasses others, especially in BLEU-1, BLEU-2, BLEU-3, SPICE, and CIDEr, highlighting advanced context interpretation and description generation capabilities. Furthermore, this is further supported by competitive performance in BLEU-4, ROUGE-L, and METEOR.

The integration of a 9-layer GRU facilitates the handling of complex temporal dynamics and ensures the maintenance of comprehensive context during caption generation. The results underscore the potential of the proposed model in progressing image captioning research.

The approaches are sorted based on CIDEr metrics, and the highest score is indicated with bold fonts in each column. The proposed approach outperforms BLEU-1, BLEU-2, SPICE, and CIDEr metrics. Figure 3.2 shows the ground truth and generated captions by the proposed approach for four images. From those results, we observe that our proposed approach is capable of capturing image information with correct and descriptive captions. For instance, in the first image (Figure 3.2 (a)), the generated caption can successfully describe the *chair* and *umbrella* with its color in the image. In the second image (Figure 3.2 (b)), the proposed approach identifies a *branch* and the action of *sitting*; in the third image (Figure 3.2 (c)), it identifies *cattle* and the action of *grazing*. In the fourth image (Figure 3.2 (d)), the proposed approach generates the words *surfboard* and *row*, which accurately describe the content of the image. Images show that our proposed approach can generate natural sentences related to the image.

Chapter 4
Leveraging Pre-trained 3D-CNNs for Video Captioning

Video captioning is a visual understanding task that aims to generate grammatically and semantically accurate descriptions. One of the main challenges in video captioning is capturing the complex dynamics present in videos. This study addresses this challenge by leveraging pre-trained 3D-CNNs. These networks are particularly effective at modeling such dynamics, enhancing video contextual understanding. We evaluated the approach on the MSVD dataset, with commonly utilized performance metrics in video captioning including CIDEr, BLEU-1 through BLEU-4, ROUGE-L, METEOR, and SPICE. The results show significant improvements across all these metrics, proving the advantage of pre-trained 3D-CNNs in enhancing video captioning accuracy.

4.1. Introduction

Video captioning is a task that involves generating descriptions for video frames by leveraging techniques from natural language processing and computer vision fields. These descriptions are expected to be grammatically correct and semantically accurate. Recently, there has been increased attention on video captioning studies due to their potential applications in video understanding, video retrieval, and video caption generation [5, 53-55].

Earlier studies in captioning have explored various approaches, including template-based, retrieval-based, and deep learning-based. One template-based approach uses a predefined template to translate semantic representation into a caption [56]. The retrieval-based approach employs a compositional semantics language model that breaks down video descriptions into subjects, verbs, and objects. These elements are then transformed into word vectors, effectively capturing the meaning of the content [57].

Recently, deep learning-based approaches have emerged as valuable tools for generating more accurate captions [24, 25, 27-29, 41, 43, 44, 58, 59]. These approaches leverage deep learning to manage the complexity of videos, including diverse objects, scenes, and actions. Various deep learning-based encoder-decoder architectures have been proposed. These architectures typically combine CNNs to extract features and RNNs for caption generation [13, 60-66]. There are various CNN architectures commonly employed in the encoder for feature extraction from video frames to feed RNN based decoders [8, 9, 31, 67]. However, conventional RNNs encounter challenges such as vanishing and exploding gradient issues, limiting their ability to process long input sequences due to short-term memory. Two types of RNNs have been proposed to address these challenges: LSTM and GRU. LSTM networks introduce three gates: the input gate, the forget gate, and the output gate. These gates, along with two states known as the hidden state and memory cells, enable LSTMs to capture long-term dependencies in sequences effectively. On the other hand, GRU networks consist of a hidden state and two gates: the update and the reset gate. GRUs can dynamically determine, by utilizing these gates, the amount of information to retain from previous time steps and update their hidden state accordingly. This enables GRUs to model dependencies in sequences with varying lengths.

A video captioning approach that utilizes the encoder-decoder architecture incorporates a hierarchical recurrent neural encoder (HRNE) with a two-layer LSTM [68]. The HRNE extracts temporal features from video frames, which serve as input for the LSTM-based decoder that generates captions. The LSTM hidden state and memory cell are carried forward to the next step, except when a new video time boundary is detected.

The Sequence-to-Sequence Video-to-Text (S2VT) approach was proposed for video captioning to capture the temporal structure of videos and represent them as fixed-length vectors. This S2VT approach employs LSTMs in both its encoder and decoder, facilitating the encoding of the temporal structure of video and the generation of captions [69].

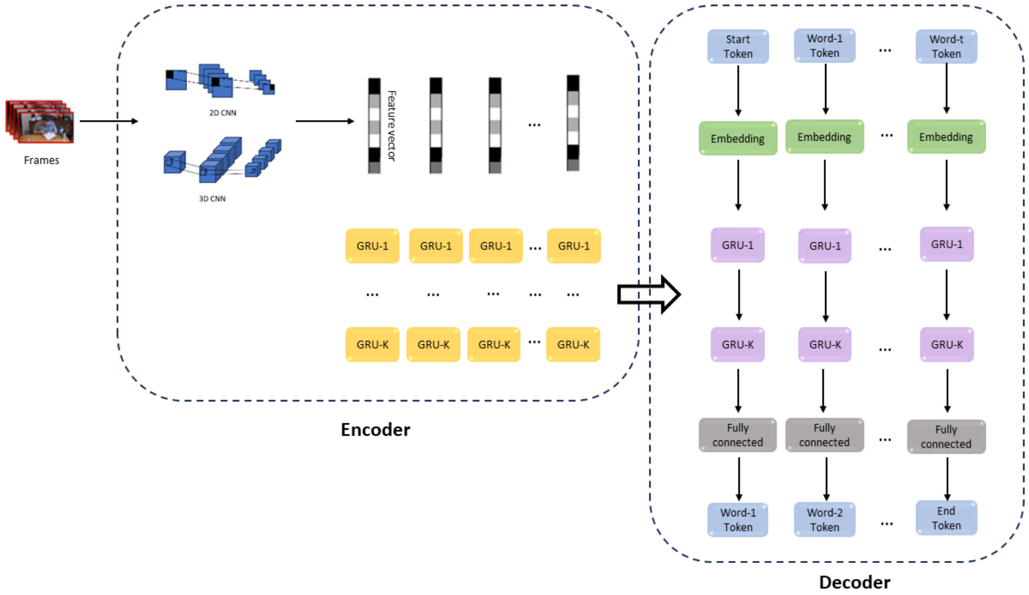


Figure 4.1 Proposed pre-trained 3D-CNNs-based approach for video captioning

In this paper, we propose a video captioning approach with a combination of two-dimensional (2D) and 3D-CNN architectures and multi-layer GRU to extract features of the videos on the encoder side. Inception-v3 as 2D-CNN is employed to extract appearance features from video frames, whereas S3D, R3D, P3D, and MVIT as 3D-CNNs are utilized for the motion features. Then, a multi-layer GRU is employed to preserve the semantic information of the video and leverage contextual information more effectively. On the decoder side, a multi-layer GRU is utilized to generate more accurate captions by leveraging its ability to compute complex representations. Experimental results are obtained on the MSVD dataset using various evaluation metrics, including BLEU-n [17], CIDEr [21], METEOR [18], ROUGE-L [19], and SPICE [20]. These metrics are used to measure the accuracy of the proposed approach on captioning performance and to compare with state-of-the-art approaches.

4.2. Proposed 3D-CNN based Video Captioning Approach

Here, we introduce our proposed approach as shown in Figure 4.1 for video captioning based on sequence-to-sequence learning which utilizes pre-trained 3D-CNNs.

The proposed video captioning approach is employed under the encoder-decoder framework. In this framework, the encoder extracts visual attributes from videos.

These extracted attributes are then fed into the decoder, which generates descriptive captions detailing events and scenes corresponding to relevant parts of the video.

For each iteration, the multi-layer GRU of the encoder receives the updated hidden state from the previous iteration until it reaches the last feature vectors. The final hidden state of the multi-layer GRU in the encoder is then passed to the decoder for caption generation. The video decoder consists of an embedding layer, a multi-layer GRU, and a fully connected layer. Caption generation begins with a predefined start token at $t = 0$ and continues for a variable length T. The embedding layer transforms each token into a meaningful latent vector containing linguistic features. The latent vector is then provided as input to the first GRU layer. The output from this layer is then transferred to the following layer.

This procedure is carried out K times, with K denoting the total count of GRU layers. The output of the multi-layer GRU is then directed into a fully connected layer, which calculates the prediction probabilities and generates the subsequent word in the caption. The fully connected layer generates the token for the first word (word- 1), which will be used in the following step. This word generation procedure continues for T iterations until the end token is reached.

All generated tokens are converted into their corresponding words to form a caption. To evaluate the impact on captioning performance, we varied the number of GRU layers, testing configurations with 1, 2, and 4 layers on both the encoder and the decoder sides.

4.3. Experimental Evaluations

We chose the MSVD dataset for the evaluation of our proposed video captioning approach due to its extensive reference captions. The performance of the video captioning approaches is evaluated using several metrics, including BLEU-n (n = 1, 2, 3, 4), METEOR, ROUGE-L, SPICE, and CIDEr. CIDEr is often used to sort the results in image and video captioning tasks due to its better correlation with human judgment than BLEU-n, METEOR, SPICE, and ROUGE-L. For our evaluation, we prioritize the CIDEr metric to sort results, as it aligns more closely with human judgment compared to the other metrics.

Table 4.1 Comparison of different 3D-CNN architectures with Inception-v3

Design	# of Layers	CIDEr	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE-L	METEOR	SPICE
S3D + Inception-v3	1	0.860	0.494	0.588	0.678	0.802	**0.712**	**0.350**	0.058
	2	**0.863**	0.493	0.588	0.685	0.807	**0.712**	**0.350**	0.060
	4	0.789	0.492	**0.593**	**0.693**	**0.814**	0.702	0.335	**0.064**
R3D + Inception-v3	1	0.770	0.442	0.547	0.618	0.780	0.692	0.330	0.054
	2	0.809	0.453	0.550	0.654	0.784	0.700	0.330	0.055
	4	0.850	**0.502**	0.585	0.684	0.808	0.711	0.339	0.061
P3D + Inception-v3	1	0.785	0.462	0.561	0.651	0.774	0.700	0.329	0.054
	2	0.822	0.478	0.576	0.672	0.793	0.708	0.337	0.058
	4	0.808	0.477	0.584	0.684	0.805	0.704	0.330	0.063
MVIT + Inception-v3	1	0.803	0.458	0.561	0.661	0.786	0.700	0.330	0.055
	2	0.716	0.465	0.562	0.653	0.782	0.699	0.333	0.056
	4	0.820	0.482	0.582	0.680	0.801	0.708	0.333	0.058
Inception-v3	4	0.715	0.491	0.591	0.692	0.813	0.701	0.334	0.063
S3D	4	0.788	0.491	0.592	0.691	0.813	0.701	0.335	0.063
R3D	4	0.513	0.370	0.471	0.574	0.720	0.651	0.288	0.043
P3D	4	0.230	0.270	0.373	0.488	0.670	0.621	0.240	0.038
MVIT	4	0.181	0.330	0.490	0.601	0.742	0.611	0.240	0.040

Table 4.1 comprehensively evaluates various 3D-CNN architectures paired with Inception-v3, using CIDEr, BLEU (1-4), ROUGE-L, METEOR, and SPICE metrics. The S3D+Inception-v3 Multi-layer GRU with 2 layers demonstrated superior performance, yielding a CIDEr score of 0.863, which indicates its enhanced ability to generate accurate descriptions of videos aligned with human annotations. Furthermore, it showed consistent performance across the BLEU-3, BLEU-2, and BLEU-1 metrics. The four-layer S3D+Inception-v3 Multi-layer GRU outperformed in terms of the SPICE metric, highlighting its proficiency in evaluating semantic content. Moreover, the R3D+Inception-v3 Multi-layer GRU with 4 layers achieved a remarkable BLEU-4 score of 0.502.

Table 4.2 benchmarks the proposed S3D with a 2-layer GRU against state-of-the-art approaches on the MSVD dataset. Remarkably, the proposed approach achieves the highest CIDEr 0.863 and METEOR 0.350 scores, indicating enhanced video description quality.

Although our approach excels in BLEU-4 0.493, indicating relevant and coherent long caption generation, it is outperformed by [70] in the metrics BLEU-1 to BLEU-3. This demonstrates that their method generates short captions more accurately.

Table 4.2 Performance metric comparison of the proposed approach with state-of-the-art architectures on the MSVD dataset.

	CIDEr	BLEU-4	BLEU-3	BLEU-2	BLEU-1	METEOR
[68]	-	0.438	0.551	0.663	0.792	0.331
[71]	0.635	0.425	-	-	-	0.324
[72]	0.517	0.419	0.526	0.647	0.800	0.296
[70]	0.658	0.499	0.604	0.704	0.815	0.326
[73]	-	0.453	0.554	0.660	0.788	0.310
Proposed S3D with 2-layer GRU	0.863	0.493	0.588	0.685	0.807	0.350

The results emphasize the advanced semantic caption generation of the proposed approach while comparing the competitive domain of video captioning architectures.

Chapter 5
Knowledge Distillation for Efficient Audio-Visual Video Captioning

Automatically describing audio-visual content with texts, namely video captioning, has received significant attention due to its potential applications across diverse fields. Deep neural networks are the dominant methods, offering state-of-the-art performance. However, these methods are often undeployable in low-power devices like smartphones due to the large size of the model parameters. In this study, we propose to exploit simple pooling front-end and down-sampling algorithms with knowledge distillation for audio and visual attributes using a reduced number of audio-visual frames. With the help of knowledge distillation from the teacher model, our proposed method greatly reduces the redundant information in audio-visual streams without losing critical contexts for caption generation. Extensive experimental evaluations on the MSR-VTT dataset demonstrate that our proposed approach significantly reduces the inference time by about 80% with a small sacrifice (less than 0.02%) in captioning accuracy.

5.1. Introduction

Audio-visual video captioning aims to generate grammatically and semantically meaningful sentences for the content of audio-visual media, driven by applications such as video indexing or retrieval and virtual assistants for visually and hearing impaired people [24, 28].

This task involves several challenges, such as identifying objects and scenes in the video frame, extracting audio attributes, and audio-visual fusion to describe the content with certain grammatical structures and semantics [25, 74-76].

These issues could be addressed with the release of large-scale datasets and advances in deep learning, which has led to the development of highly complex networks with improved caption generation. However, this can also lead to high computational cost due to the increased complexity of the networks and scale of the datasets. One approach to overcome this issue is to use efficient audio and visual feature extraction networks as they provide faster inference time [77]. These networks can be categorized into four classes: namely, model compression [7, 78], knowledge distillation [79-81], efficient networks [11, 82], and simple pooling front-ends (SimPFs) [83]. A framework that applies passive filter pruning to reduce the number of convolutional filters is proposed for a compressed CNN [78].

Similarly, a low-complexity CNN architecture is presented in [7], by reducing model parameters and memory usage. A BERT architecture is proposed as a teacher network that provides soft labels to guide a seq2seq network for audio speech recognition [79]. In a highway deep neural network, knowledge distillation and teacher-student training are leveraged to achieve improved accuracy with a reduced number of parameters [80]. PANNs [11], which are trained on AudioSet [84], can be transferred to audio-related tasks such as audio classification and captioning [85-87]. SimPFs are employed to reduce the required number of audio frames by reducing floating point operations on a network for efficient audio classification [83].

For visual feature extraction, knowledge distillation is used in [88] to generate soft labels for simpler networks to be deployed on a device with low computing resources. Similarly, knowledge distillation with an attention mechanism is used in [89], which groups high-dimensional features into low-dimensional vectors. Furthermore, [90] uses all the visual frames in a video to train the teacher network. The student network then uses uniformly down-sampled frames and mimics the teacher for efficient video classification.

In this study, we propose an efficient audio-visual captioning method based on the teacher-student network, which uses knowledge distillation for audio and visual feature extraction with a reduced number of frames, leading to substantially improved captioning efficiency. More specifically, the PANNs network [11] is used with SimPF [83] for audio feature extraction, while Inception-v3 CNN architecture [91] with down-sampling is utilized for visual feature extraction [58].

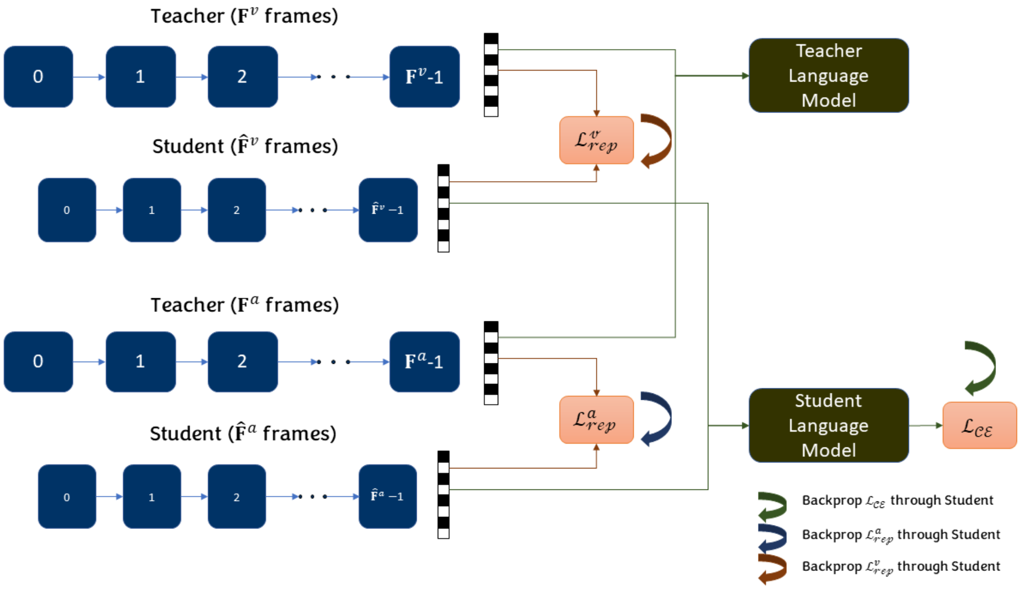

Teacher ($\mathbf{F}^v$ frames)
0
1
2
$\mathbf{F}^v$-1
Student ($\hat{\mathbf{F}}^v$ frames)
0
1
2
$\hat{\mathbf{F}}^v$ −1
Teacher ($\mathbf{F}^a$ frames)
0
1
2
$\mathbf{F}^a$-1
Student ($\hat{\mathbf{F}}^a$ frames)
0
1
2
$\hat{\mathbf{F}}^a$ −1
$\mathcal{L}^v_{rep}$
$\mathcal{L}^a_{rep}$
Teacher Language Model
Student Language Model
$\mathcal{L}_{\mathcal{CE}}$
Backprop $\mathcal{L}_{\mathcal{CE}}$ through Student
Backprop $\mathcal{L}^a_{rep}$ through Student
Backprop $\mathcal{L}^v_{rep}$ through Student

Figure 5.1 Proposed knowledge distillation based approach for audio-visual video captioning

The language model uses simple stacked GRUs [59] with dropouts [4] and residual connections [27, 36]. The student network is first trained and fine-tuned with the cross-entropy loss. To further improve the captioning accuracy, the representation loss is also used along with the cross-entropy loss. The experiments show that knowledge distillation can speed up audio-visual feature extraction with a negligible drop in captioning accuracy.

5.2. Proposed Knowledge Distillation based Audio-Visual Video Captioning Approach

This section presents the proposed video captioning approach based on the teacher-student model, as illustrated in Figure 5.1.

In video captioning, a sequence of words needs to be predicted from a vocabulary using audio and visual attributes.

The teacher network utilizes N^a audio frames $F^a = \left(F^a_0, F^a_1, \dots F^a_{N_a-1}\right)$ and N^v video frames $F^v = \left(F^v_0, F^v_1, \dots F^v_{N_v-1}\right)$ of the video **V** to predict a caption which can be stated using a neural network *n*:

$$P\left(\hat{\mathbf{Y}}\middle|V\right\} = n(F^a, F^v) \tag{5.1}$$

where $\hat{Y}$ denotes a series of words as $(\widehat{y_0}, \widehat{y_1}, \dots \widehat{y_{N^c}})$ and N^c refers to the number of words in the caption.

We employ Inception-v3 CNN architecture pre-trained on the ImageNet dataset to extract features from visual frames. The architecture resizes the images to $3 \times 299 \times 299$, then the average pooling layer outputs a latent vector consisting of 2048 units. Similarly, audio features are extracted with PANNs CNN architecture containing 10 stacked CNN layers pre-trained on AudioSet. An RNN based network that utilizes audio and visual features from the Inception-v3 and PANNs is used as a language model to generate captions. We employ a mean operator and acquire latent vectors from time-series input, which describe audio and visual features. These latent vectors are concatenated and fed to the RNN based network consisting of embedding, GRUs, and linear layers. Moreover, residual connections and dropouts are applied between layers to maintain gradient flow from the lower to upper layers. The teacher network is trained with the cross-entropy loss denoted as $\mathcal{L}_{\mathcal{CE}}$.

The student network is similar to the teacher, where SimPF and down-sampling algorithms are employed to reduce the number of audio and visual frames by a compression rate in a video. Specifically, we use the spectral pooling method of SimPF, which computes the discrete Fourier transform (DFT) of the audio frames F^a and then crops the center with a bounding box with the shape of (S, qN^a) where S refers to the dimension of the spectral feature to get $\widetilde{F^a_{crop}}$.

Then the output of the inverse discrete Fourier transform (IDFT) $\widehat{F^a}$ is taken as the compressed audio, as shown below,

$$\begin{aligned} \widetilde{F^a} &= DFT(F^a) \\ \widetilde{F^a_{crop}} &= F^a(S, qN^a) \\ \widehat{F^a} &= IDFT\left(\widetilde{F^a_{crop}}\right). \end{aligned} \tag{5.2}$$

Down-sampling is performed on F^v to obtain compressed visual frames $\widehat{F^v}$,

$$\widehat{F^v} = F^v(m/q), m = 0,1,2, \dots, N_v - 1 \tag{5.3}$$

where q denotes the compression rate, ranging from 0 to 1.

We extract audio and visual features from compressed frames using PANNs and Inception-v3. Then, latent vectors are acquired with a mean operator. We employ knowledge distillation from the teacher network to increase the accuracy of caption generation. A neural network with two hidden layers is utilized to increase the resemblance of latent vectors to the teacher. The network is trained to minimize the L1 loss between student and teacher latent vectors. We denote this loss as $\mathcal{L}_{rep}$ where rep refers to representation. We train the teacher network, and then the teacher guides the optimization of the parameters of the student network. In this study, we train the student-teacher network with the following losses:

$\mathcal{L}_{rep}$: The student network is only trained by the $\mathcal{L}_{rep}$ loss and is learned to mimic the audio-visual features of the teacher network. Then, the language model is trained with the updated neural network.

$\mathcal{L}_{rep} + \mathcal{L}_{\mathcal{CE}}$: We employ both $\mathcal{L}_{rep}$ and $\mathcal{L}_{\mathcal{CE}}$ losses to minimize the representation loss and maximize the captioning accuracy.

5.3. Experimental Evaluations

The proposed approach is evaluated on the MSR-VTT dataset [23], which initially consists of 10,000 videos, each with 20 ground truth captions. However, by the time the experiments are executed, only 5,074 and 2,123 videos are available from the training and testing sets, respectively. Several performance metrics are employed to measure the accuracy of the video captioning approach, including METEOR, BLEU, CIDEr, and ROUGE-L, and SPICE.

The ranking of the results is based on a final SCORE which is calculated as an average of all performance metrics. In calculating the final SCORE, we used the mean of the BLEU scores. For the experiments, the visual frames of the videos are resized into the shape of $3 \times 299 \times 299$. We utilized tokenization and punctuation removal on the ground truth captions of the training set. The latent vector size of the layers in the language models is set to 2,576, and the dimension of the linear layer output is equal to the vocabulary length. We evaluated the proposed approach with $0.2, 0.4, 0.6,$ and 0.8 compression ratios. The accuracy and time consumption of the teacher and student

Table 5.1 Performance metric evaluation results on the MSR-VTT test set

	BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**	**CIDEr**	**METEOR**	**ROUGE-L**	**SPICE**	**SCORE**	**Diff (%)**
Student (q=0.2) $\mathcal{L}_{rep}$	0.722	0.555	0.422	0.311	0.267	**0.236**	**0.554**	0.045	0.321	0.127
Student (q=0.4) $\mathcal{L}_{rep}$	0.715	0.546	0.411	0.294	0.223	0.234	0.539	0.043	0.306	0.168
Student (q=0.6) $\mathcal{L}_{rep}$	0.709	**0.542**	**0.412**	**0.300**	**0.232**	0.231	0.543	**0.041**	**0.308**	0.163
Student (q=0.8) $\mathcal{L}_{rep}$	0.719	0.550	0.413	0.300	0.256	0.235	0.545	0.046	0.315	0.144
Student (q=0.2) $\mathcal{L}_{rep} + \mathcal{L}_{\mathcal{CE}}$	**0.766**	**0.613**	**0.476**	**0.357**	**0.375**	**0.256**	**0.585**	**0.054**	**0.365**	**0.008**
Student (q=0.4) $\mathcal{L}_{rep} + \mathcal{L}_{\mathcal{CE}}$	0.774	0.618	0.473	0.348	0.359	0.256	0.582	0.055	0.361	0.019
Student (q=0.6) $\mathcal{L}_{rep} + \mathcal{L}_{\mathcal{CE}}$	0.769	0.616	0.478	0.357	0.375	0.258	0.586	0.055	0.366	0.005
Student (q=0.8) $\mathcal{L}_{rep} + \mathcal{L}_{\mathcal{CE}}$	0.765	0.614	0.479	0.358	0.366	0.255	0.583	0.054	0.362	0.016
Teacher	0.760	0.612	0.473	0.352	0.397	0.254	0.583	0.054	0.368	0.000

networks are measured with the test set of the MSR-VTT dataset under the $\mathcal{L}_{rep}$, and $\mathcal{L}_{rep} + \mathcal{L}_{\mathcal{CE}}$ losses.

In the evaluations, we compressed the frames on the student networks to enable faster inference time. The results for the students and teacher networks are given in Table 5.1, while time consumptions are shown in Table 5.2.

Using only the $\mathcal{L}_{rep}$ loss resulted in poor captioning performance in all performance metrics regarding the teacher network, as seen in Table 5.1. Notably, among the student networks trained with the $\mathcal{L}_{rep}$ loss, the compression rate of 0.2 has achieved the highest final SCORE. However, the combination of the $\mathcal{L}_{rep}$ and $\mathcal{L}_{\mathcal{CE}}$ losses in the student networks offered an accuracy approaching the level of the teacher model across all performance metrics. The captioning accuracy of the student network is increased from 0.321 to 0.365 with $\mathcal{L}_{rep} + \mathcal{L}_{\mathcal{CE}}$ under the compression rate of 0.2. The difference between the accuracy of the teacher and student network dropped from 0.127% to 0.008%. However, the student network with a 0.4 compression rate leveraged the final SCORE from 0.306 to 0.361, which is still lower than that of the compression rate at 0.2. We achieved the highest final SCORE at 0.366 using the student network with a compression rate of 0.6. This is followed by the compressed student network with a compression rate of 0.8, with a final SCORE of 0.362.

Table 5.2 Time consumption evaluation results on random 100 videos from the MSR-VTT test set

Network	average time consumption (s)	Diff (%)
Student (q=0.2)	2.77	79.1
Student (q=0.4)	5.65	57.4
Student (q=0.6)	8.31	37.4
Student (q=0.8)	11.03	16.9
Teacher	13.28	0.0

Furthermore, the student networks with compressed audio and visual frames scored higher across some metrics than the teacher. This indicates that student networks can generate accurate captions similar to the teacher. In Table 5.1, we present the time consumption of feature extraction for both audio and visual frames from randomly selected 100 videos from the test set of the MSR-VTT dataset. The compression rate 0.8 reduces feature extraction time up to 16.9%, while 0.6 compression rate decreases the audio-visual feature extraction time by about 37.4%. Similarly, 0.4 and 0.2 have reduced the inference time by 57.4% and 79.1%, respectively. Table 5.2 shows that the student networks reduce inference time significantly compared to the teacher network.

Chapter 6

Conclusions and Future Research

In this thesis, we delved into the dynamic field of image and video captioning, exploring various innovative techniques. Our focus was on advanced neural networks, attention mechanisms, and efficient down-sampling methods to enhance captioning accuracy and speed. This research spanned diverse strategies, contributing significantly to the evolution of the field. As we conclude, it is crucial to reflect on our key findings and contributions. This final chapter synthesizes our insights, highlighting major advancements and their implications for future research. We aim to show how our work contributes to the image and video captioning.

6.1. Multi-layer Gated Recurrent Unit based Recurrent Neural Network for Image Captioning

Encoder-decoder frameworks often encounter difficulties in efficiently extracting and utilizing contextual information from encoded data, causing inadequate performance in caption generation. To address these issues, in this paper, we have introduced a novel image captioning approach utilizing the NASNet-Large CNN encoder and a multi-layer GRU based decoder under the init-inject architecture. This modification substantially enhances the ability of the decoder to modulate the relevant information flow within the unit, thereby addressing the long-standing issue of RNN decoders challenged by managing long-term complex dependencies. The outcome is an improved decoder capable of producing semantically consistent and contextually accurate captions. Experimental results obtained from comprehensive evaluations in the MSCOCO dataset validate the effectiveness of our approach. Regarding the different CNN based encoders considered in the image captioning system, NASNet-Large architecture outperforms the other compared architectures in terms of seven performance metrics (i.e., BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L, METEOR, and CIDEr). The empirical analysis indicates that multi-layer GRU based decoders can yield higher performance compared to single-layer. The performance improvements can be achieved as the number of layers increases up to 9 layers. However, there is a subtle trend of a decrease with 12 and 15 layers. This system was developed to respond to significant challenges in the image captioning field, particularly in generating semantically consistent and grammatically accurate captions. Our future work will focus on the implementation of attention mechanisms to enhance caption generation by prioritizing key parts of the input image.

6.2. Leveraging Pre-trained 3D-CNNs for Video Captioning

In this study, a video captioning approach has been developed under the encoder-decoder based sequence-to-sequence approach. Different 2D and 3D-CNN architectures were used to extract the features of the video frames, and a multi-layer GRU was used to process the features and generate the video caption. The evaluations in the MSVD dataset show that the proposed approach improves the accuracy of 3D-CNN architectures in generating meaningful captions. We plan to explore ensembles of 3D-CNN architectures in our future study. Additionally, an evaluation of the feature extraction and representation capabilities of these architectures will be conducted to provide insights into their strengths and weaknesses.

6.3. Knowledge Distillation for Efficient Audio-Visual Video Captioning

In this chapter, we have presented a simple pooling front-end and down-sampling method to reduce the number of audio and visual frames in a video for video captioning. Furthermore, we have proposed a teacher-student based-network to leverage the accuracy of caption generation with knowledge distillation. We used $\mathcal{L}_{\mathcal{rep}}$ representation and $\mathcal{L}_{\mathcal{CE}}$ cross-entropy loss for network training. The proposed approach is evaluated on the MSR-VTT dataset. Experimental results show that the proposed approach significantly reduces the inference time with a negligible drop in captioning accuracy.

6.4. Future Research

Integration of Transformers: Given the advancements in neural network architectures, the integration of transformer models in both image and video captioning presents a promising avenue. Transformers, known for their effectiveness in handling sequential data and capturing long-range dependencies, could significantly elevate captioning accuracy. Future research could focus on customizing transformer architectures to better suit the specific nuances of image and video data, potentially enhancing the semantic and contextual understanding in caption generation.

Neural Architecture Search (NAS): The application of NAS could be a pivotal step in optimizing the neural network architectures used in captioning tasks. NAS automates the process of architectural engineering, potentially discovering novel and more efficient network structures that outperform manually designed models. Implementing NAS could lead to the development of optimized models for both image and video captioning, providing a balance between computational efficiency and captioning accuracy.

Advanced Attention Mechanisms: Building upon the success of attention mechanisms like Bahdanau and Transformer based attentions, future research can delve into more advanced attention models. These could include multi-head attention or spatial-temporal attention mechanisms, particularly for video captioning, to better capture the dynamics and intricacies of video data.

Multimodal Learning: Exploring multimodal learning strategies, especially in the context of video captioning, can enhance the system's ability to understand and integrate different types of data (e.g., audio, visual, and textual). This approach can lead to a more comprehensive and accurate representation of the content, improving the quality of generated captions.

Knowledge Distillation and Transfer Learning: Further research into knowledge distillation and transfer learning can be beneficial. This includes exploring how efficiently knowledge can be transferred from complex models (teachers) to simpler, more efficient models (students) without significant loss in performance, and how these methods can be adapted to different types of captioning tasks.

Exploration of Loss Functions: Investigating the impact of various loss functions on the performance of captioning models can provide insights into more effective training strategies. This includes experimenting with different combinations of loss functions like cross-entropy, representation loss, and others, to fine-tune the balance between the accuracy of captions and the computational efficiency of the models.

Real-world Application and Deployment: Practical application and deployment of developed models, especially in assistive technologies for the visually impaired, is a crucial area for future research. This includes refining the models for real-world scenarios, ensuring they are robust, user-friendly, and adaptable to various environments and user needs.

In summary, the future research of this thesis will be geared towards leveraging advanced neural network architectures and techniques to further enhance the accuracy and efficiency of captioning systems. The integration of these cutting-edge technologies holds the promise of developing more sophisticated, accurate, and user-friendly captioning tools for a wide array of applications.

References

[1] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. J. A. C. S. Laga, "A comprehensive survey of deep learning for image captioning," vol. 51, no. 6, pp. 1-36, 2019.

[2] J. Devlin *et al.*, "Language models for image captioning: The quirks and what works," 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[4] S. Wager, S. Wang, and P. S. J. A. i. n. i. p. s. Liang, "Dropout training as adaptive regularization," vol. 26, 2013.

[5] Z. Guo, L. Gao, J. Song, X. Xu, J. Shao, and H. T. Shen, "Attention-based LSTM with semantic consistency for videos captioning," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 357-361.

[6] Q. Wang and A. B. J. a. p. a. Chan, "Cnn+ cnn: Convolutional decoders for image captioning," 2018.

[7] A. Singh and M. D. J. a. p. a. Plumbley, "Low-complexity CNNs for acoustic scene classification," 2022.

[8] S. Targ, D. Almeida, and K. J. a. p. a. Lyman, "Resnet in resnet: Generalizing residual architectures," 2016.

[9] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251-1258.

[10] Y. Koizumi, Y. Ohishi, D. Niizumi, D. Takeuchi, and M. J. a. p. a. Yasuda, "Audio captioning using pre-trained large-scale language model guided by audio-based similar caption retrieval," 2020.

[11] Q. Kong *et al.*, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," vol. 28, pp. 2880-2894, 2020.

[12] L. Zhang, S. Wang, B. J. W. I. R. D. M. Liu, and K. Discovery, "Deep learning for sentiment analysis: A survey," vol. 8, no. 4, p. e1253, 2018.

[13] Ş. A. Akosman, M. Öktem, Ö. T. Moral, and V. Kılıç, "Deep Learning-based Semantic Segmentation for Crack Detection on Marbles," in *2021 29th Signal Processing and Communications Applications Conference (SIU)*, 2021, pp. 1-4: IEEE.

[14] J. Chung, C. Gulcehre, K. Cho, and Y. J. a. p. a. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014.

[15] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International conference on machine learning*, 2013, pp. 1310-1318: Pmlr.

[16] M. Tanti, A. Gatt, and K. P. J. N. L. E. Camilleri, "Where to put the image in an image caption generator," vol. 24, no. 3, pp. 467-489, 2018.

[17] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311-318.

[18] S. Banerjee and A. J. P. o. A.-W. Lavie, "Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments," pp. 65-72.

[19] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74-81.

[20] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, 2016, pp. 382-398: Springer.

[21] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566-4575.

[22] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 2014, pp. 740-755: Springer.

[23] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5288-5296.

[24] B. Fetiler, Ö. ÇAYLI, Ö. T. Moral, V. KILIÇ, and O. J. A. B. v. T. D. Aytuğ, "Video captioning based on multi-layer gated recurrent unit for smartphones," no. 32, pp. 221-226, 2021.

[25] R. Keskin, Ö. ÇAYLI, Ö. T. MORAL, V. KILIÇ, and O. J. A. B. v. T. D. Aytuğ, "A benchmark for feature-injection architectures in image captioning," no. 31, pp. 461-468, 2021.

[26] B. Makav and V. Kılıç, "A new image captioning approach for visually impaired people," in *2019 11th International Conference on Electrical and Electronics Engineering (ELECO)*, 2019, pp. 945-949: IEEE.

[27] S. AYDIN, Ö. ÇAYLI, V. KILIÇ, and O. J. A. B. v. T. D. Aytuğ, "Sequence-to-sequence video captioning with residual connected gated recurrent units," no. 35, pp. 380-386, 2022.

[28] U. Betül, Ö. ÇAYLI, V. KILIÇ, and O. J. A. B. v. T. D. Aytuğ, "Resnet based deep gated recurrent unit for image captioning on smartphone," no. 35, pp. 610-615, 2022.

[29] B. Makav and V. Kılıç, "Smartphone-based image captioning for visually and hearing impaired," in *2019 11th international conference on electrical and electronics engineering (ELECO)*, 2019, pp. 950-953: IEEE.

[30] R. Kiros, R. Salakhutdinov, and R. S. J. a. p. a. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," 2014.

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818-2826.

[32] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697-8710.

[33] L. S.-T. J. N. c. Memory, "Long short-term memory," vol. 9, no. 8, pp. 1735-1780, 2010.

[34] A. Mathews, L. Xie, and X. He, "Semstyle: Learning to generate stylised image captions using unaligned text," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8591-8600.

[35] Q. You, H. Jin, and J. J. a. p. a. Luo, "Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions," 2018.

[36] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128-3137.

[37] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. J. a. p. a. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," 2014.

[38] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. J. a. p. a. Murphy, "Optimization of image description metrics using policy gradient methods," vol. 5, 2016.

[39] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156-3164.

[40] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625-2634.

[41] M. Baran, Ö. T. Moral, and V. J. A. B. v. T. D. KILIÇ, "Akıllı telefonlar için birleştirme modeli tabanlı görüntü altyazılama," no. 26, pp. 191-196, 2021.

[42] H. Wang, H. Wang, and K. J. N. Xu, "Evolutionary recurrent neural network for image captioning," vol. 401, pp. 249-256, 2020.

[43] R. Keskin, Ö. T. Moral, V. Kılıç, and A. Onan, "Multi-gru based automated image captioning for smartphones," in *2021 29th Signal Processing and Communications Applications Conference (SIU)*, 2021, pp. 1-4: IEEE.

[44] V. J. S. U. J. o. C. KILIÇ and I. Sciences, "Deep gated recurrent unit for smartphone-based image captioning," vol. 4, no. 2, pp. 181-191, 2021.

[45] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, "Stylenet: Generating attractive visual captions with styles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3137-3146.

[46] K. Umemura *et al.*, "Tell as you imagine: Sentence imageability-aware image captioning," in *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27*, 2021, pp. 62-73: Springer.

[47] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2407-2415.

[48] Y. H. Tan and C. S. J. N. Chan, "Phrase-based image caption generator with hierarchical LSTM network," vol. 333, pp. 86-100, 2019.

[49] M. Heidari, M. Ghatee, A. Nickabadi, A. J. C. Pourhasan Nezhad, C. Practice, and Experience, "Diverse and styled image captioning using singular value decomposition-based mixture of recurrent experts," vol. 34, no. 22, p. e6866, 2022.

[50] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048-2057: PMLR.

[51] D. W. Otter, J. R. Medina, J. K. J. I. t. o. n. n. Kalita, and l. systems, "A survey of the usages of deep learning for natural language processing," vol. 32, no. 2, pp. 604-624, 2020.

[52] R. Staniūtė and D. J. A. S. Šešok, "A systematic literature review on image captioning," vol. 9, no. 10, p. 2024, 2019.

[53] C. Gan, T. Yao, K. Yang, Y. Yang, and T. Mei, "You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 923-932.

[54] F. Shen, C. Shen, Q. Shi, A. Van Den Hengel, and Z. Tang, "Inductive hashing on manifolds," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1562-1569.

[55] Ö. Çayli, X. Liu, V. Kiliç, and W. Wang, "Knowledge Distillation for Efficient Audio-Visual Video Captioning," in *2023 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 745-749: IEEE.

[56] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. J. a. p. a. Saenko, "Translating videos to natural language using deep recurrent neural networks," 2014.

[57] S. Guadarrama *et al.*, "Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2712-2719.

[58] Ö. Çaylı, B. Makav, V. Kılıç, and A. Onan, "Mobile application based automatic caption generation for visually impaired," in *Intelligent and Fuzzy Techniques: Smart and Innovative Solutions: Proceedings of the INFUS 2020 Conference, Istanbul, Turkey, July 21-23, 2020*, 2021, pp. 1532-1539: Springer.

[59] Ö. Çaylı, V. Kılıç, A. Onan, and W. Wang, "Auxiliary classifier based residual rnn for image captioning," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 1126-1130: IEEE.

[60] M. Amaresh and S. Chitrakala, "Video captioning using deep learning: an overview of methods, datasets and metrics," in *2019 International Conference on Communication and Signal Processing (ICCSP)*, 2019, pp. 0656-0661: IEEE.

[61] Ö. T. Moral, V. Kiliç, A. Onan, and W. Wang, "Automated Image Captioning with Multi-layer Gated Recurrent Unit," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 1160-1164: IEEE.

[62] Ö. B. Mercan and V. Kılıç, "Deep learning based colorimetric classification of glucose with au-ag nanoparticles using smartphone," in *2020 Medical Technologies Congress (TIPTEKNO)*, 2020, pp. 1-4: IEEE.

[63] Ö. B. Mercan, V. Doğan, and V. Kılıç, "Time Series Analysis based Machine Learning Classification for Blood Sugar Levels," in *2020 Medical Technologies Congress (TIPTEKNO)*, 2020, pp. 1-4: IEEE.

[64] V. Kılıç, X. Zhong, M. Barnard, W. Wang, and J. Kittler, "Audio-visual tracking of a variable number of speakers with a random finite set approach," in *17th International Conference on Information Fusion (FUSION)*, 2014, pp. 1-7: IEEE.

[65] V. Doğan, T. Isık, V. Kılıç, and N. J. A. M. Horzum, "A field-deployable water quality monitoring with machine learning-based smartphone colorimetry," vol. 14, no. 35, pp. 3458-3466, 2022.

[66] B. Sayraci, M. AĞRALI, and V. J. A. B. v. T. D. KILIÇ, "Artificial Intelligence Based Instance-Aware Semantic Lobe Segmentation on Chest Computed Tomography Images," no. 46, pp. 109-115, 2023.

[67] V. Doğan, M. Evliya, L. N. Kahyaoglu, and V. J. T. Kılıç, "On-site colorimetric food spoilage monitoring with smartphone embedded machine learning," vol. 266, p. 125021, 2024.

[68] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang, "Hierarchical recurrent neural encoder for video representation with application to captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1029-1038.

[69] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4534-4542.

[70] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4584-4593.

[71] L. Baraldi, C. Grana, and R. Cucchiara, "Hierarchical boundary-aware neural encoder for video captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1657-1666.

[72] L. Yao *et al.*, "Describing videos by exploiting temporal structure," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4507-4515.

[73] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4594-4602.

[74] J. Sun, X. Liu, X. Mei, M. D. Plumbley, V. Kilic, and W. J. a. p. a. Wang, "Automated audio captioning via fusion of low-and high-dimensional features," 2022.

[75] X. Mei, X. Liu, M. D. Plumbley, W. J. E. j. o. a. Wang, speech,, and m. processing, "Automated audio captioning: an overview of recent progress and new challenges," vol. 2022, no. 1, pp. 1-18, 2022.

[76] X. Liu *et al.*, "Visually-aware audio captioning with adaptive audio-visual attention," 2022.

[77] A. G. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017.

[78] A. Singh and M. D. J. a. p. a. Plumbley, "A passive similarity based CNN filter pruning for efficient acoustic scene classification," 2022.

[79] H. Futami, H. Inaguma, S. Ueno, M. Mimura, S. Sakai, and T. J. a. p. a. Kawahara, "Distilling the knowledge of BERT for sequence-to-sequence ASR," 2020.

[80] L. Lu, M. Guo, and S. Renals, "Knowledge distillation for small-footprint highway networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4820-4824: IEEE.

[81] K. Choi, M. Kersner, J. Morton, and B. Chang, "Temporal knowledge distillation for on-device audio classification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 486-490: IEEE.

[82] J. Liang, T. Zhang, and G. J. I. A. Feng, "Channel compression: Rethinking information redundancy among channels in CNN architecture," vol. 8, pp. 147265-147274, 2020.

[83] X. Liu, H. Liu, Q. Kong, X. Mei, M. D. Plumbley, and W. Wang, "Simple pooling front-ends for efficient audio classification," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1-5: IEEE.

[84] J. F. Gemmeke *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2017, pp. 776-780: IEEE.

[85] X. Liu *et al.*, "Leveraging pre-trained BERT for audio captioning," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 1145-1149: IEEE.

[86] X. Mei, X. Liu, J. Sun, M. D. Plumbley, and W. Wang, "Diverse audio captioning via adversarial training," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8882-8886: IEEE.

[87] X. Mei *et al.*, "An encoder-decoder based audio captioning system with transfer and reinforcement learning," 2021.

[88] P. Ostyakov *et al.*, "Label denoising with large ensembles of heterogeneous neural networks," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0-0.

[89] R. Lin, J. Xiao, and J. Fan, "Nextvlad: An efficient neural network to aggregate frame-level features for large-scale video classification," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0-0.

[90] S. Bhardwaj, M. Srinivasan, and M. M. Khapra, "Efficient video classification using fewer frames," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 354-363.

[91] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, 2015, pp. 448-456: pmlr.

Appendices

Publications from the Thesis

SCIE Journal Articles

1. **Ö Çaylı**, V Kılıç, A Onan, W Wang. Multi-layer Gated Recurrent Unit based Recurrent Neural Network for Image Captioning. In *International Journal of Pattern Recognition and Artificial Intelligence. (under review)*

National Journal Articles

1. B Fetiler, **Ö Çaylı**, V Kılıç. Leveraging Pre-trained 3D-CNNs for Video Captioning. In *European Journal of Science and Technology. (Accepted)*

2. M Kılcı, **Ö Çaylı**, V Kılıç (2023). Fusion of High-Level Visual Attributes for Image Captioning. In *European Journal of Science and Technology, (52), 161-168.*

3. B Uslu, **Ö Çaylı**, V Kılıç, A Onan (2022). Resnet based Deep Gated Recurrent Unit for Image Captioning on Smartphone. In *European Journal of Science and Technology, (35) 610-615.*

4. S Aydın, **Ö Çaylı**, V Kılıç, A Onan (2022). Sequence-to-Sequence Video Captioning with Residual Connected Gated Recurrent Units. In European Journal of Science and Technology, (35), 380-386.

5. B Fetiler, **Ö Çaylı**, ÖT Moral, V Kılıç, A Onan (2021). Video Captioning Based on Multi-layer Gated Recurrent Unit for Smartphones. In *European Journal of Science and Technology, (32), 221-226.*

6. R Keskin, **Ö Çaylı**, ÖT Moral, V Kılıç, A Onan (2021). A Benchmark for Feature-injection Architectures in Image Captioning. In *European Journal of Science and Technology, (31), 461-468.*

Conference Papers

1. **Ö Çaylı**, X Liu, V Kılıç, W Wang. Knowledge Distillation for Efficient Audio-Visual Video Captioning. In *2023 Proceedings of the 31th European Signal Processing Conference.*

2. **Ö Çaylı**, V Kılıç, A Onan, W Wang. Auxiliary Classifier based Residual RNN for Image Captioning. In *2022 Proceedings of the 30th European Signal Processing Conference.*

3. E Çil, **Ö Çaylı**, V Kılıç. Bahdanau Attention based Residual LSTM for Video Captioning on Smartphones. In *2023 Proceedings of the 2nd International Conference on New Horizons in Science.*

4. A Şahan, **Ö Çaylı**, V Kılıç. Utilizing Bahdanau Attention-Based Residual LSTM for Image Captioning on Mobile Devices. In *2023 Proceedings of the 2nd International Conference on New Horizons in Science.*

5. **Ö Çaylı**, B Makav, V Kılıç, A Onan. Natural Language Description of Images Using a Smartphone Application. In *2020 Proceedings of the 2nd International Eurasian Conference on Science, Engineering and Technology*

6. **Ö Çaylı**, B Makav, V Kılıç, A Onan. Mobile application based automatic caption generation for visually impaired. In *2020 Proceedings of the 2nd Intelligent and Fuzzy Techniques: Smart and Innovative Solutions*