Please cite the Published Version

A transformer-based Urdu image caption generation

Muhammad Hadi[2] · Iqra Safder[1] · Hajra Waheed[1] · Farooq Zaman[2] · Naif Radi Aljohani[3] · Raheel Nawaz[4] · Saeed Ul Hassan[2,5] · Raheem Sarwar[6]

Abstract

Image caption generation has emerged as a remarkable development that bridges the gap between Natural Language Processing (NLP) and Computer Vision (CV). It lies at the intersection of these fields and presents unique challenges, particularly when dealing with low-resource languages such as Urdu. Limited research on basic Urdu language understanding necessitates further exploration in this domain. In this study, we propose three Seq2Seq-based architectures specifically tailored for Urdu image caption generation. Our approach involves leveraging transformer models to generate captions in Urdu, a significantly more challenging task than English. To facilitate the training and evaluation of our models, we created an Urdu-translated subset of the flickr8k dataset, which contains images featuring dogs in action accompanied by corresponding Urdu captions. Our designed models encompassed a deep learning-based approach, utilizing three different architectures: Convolutional Neural Network (CNN) + Long Short-term Memory (LSTM) with Soft attention employing word2Vec embeddings, CNN+Transformer, and Vit+Roberta models. Experimental results demonstrate that our proposed model outperforms existing state-of-the-art approaches, achieving 86 BLEU-1 and 90 BERT-F1 scores. The generated Urdu image captions exhibit syntactic, contextual, and semantic correctness. Our study highlights the inherent challenges associated with retraining models on low-resource languages. Our findings highlight the potential of pre-trained models for facilitating the development of NLP and CV applications in low-resource language settings.

Keywords Caption generation · Deep learning · Natural Language processing - NLP · Transformers · Urdu image caption generation

✉ Raheem Sarwar
R.Sarwar@mmu.ac.uk

Muhammad Hadi
msds20029@itu.edu.pk

Iqra Safder
iqra.safder@nu.edu.pk

Hajra Waheed
hajra.waheed@nu.edu.pk

Farooq Zaman
phdcs18002@itu.edu.pk

Naif Radi Aljohani
nraljohani@kau.edu.sa

Raheel Nawaz
raheel.nawaz@staffs.ac.uk

Saeed Ul Hassan
saeedulhassan@gmail.com

1   Department of Computer Science, National University of Computer & Emerging Sciences, Lahore, Pakistan

2   Department of Computer Science, Information Technology University, Lahore, Pakistan

3   Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

4   Staffordshire University, Stoke-on-Trent, UK

5   Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, UK

6   Faculty of Business and Law, Manchester Metropolitan University, Manchester, UK

1 Introduction

Image captioning. Image caption generation involves comprehending the visual scenes captured in an image and the subsequent generation of the correct sentences that describe them. Image captioning is a technology that combines computer vision (CV) and natural language processing (NLP) to generate descriptive and contextually relevant textual descriptions for images. The system takes an image as its input. Computer vision techniques are applied to extract meaningful features from the image. These features represent objects, scenes, or other relevant information within the image. NLP models generate descriptive captions based on the extracted image features. The captioning model considers the context of the image to produce coherent and contextually relevant descriptions. This involves understanding relationships between objects, recognizing actions, and interpreting the scene. The generated captions are often evaluated using metrics such as BLEU (Bilingual Evaluation Understudy) (Antol et al. 2015).

Applications of image captioning. Image captioning makes visual content more accessible to individuals with visual impairments. It provides textual descriptions of images and helps them understand the content. Image captions are used in content-based image retrieval systems, enabling users to find images based on textual descriptions rather than keywords. Image captioning is widely used on social media platforms. It allows users to add context to their photos and helps index and search content. Search engines can use image captions to improve the accuracy of image searches. Users can input text to find images relevant to their query. On platforms where user-generated content is shared, image captions can be analyzed for content moderation, ensuring that inappropriate or harmful content is filtered out. In the medical field, image captioning can be applied to medical imaging to provide detailed descriptions of scans and images for better patient understanding and communication among healthcare professionals. Image captioning is used in e-learning and educational materials, providing explanations for visual content to enhance learning experiences. It aids in computer vision applications, such as self-driving cars, robotics, and drones, by providing descriptions of the environment. Image captioning assists in organizing personal photo collections by automatically generating descriptions for images (Goodrum 2000; Bouchard et al. 2008).

Challenges. Most existing studies on image captioning tasks are focused on resource-rich languages such as English, French, and Chinese. Although the applications of the Image captioning task are not limited to resource-rich languages (Karpathy et al. 2014; Li et al. 2022; Sharma et al. 2018; Li et al. 2020; Bakar et al. 2023; Saadany et al. 2023;

Hassan et al. 2023b), limited attention has been paid to low-resource languages such as Urdu.

Processing the Urdu language presents several challenges, primarily due to the language's unique characteristics and limited digital resources. (i) Urdu is considered a low-resource language in the context of NLP. Compared to languages like English, Spanish, or Chinese, there are fewer digital resources available, including text corpora, language models, and annotated datasets. This scarcity hinders the development of NLP tools and models for Urdu. (ii) Urdu is a morphologically rich language, meaning that words can change significantly based on tense, gender, number, and other linguistic factors. This complexity makes part-of-speech tagging, stemming, and lemmatization more challenging. (iii) Urdu uses the Perso-Arabic script, which is written from right to left and includes a rich set of ligatures and diacritics. (iv) Handling the script's complexity in text processing, including tokenization and character encoding, can be challenging. (v) There is a lack of standardized language resources and guidelines for Urdu. This makes it difficult to develop consistent language models and tools. (vi) Urdu speakers often engage in code-switching, which means mixing Urdu with other languages, such as English or Hindi, in their conversations. This complicates language understanding and modelling. (vii) Recognizing named entities in Urdu text is challenging due to the absence of well-annotated datasets and the diversity of names and entities used in Urdu-speaking regions. (viii) Like many languages, Urdu exhibits word and phrase ambiguity, where the same word or phrase can have multiple meanings based on the context. Resolving such ambiguities is challenging for NLP models. (ix) Languages evolve, and Urdu is no exception. Keeping NLP resources and models up-to-date with the evolving language is a challenge (Sarwar 2022; Sarwar and Hassan 2021; Limkonchotiwat et al. 2020; Safder et al. 2021; Mohammad et al. 2019; Khan et al. 2020; Sabah et al. 2023; Silva et al. 2023; Hassan et al. 2023a; Mohamed et al. 2023; Mohamed and Sarwar 2022).

Transformer models have revolutionized the field of natural language processing (NLP) due to their exceptional ability to capture long-range dependencies and generate coherent captions, among other tasks. Traditional NLP models, such as recurrent neural networks (RNNs), have limitations in capturing long-range dependencies due to their sequential nature. Transformers, however, use a self-attention mechanism that allows them to weigh the importance of different words in a sequence, regardless of their distance. This parallelism makes them highly efficient for sequence modelling tasks. Transformers can understand the nuances and context of natural language, making them exceptionally adept at tasks like text classification, and language translation. Their ability to consider the entire input sequence when

making predictions makes them superior in understanding textual data. Transformers rely on the attention mechanism to focus on relevant parts of the input sequence when producing an output. This attention mechanism allows them to generate coherent captions by considering the entire image or text and understanding the relationships between different elements in the sequence. In the context of image captioning, transformers have shown impressive results. They can effectively process both image and text data, making them well-suited for tasks that require understanding the visual content of images and generating descriptive captions that are coherent and contextually relevant. When generating captions, transformers can connect relevant visual features to textual content and understand how different elements in the image relate to the generated text. This enables them to capture long-range dependencies between image regions and words in the captions, resulting in more meaningful and coherent descriptions (Afzal et al. 2023).

Objectives and contributions of this study. The objectives of this study are threefold: to improve the quality of Urdu image captions, to investigate the impact of transformer-based models, and to leverage Urdu word embeddings.

The primary focus of this study is to develop an image caption generation model for Urdu. The research introduces advanced image captioning methods for generating captions in the Urdu language. To facilitate this, a subset of the original Flick8k dataset (Hodosh et al. 2013) is utilized and translated into Urdu. This dataset comprises images of dogs, each associated with five Urdu captions. To ensure unbiased results, the dataset is manually partitioned into training, testing, and validation sets, preventing captions from the same image from overlapping in both the training and testing phases.

Transformers, renowned for their prowess in various natural language processing (NLP) tasks as mentioned earlier, serve as the driving force behind this study. Building upon the impressive outcomes of transformer models (Vaswani et al. 2017), the research makes the following key contributions:

Three distinct deep learning models are proposed for the task of Urdu caption generation using the Dogs dataset (Afzal et al. 2023).

- The first model employs an attention-based mechanism, extending the groundwork laid by a previous study (Xu et al. 2015). This model incorporates specific enhancements, including the integration of a pre-trained word2vec embedding layer with 512 dimensions. This model attains a BLEU score of 71.8 and a Language-Agnostic Sentence Representations (LASER) score of 73.5.

- The second model proposes the use of a Convolutional Neural Network (CNN) as an encoder and a transformer-based model as a decoder. While transformers are often utilized in pre-trained forms and fine-tuned for downstream tasks, this study implements a custom transformer-based solution and trains the decoder from scratch. This model yields a BLEU score of 78.9 and a LASER score of 79.99.

- Lastly, the study introduces a vision transformer-based encoder and a Roberta-based decoder. The vision transformer is pre-trained on an extensive collection of images from ImageNet, while Roberta is trained on a substantial Urdu corpus, focusing on masked language modelling tasks. This model is fine-tuned using the Urdu image caption dataset (Afzal et al. 2023) and significantly surpasses the previous baseline model (Afzal et al. 2023; Ilahi et al. 2020). The research underscores the effectiveness of transformers in outperforming prior models in generating Urdu image captions. Moreover, it highlights the advantages of pretraining in enhancing model performance when fine-tuning for downstream tasks with limited data. This model achieves a BLEU score of 86.0 and a LASER score of 81.79.

Objectives and contributions of this study. The main objectives of this study are three-fold: enhance Urdu image caption quality, investigate the impact of transformer-based models, and leverage Urdu word embeddings. This study focuses on preparing a URDU-based automated image caption generation model and proposes state-of-the-art image captioning methods for Urdu image caption generation. This study utilized subset (Afzal et al. 2023) from the original Flick8k dataset (Hodosh et al. 2013). This subset is translated into URDU and validated. The dataset consists of dog images and 5 URDU captions against each image. The total dataset is manually separated into train, test, and validation splits so that no captions from the same image overlap in training and testing giving biased results. As discussed earlier, transformers have shown the best results for several NLP tasks. Based on the admirable results reported by the transformers model (Vaswani et al. 2017), the following are the main contributions of this research.

This research introduces three Deep Learning models designed for the task of generating captions for Urdu images, which were trained using the Dogs dataset (Afzal et al. 2023).

- The first model employs an attention-based mechanism, building upon an existing study (Xu et al. 2015) with specific modifications, including the incorporation of a pre-trained word2vec embedding layer with 512 dimensions. This model achieved a BLEU score of 71.8 and a

Language-Agnostic Sentence Representations (LASER) score of 73.5.

- The second model proposes the use of a Convolutional Neural Network (CNN) as an encoder and a transformer-based model as a decoder. While transformers are commonly utilized in pre-trained forms and fine-tuned for downstream tasks, this study employs a custom transformer-based implementation, training the decoder from scratch. With this model, the research reports a BLEU score of 78.9 and a LASER score of 79.99.

- Lastly, the study introduces a vision transformer-based encoder and a Roberta-based decoder. The vision transformer is pre-trained on a substantial collection of images from ImageNet, while Roberta is trained on a vast Urdu corpus for masked language modelling tasks. This model is fine-tuned using the Urdu image caption dataset (Afzal et al. 2023) and demonstrates superior performance, surpassing the previous baseline model (Afzal et al. 2023; Ilahi et al. 2020). This highlights the effectiveness of transformers in outperforming prior models for generating Urdu image captions by a significant margin. The research also highlights the advantages of pretraining in enhancing model performance when fine-tuning for downstream tasks with limited data. With this model, the research achieved a BLEU score of 86.0 and a LASER score of 81.79.

2 Related work

The problem of generating image captions was initially addressed by traditional machine learning techniques that relied on extracting hand-crafted features from images, such as Local Binary Patterns (LBP) (Ojala et al. 2000). Scale-Invariant Feature Transform (SIFT) (Lowe 1999) was then introduced, followed by the Histogram of Oriented Gradients (HOG) (Dalal and Triggs 2005), that proposed classification using machine learning classifiers such as Support Vector Machines (SVM). I2T: Image Parsing to Text Description introduced image parsing framework (Yao et al. 2010). This framework converted images to parse graphs, using and-or graphs, (AOG), into visual knowledge representation created using edge detection, SIFT, and histograms passed to the text generation engine.

A proposed framework called Im2Text, which was both simple and efficient, addressed the image caption generation problem (Ordonez et al. 2011). In their work, the authors employed object detection combined with TF-IDF. Subsequently, deep learning methods emerged as a significant breakthrough for this problem, offering a wide array of solutions (Afzal et al. 2023). These solutions encompassed various approaches, including a) template-based image captioning, b) retrieval-based methods, c) dense image captioning, d) whole scene-based image captioning, e) encoder-decoder-based models, and f) attention-based techniques.

Another study introduced a multi-modal approach for representing both text and image (Kiros et al. 2014). This approach involved jointly learning image features alongside textual descriptions, eliminating the requirement for image pyramids. Additionally, another multi-modal method (Karpathy et al. 2014) was proposed, which utilized visual regions from an image, aligned and transformed them into visual embeddings, and employed RNNs for text generation. However, both approaches faced significant challenges such as large pre-trained word embeddings and RNN long-term vanishing gradient issues.

An image captioning approach presented in Chen and Lawrence Zitnick (2015) was multi-modal, but instead of separate embeddings, the authors utilized a combined multi-modal embedding of the image. Another notable contribution to the research community was the introduction of the "Learning like a Child" model, which was one of the earliest few-shot learning approaches for image caption generation (Mao et al. 2015). Additionally, a Generative Adversarial Network (GAN)-based image-caption generator was proposed in Dai et al. (2017), where the model was trained with a single caption per image during training. In a similar vein, another captioning method was presented in a study that used multiple captions per image in their training data, leading to improved results (Shetty et al. 2017)

A scene graph-based approach for image caption generation used a scene graph generation module for images and attention-based LSTM that takes encoded sub-graphs and generates caption (Zhong et al. 2020). The authors reported a BLEU4 score of 35 on the MSCOCO dataset. Attention-based image captioning was first introduced in Xu et al. (2015). The authors used hard and soft attention for image captioning using a CNN feature extractor with an attention layer that was responsible for mapping information from hidden layers of CNN to LSTM during generation. At each time step, the LSTM model directs its attention to the specific features that contribute to the generation of a particular word.

An image caption generator on fashion items was proposed in Yang et al. (2020), to propose a method that can learn the rich attributes of an image using reward-based learning. Moreover, many other methods were also introduced, such as semi-supervised caption generation (Chen et al. 2016), and self-supervise methods (Herdade et al. 2019) that significantly improved the state-of-the-art techniques for image caption generation and have opened up new avenues for research in this domain.

The minds behind "Attention is All You Need" presented a novel approach called Transformers (Vaswani et al. 2017). This architecture defined a method called multi-head attention and used feed-forward layers. Transformers provided a huge boost to natural language processing and computer vision problems (Vaswani et al. 2017). Subsequently, CPTR a fully attention-based image caption generation model was also proposed (Herdade et al. 2019). In this study, the authors proposed an Object Relation Transformer architecture that maps spatial relations between input and attention-based weights for objects. Another study proposed VisualBERT, a simple framework for vision-text tasks (Li et al. 2019). The authors proposed a pre-training technique of BERT-based architecture that can be further fine-tuned for several vision-text tasks like captioning, visual question answering (VQA), visual common-sense reasoning (VCR) (Sap et al. 2020).

A model called OSCAR (Li et al. 2020), demonstrates that pre-training a model and fine-tuning it for domain-specific tasks significantly enhances performance in vision-text tasks. The proposed architecture leverages object detection tags to establish semantic alignments between images and text, leading to state-of-the-art results on the MS-COCO dataset (Lin et al. 2014). Another study (Cornia et al. 2020) introduced meshed connections between the multi-layer encoder and decoder, improving the model's ability to generate coherent and contextual captions. Addressing the issue of forgetting previous tasks, the authors of Del Chiaro et al. (2020) proposed a solution called Recurrent Attention to Transient Tasks (RATT) based on continual learning. This method aimed to mitigate the problem by incorporating recurrent attention mechanisms. Furthermore, a bootstrapping mechanism for large-scale image-text pretraining, known as BLIP, was proposed (Li et al. 2022). Meanwhile, OFA (Wang et al. 2022), a sequence-to-sequence learning framework that jointly pretrains on multi-modal tasks, eliminating the need for finetuning. This approach offers a relatively simple model architecture with fewer parameters, making it adaptable to unseen tasks with ease.

Despite being spoken in multiple countries with some of the largest populations globally, the Urdu language has received limited attention in the literature, especially in the field of Urdu image captioning. Limited evidence can be observed in the existing literature, focusing on this area. Generally, data-driven solutions and methods rely on the availability of a well-curated dataset as an initial step. However, Urdu is a low-resource language, and limited and scarce resources are available to construct such datasets. Furthermore, as mentioned earlier in the Introduction, the processing of Urdu datasets presents a challenging task due to the lack of tools and resources dedicated to the language. In addition to the scarcity of published work and literature,

the absence of a large-scale dataset poses a significant challenge for Urdu caption generation. The existing datasets for Urdu are typically small in size, making it difficult to train state-of-the-art methods effectively. Furthermore, the available Urdu datasets for supervised learning often suffer from poor annotation quality, further hindering progress in this area.

A study proposed a resnet50-based encoder for feature extraction and a GRU-based decoder for text generation (Ilahi et al. 2020). The model utilized a soft attention layer within the hidden layers, which allowed for focusing on relevant information during the caption generation process. Additionally, the authors introduced a grammar correction layer at the end to ensure that the generated outputs were not only conceptually correct but also aligned accurately with the intended meaning.

In another study, encoder-decoder architecture with soft attention was used (Afzal et al. 2023). In this study, the authors used resnet101 with an LSTM-based decoder including soft attention. The authors further applied early stopping based on a BLEU score of the validation set, along with gradient clipping to avoid exploding gradients. The study reported improvements in the performance of a dataset that was introduced in the same study. However, our study proposes three Seq2Seq-based architectures specifically tailored for Urdu image caption generation that significantly outperform the existing solutions. Our models encompassed a deep learning-based approach, utilizing three different architectures: CNN+LSTM with Soft attention employing word2Vec embeddings, CNN+Transformer, and Vit+Roberta models. Experimental results demonstrate that our proposed model outperforms existing state-of-the-art approaches, achieving an 86 BLEU-1 score and a 90 BERT-F1 score. The generated Urdu image captions exhibit syntactic, contextual, and semantic correctness.

Another study introduced SmallCap, a model producing captions by integrating input images with related captions from a stored database which provides a lightweight, swift-to-train solution (Ramos et al. 2023). It relies on cross-attention layers between a CLIP encoder and GPT-2 decoder, learning minimal parameters. SmallCap exhibits domain transferability sans extra finetuning, utilizing diverse data in a training-free manner. Tests on COCO showcase competitive performance, extending to new domains through data retrieval. Leveraging human-labeled and web data enhances results across various domains, demonstrating efficacy in handling novel visual concepts, as seen in the nocaps benchmark.

Another study explores novel transformer-based techniques for image captioning (Dubey et al. 2023). It introduces two modules: Label Attention (LAM) focusing on objects and Geometrically Coherent Proposal (GCP) for

object scale and position. These enforce object relevance in the environment, enhancing image perception and language association. LAM links object classes to a dictionary via self-attention, while GCP ensures coherence using object geometry ratios. The proposed framework, LATGeO, generates captions by analyzing object relationships. Tested on MSCOCO, LATGeO proves the importance of objects' context and visual attributes, producing enhanced and meaningful captions by linking object features with their geometrically coherent representation and associated labels.

Another paper diverges from traditional Transformer-based approaches, introducing Semantic-Conditional Diffusion Networks (SCD-Net) tailored for captioning (Luo et al. 2023). Utilizing cross-modal retrieval, semantically relevant sentences inform a Diffusion Transformer's learning process, enhancing semantic richness. SCD-Net employs stacked Diffusion Transformer structures to bolster output

sentences progressively, refining visual-language alignment and linguistic coherence. A novel self-critical sequence training stabilizes the diffusion process, guided by an autoregressive Transformer model. COCO dataset experiments showcase SCD-Net's promising potential in addressing the intricacies of image captioning.

Moreover, another study evaluated the results of random data split. The employed splitting method resulted in overlapping captions, as the dataset used contained five captions for each image. This approach could lead to the same images appearing in both the train and test sets, which may artificially boost scores on the validation and test sets. However, it is important to note that such datasets are typically shared with pre-defined splits for the training, validation, and test sets, which ensures a more effective evaluation of models' generalization abilities. Afzal et al. (2023) used flicker8k data split and produced results on the test set. A summary of the main differences between image captioning research is given in Table 1.

2.1 Summary

Image captioning is a challenging task in artificial intelligence that involves generating textual descriptions for images. Over the years, various image captioning techniques have been developed, and they can be broadly categorized into traditional and transformer-based approaches. Here's an overview of both:

Traditional image captioning techniques.

Template-based approaches: These methods use predefined templates to create captions for images. They fill in the templates with relevant words or phrases based on the image content. While simple, they cannot generate creative and contextually rich captions.

Statistical language models: Traditional models like Hidden Markov Models (HMMs) and n-gram models have been used for image captioning. They rely on statistical patterns in language and image data to generate captions.

Feature engineering: These techniques extract handcrafted features from images, such as SIFT or HOG descriptors, and combine them with textual features for caption generation. However, they often struggle with complex scenes and lack the ability to understand high-level concepts.

Deep learning models: Convolutional Neural Networks (CNNs) for image feature extraction and Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks for text generation have been a popular combination. CNNs capture image features, and RNNs/LSTMs generate captions word by word. This sequential approach has limitations in handling long-range dependencies and generating diverse and coherent captions.

Transformer-based image captioning techniques.

Table 1 Main Differences in deep learning based image captioning research

Research	Method Used	Dataset	Limitations	Advantages
Aneja et al. (2018)	CNN-LSTM	MS COCO	Limited multilingual support	High-quality captions
			Difficulty with rare concepts	Robust object recognition
			Limited context modelling	Well-established benchmark dataset
He et al. (2020)	Transformer-based	ImageNet	Limited contextual awareness	Efficient parallel processing
			Lack of fine-grained details	Scalable architecture
			Requires large-scale pretraining	Strong performance on common tasks
Vaswani et al. (2017)	Attention Mechanisms	Flickr30k	Limited diversity in captions	Smoother attention and coherence
			Performance drops with rare data	Suitable for specific applications
			Not as competitive on benchmarks	Customizable attention mechanisms
Yan et al. (2020)	Reinforcement Learning with Policy Gradient	Visual Genome	Increased computational cost	Improved fluency in captions
			Challenges with reward shaping	Better adaptation to diverse data
			Long training time	Enhanced caption quality

Vision transformers (ViTs): Vision Transformers, inspired by the success of transformer models in NLP, have been adapted for image captioning. They treat images as sequences of patches and use transformer architectures like the original Transformer, BERT, or GPT for caption generation.

Alignment-based models: These models utilize cross-modal attention mechanisms to align image regions with words in captions. By establishing meaningful correspondences between visual and textual elements, they can generate more contextually relevant captions.

Pretrained models: Transformers pre-trained on large-scale datasets (e.g., ImageNet and COCO) have become popular for image captioning. These models benefit from the pretraining's generalization and are fine-tuned on specific captioning tasks.

Dense captioning: Instead of generating a single caption, dense captioning generates multiple captions for different regions or objects within an image. This is useful for describing all relevant details in a complex scene.

Refinement networks: Some transformer-based approaches use additional modules for caption refinement. These modules help improve the fluency and coherence of generated captions.

Multimodal models: Integrating vision and language understanding, these models often combine image and text encoders to extract rich representations from both modalities. They can capture complex relationships between visual and textual information.

Generative Adversarial Networks (GANs): GAN-based models can be used for image captioning by training a generator network to produce captions that discriminate against real and fake captions. This can improve the quality of generated captions.

Transformer-based approaches have shown significant advancements in image captioning due to their ability to capture long-range dependencies, handle multimodal data effectively, and leverage pre-trained models. These techniques have outperformed traditional methods and continue to drive progress in the field of image captioning.

3 Methodology

3.1 Overview

This section presents an overview of the image captioning task and briefly explains our methodology with the help of Algorithm 1.

This study presents three distinct approaches for Urdu image captioning. Our experiments and analysis are based on the Urdu version of the flicker8k dataset (Afzal et al. 2023), and further details about this dataset are provided in the corresponding subsections. To enhance the linguistic representation, a Word2Vec embedding model was trained specifically with Urdu texts. We conducted experiments using both CNN with attention and LSTM models in combination with the Word2Vec Urdu embeddings. Additionally, we explored the utilization of a Transformer model integrated with a CNN module. Further subsections comprehensively discuss the deployed proposed models and datasets. The data and code used in this study can be used for future research via GitHub.[1]

In image captioning, preprocessing steps are essential to prepare both the image and text data for training a model that can generate descriptive captions for images. Below, we have elaborated on the preprocessing steps for both image and text data, including tokenization, sequence padding, and embedding generation using pre-trained Urdu word embeddings.

Image data preprocessing. The first step is to load and standardize the image data. Images are typically resized to a uniform dimension to ensure consistent input size for the model. Feature extraction is performed using a desired model. The model is typically truncated before the final classification layer, and the output from the last layer is used as image features. This results in a fixed-size feature vector for each image, which can be used as input to the captioning model.

Text data preprocessing. The text data, which includes image captions, needs to be tokenized into individual words. Tokenization divides the text into meaningful units, such as words or subword units. WordPiece was used to handle complex Urdu morphology and Word2Vec was trained on the Urdu corpus. In order to train models efficiently, sequences (sentences) need to be of the same length. Padding is applied to ensure uniform sequence lengths. For example, if the maximum caption length is 20 words, shorter captions are padded with a special token (e.g., $< PAD >$) to reach the maximum length.

Word embeddings are used to represent words as dense vectors. These embeddings are typically trained on a large corpus of text and can capture semantic relationships between words. The pre-trained embeddings can be loaded and used for the captioning model. Each word in a caption is replaced with its corresponding word embedding, resulting in a sequence of embedded vectors. Special tokens such as $< START >$ and $< END >$ are added to the caption sequences. These tokens indicate the beginning and end of a caption. They are important for sequence generation. Now, the preprocessed image features and text data are fed into the image captioning model. The model takes the image

---

[1] https://github.com/researchcode-1/Urdu_image_captioning.

features as input and generates a sequence of words to form a descriptive caption for the image. The model is trained to minimize the difference between the generated caption and the reference (ground-truth) caption using loss functions like cross-entropy loss. These preprocessing steps are crucial for training a successful image captioning model, as they enable the model to learn the relationships between image content and textual descriptions. The model learns to generate coherent and contextually relevant captions based on the preprocessed data.

3.2 Proposed models

3.2.1 CNN+AttenLSTM with Word2Vec embeddings

Word2Vec embeddings. To facilitate the vectorization of our input data, we harnessed Urdu Embeddings and conducted a series of experiments involving various embedding dimensions. A Word2Vec model was meticulously trained on the UR-Mono Urdu dataset, which was then subjected to multiple experiments using different embedding sizes, including 64, 128, and 512. The ultimate choice was an embedding size of 512, primarily to align with the prerequisites of the subsequent phase, which revolves around the training of the encoder-decoder or transformer model, demanding an input embedding size of 512. Additionally, several window sizes, such as 2, 5, 12, 32, and 50, were evaluated. Given the common occurrence of short sentences in our Urdu dataset (UR-Mono), a window size of 5 emerged as the top-performing option.

As a starting point, we embraced a CNN-LSTM model (Xu et al. 2015) for the task of Urdu Image caption generation. This model seamlessly integrates soft attention by employing CNN as an encoder and LSTM with an attention mechanism to ensure equitable focus on all image features throughout the temporal dimension. The model is equipped with trainable embedding layers, enabling concurrent training of embeddings alongside caption generation tasks. While this customization can yield substantial advantages when dealing with expansive caption datasets, our selected subset lacked an ample number of captions to yield satisfactory performance. Consequently, we turned to pre-trained Word2Vec embeddings, a proven asset in various machine-learning domains, including machine translation and text summarization (Zaman et al. 2020). In our approach, we made use of pre-trained Urdu-specific embeddings, meticulously trained on the UR-Mono dataset (Jawaid et al. 2014). The model architecture is visually depicted in Fig. 4.

Within the realm of language modelling for image captioning, the incorporation of ResNet50 becomes pertinent, primarily owing to its adeptness in mitigating vanishing gradient issues through residual connections. For image processing, we enlisted ResNet50 to encode input images and effectively extract salient features. Concretely, we harnessed the feature extraction modules of ResNet50 while excluding the final fully connected layer, typically responsible for classification. This workflow resulted in the emergence of a 3D tensor that aptly represented image features, marked by dimensions of $14 \times 14 \times 2048$. Subsequently, this tensor was thoughtfully reshaped into a 2D tensor bearing dimensions of 196×2048, ensuring compatibility for downstream processing.

On the decoding front, the arsenal of choice comprised LSTM with a soft attention mechanism, recognized for its proficiency in sequence-to-sequence modelling. The LSTM-based decoder is marked by an assembly of stacked LSTM layers, intricately intertwined with attention mechanisms to seamlessly interact with the encoded image features.

3.2.2 CNN + transformer model

In this variant of our model, the transformative potential of the Transformer model was harnessed for the noble pursuit of Urdu caption generation. When dissected into its constituent components, the image processing arm featured InceptionV3, while the text-based counterpart showcased Transformer (Afzal et al. 2023). The learning capabilities for Urdu image captioning saw marked improvements, courtesy of the insightful inclusion of self-attention within the transformer architecture. This delightful blend of image and text processing finds its visual manifestation in Fig. 1.

The training phase kicks off with the processing of input images by InceptionV3, effectively yielding a feature vector tailored to the image. On the flip side, the input embeddings are seamlessly relayed to the Transformer model, in tandem with the image vector. Within the decoder, a smart strategy unfolds, with words bearing the highest probability adorning the scene until the sentinel end-of-sentence token or the prescribed maximum sequence length is judiciously reached. This decision-making capacity of the decoder rests on the solid foundation of the Transformer, which comes equipped with the prowess to shine in training scenarios involving expansive datasets within relatively constrained time frames. With the transformer model in play, experimental results boast relatively elevated scores compared to its predecessors, providing the necessary impetus to explore the deployment of other Transformer variants, such as Roberta and Vision Transformer.

3.2.3 Vit+Roberta

In this particular variant of our experimentation, a transformative aura engulfs our entire pipeline, powered by the indomitable Transformer models. Transformer models have

Fig. 1 Proposed CNN+Transformer model Architecture
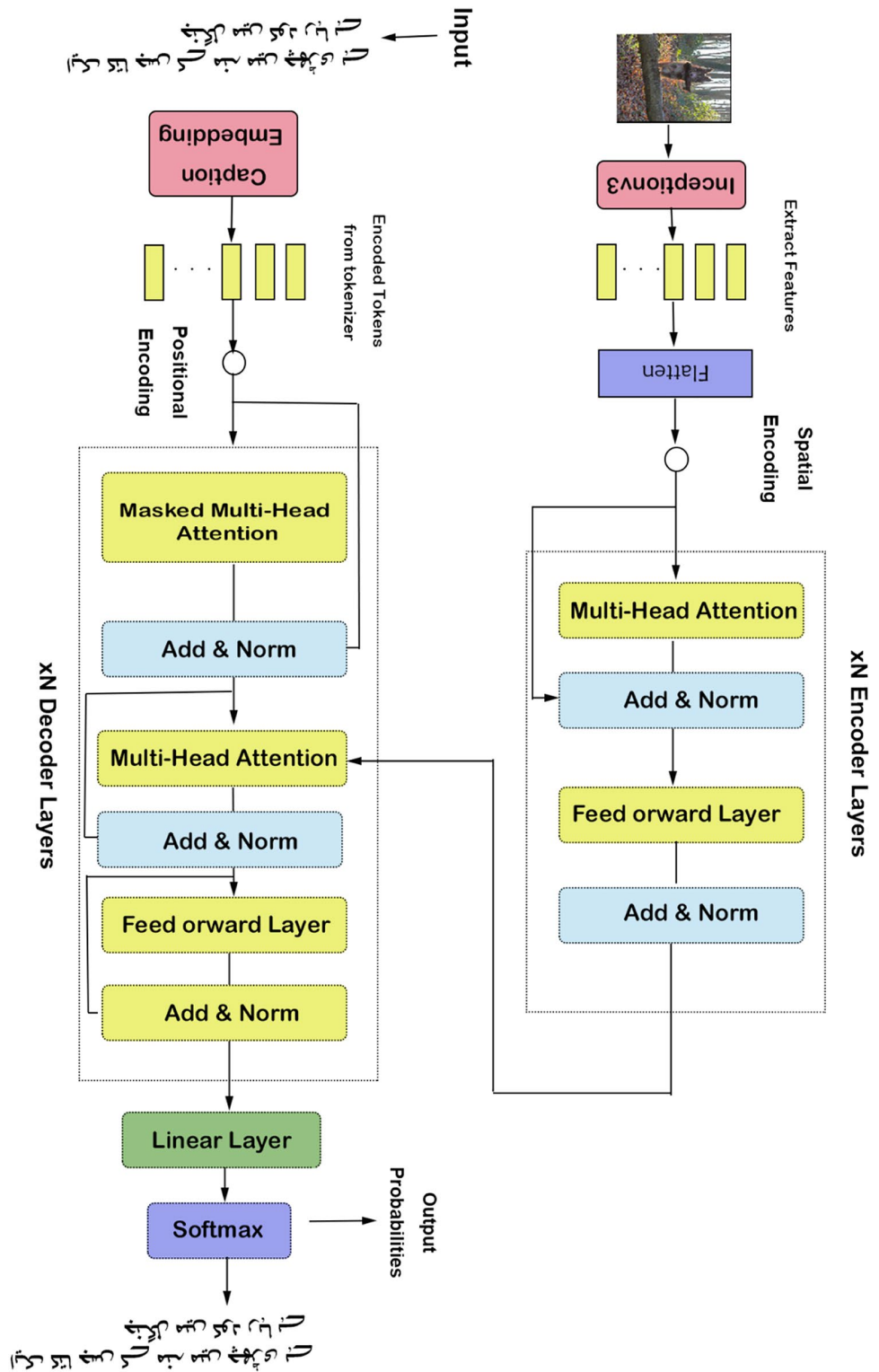
established a firm foothold in a myriad of NLP and computer vision tasks, including image classification and object detection. The foundation of image processing lays claim to the Vision Transformer (ViT) (Dosovitskiy et al. 2021), an ensemble of spectacular promise. The architectural blueprint is masterfully conveyed in Fig. 2. Our Transformer-based image encoder steps into the limelight, characterized by its engagement with 3D images as input. To usher in a coherent representation of the input, the encoder meticulously dissects the image into a sequence of input tokens. This
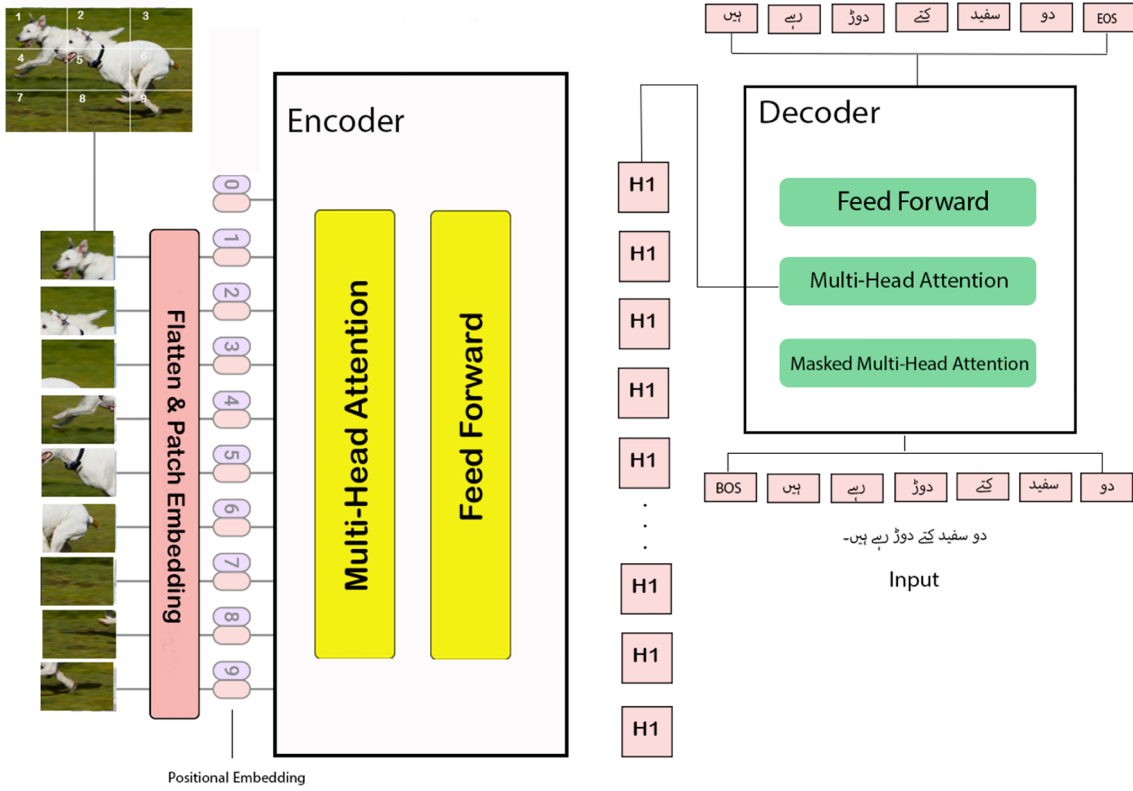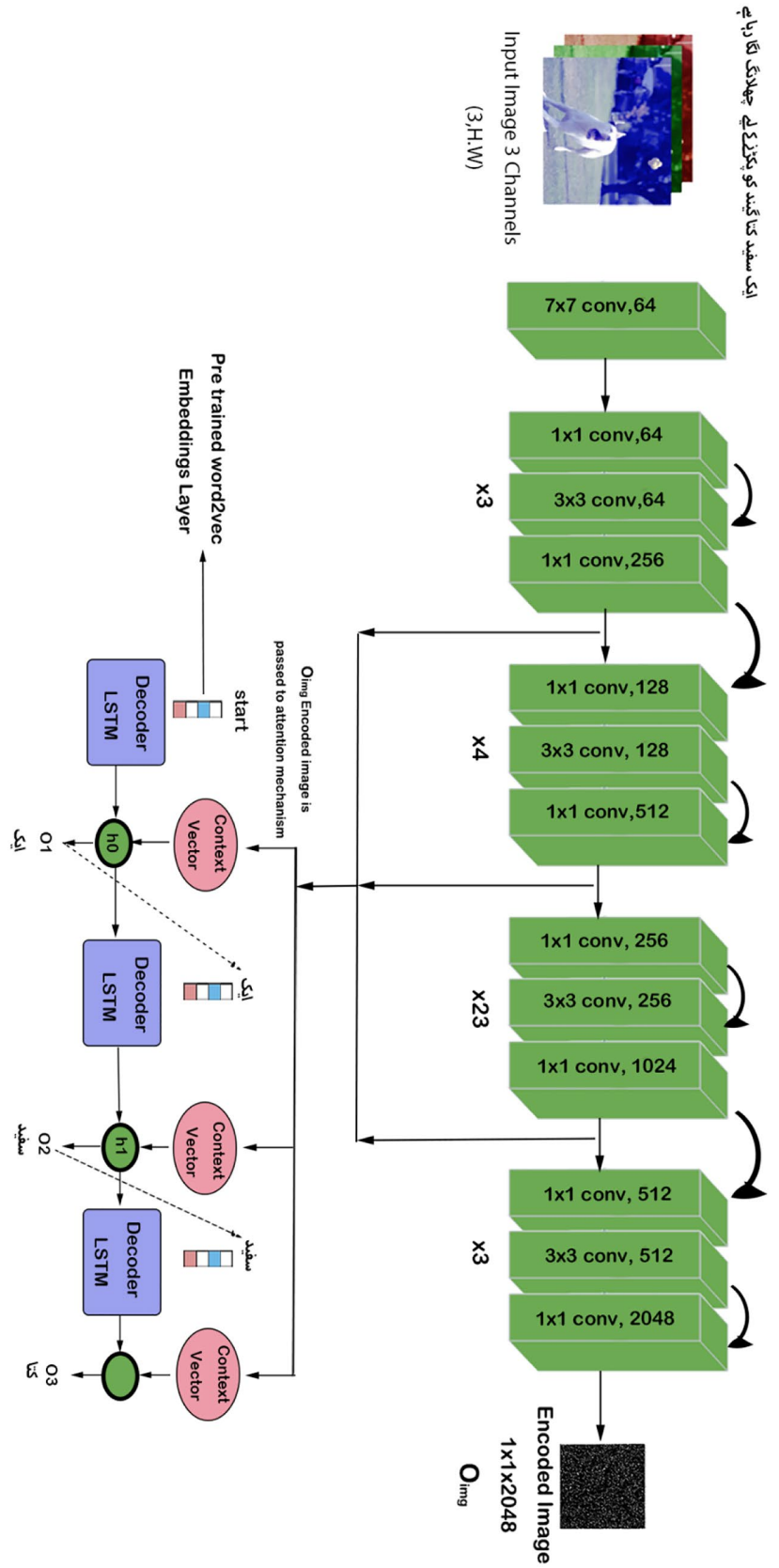
Fig. 2 Model Architecture of Vit+Roberta

Fig. 3 Images and their ground truth captions, predictions are presented in Table 7

Fig. 4 CNN+AttenLSTM with Word2Vec embeddings

granular division of the image unfolds through the act of slicing it into patches, guided by a patch size parameter denoted as $N = \frac{HW}{p}$. This approach ingeniously partitions the image into fixed-size patches while steadfastly adhering to the pre-defined height (H) and width (W) dimensions. These patches are duly flattened into vectors, culminating in the creation of a patch embedding, a collection of flattened image patch vectors. The Vision Transformer introduces an iconic [CLS] token, donned with the responsibility of spearheading a classification task, effectively summarizing the wealth of information hidden within the patch embeddings, and painting the holistic picture of the image. The journey marches forward as positional embedding enters the fray, harmoniously aligning with the patch embedding and ultimately finding its place in the encoder layers. A series of attention mechanisms take centre stage, bolstered by self-attention blocks that set the stage for a fully connected linear layer. This stage offers attention mechanisms the leeway to distribute their focus on the outputs by seamlessly introducing a weighted sum to the proceedings. In the domain of self-attention, the spotlight shines on the keys and values, both sourced from the same sequence, lending an air of distinction to the process. To tame the turbulence that occasionally simmers beneath the surface of the softmax function, a scaling factor of $\frac{1}{\sqrt{dk}}$ is judiciously applied. The scaling mechanism plays the crucial role of ensuring that gradients remain steadfast and unwavering. This mechanism empowers the attention module with the capacity to treat queries, keys, and values with the nuanced distinctions they deserve. In the context of the Vision Transformer, where the customary sequential framework takes a back seat, the attention module unfurls its magic upon the image as if it were a sequence, judiciously distributing attention to different patches, harmonizing with their corresponding outputs, effectively endowing the model with the ability to apprehend the relationships and dependencies that underpin the image.

When the focus transitions to the decoder component, it is the Roberta model that takes centre stage, an evolved iteration of the iconic BERT

1: **Input:** Ur-Mono dataset (5.4 million sentences, 95.4 million tokens), Urdu Image caption dataset (Flick8k Dogs) - 1800 Images, 9,000 captions
2: **Output:** Trained Word2Vec model, Trained CNN+Transformer model
3: **Word2Vec Model Training**
4: Load Ur-Mono dataset
5: Preprocess Ur-Mono dataset
6: **for** each sentence in dataset **do**
7:     Remove punctuation
8:     Remove invalid and junk characters
9:     Remove extra spaces
10: **end for**
11: Split preprocessed text into sentences
12: Train Word2Vec model on preprocessed data
13: **CNN+Transformer Model Training**
14: Load Urdu Image caption dataset
15: Preprocess Urdu Image caption dataset
16: **for** each caption in dataset **do**
17:     Remove punctuation
18:     Remove invalid and junk characters
19:     Remove extra spaces
20:     Split caption into sentences
21:     Add start and end tokens to each caption
22: **end for**
23: Translate all captions into Urdu
24: Train CNN+Transformer model on translated captions and corresponding images

Algorithm 1 Word2Vec and CNN+Transformer Model Training

4 Results and discussion

In this section, we discuss the dataset, and the experiments conducted and provide a detailed description of the achieved results. The results are presented in a tabular format, allowing for easy comparison. Additionally, we compare our findings with other relevant models in the field. Finally, we summarize the results and initiate a discussion. To evaluate the performance of each model, we employed evaluation metrics such as BLEU and its variants. These metrics serve as quantitative measures to assess the quality of the generated captions and enable a comprehensive evaluation of the model's performance.

Table 2 Summary of dataset used for Word2Vec Embeddings

Total Sentences	Total Tokens
5.4M	95.4M

Table 3 Summary of flickr8k dataset and dog subset

Dataset	Number of Images	Number of Captions
Flickr8k	8000	40,000
Flick8k Dogs	1800	9000

Table 4 Parameter settings for CNN+AttenLSTM and Transformer models

Model	Parameter	Setting
CNN+AttenLSTM	Word2Vec Embedding	Trained for 36 epochs
	Early Stopping	Based on BLEU score and validation loss
	Adaptive Learning Rate	Initial value: 0.0001
		Best BLEU-4 score achieved: 29.1
CNN (InceptionV3) +Transformer	Encoder	InceptionV3
	Batch Size	64
	Decoder Layers	4
	Attention Heads	8
	Dropout Rate	0.1
	Optimization	Adam
	Learning Rate Scheduler	Custom rate scheduling approach
	β_1	0.9
	β_2	0.98
	ϵ	10^{-9}
	Warmup Steps	4000
	Learning Rate	0.01
	Training Epochs	12
	Max Length	512
	Beam Size	4
	Repetition Penalty	Applied
	Data Split	80:10:10 for train-validate-test
	Decoder Model	Vision Transformer + Roberta

4.1 Dataset

To train and enhance our Urdu Word2Vec embeddings, we utilized the Ur-Mono dataset (Jawaid et al. 2014). For image caption Generation we used fliker8k Dogs Urdu subset. Although both mentioned datasets were preprocessed, we performed task-specific preprocessing on the Ur-Mono

Table 5 Comparison of different models based upon their BLEU, LASER, and BERT-F1

Methods	BLEU1	BLEU2	BLEU3	BLEU4	BERTF1	MicroLASER	MacroLASER
Ilahi et al. (2020)	68.0	–	–	–	–	–	–
Afzal et al. (2023)	72.5	56.9	42.8	31.6	85.07	73.99	74.44
CNN+LSTM(Word2Vec)	71.8	52.3	38.7	29.1	81.9	73.2	73.5
CNN+Transformers	78.9	63.96	49.826	37.68	87.0	80.3	79.99
VIT+Roberta	86.0	76.7	67.57	59.02	90.61	82.05	81.79

Table 6 Comparison of inference strategies for our best model (highest values are in bold format)

Inference Strategy	BLEU1	BLEU2	BLEU3	BLEU4
VIT+Roberta Greedy	83.34	72.83	62.97	54.07
VIT+Roberta - Beam 2	**85.28**	**75.4**	**66.61**	58.38
VIT+Roberta - Beam 3	84.83	74.88	66.14	58.17
VIT+Roberta - Beam 4	84.97	75.18	66.48	**58.48**

dataset such as punctuation removal, invalid and junk character removal, extra spaces removal, and then splitting into sentences. For the fliker8k Dogs dataset, we performed the same processing as for Ur-Mono with an addition of extra steps such as adding the start and end of tokens to each caption. Moreover, the Ur-Mono dataset consists of 95.4 million tokens and 5.4 million sentences (Jawaid et al. 2014). More details can be found in the Table. 2

We used the fliker8k Dogs dataset (Afzal et al. 2023), which is a subset of the original flicker8k dataset. The original flicker8k dataset consists of 8000 images, and for each image, there are 5 textual captions. We selected a subset that sums to 1800 images with corresponding 5 textual captions for each image. We then translated all the captions across the dataset into Urdu language, such that a total of 9000 captions were translated. The dataset was then divided into three subsets, training validation and testing sets. A summary of our selected subset is provided in Table 3.

4.2 Hyperparameters

During the experimentation, various values for the hyperparameters of each model were tweaked to find the optimal one given in Table 4. In this section, we describe each hyperparameter and its values.

For the CNN+AttenLSTM model, a word2vec embedding was utilized, which was trained for 36 epochs. Early stopping was applied based on the BLEU score and validation loss to stop training if either the loss increases or the BLEU score decreases. The adaptive learning rate was applied with 0.0001 as an initial value and the best values for BLEU-4 score were achieved 29.1.

Our experiments were conducted using the InceptionV3 and Transformer models, specifically training a CNN (InceptionV3) + Transformer model. InceptionV3 served as the encoder in this setup. The model was trained with a batch size of 64, providing efficient training and utilization

of computational resources. The decoder component of our model is comprised of four decoder layers and eight attention heads, enabling the model to effectively capture dependencies and generate accurate captions. To mitigate overfitting, a dropout rate of 0.1 was applied during training. For optimization, a custom rate scheduling approach was utilized in conjunction with the Adam optimizer. The coefficients for β_1 and β_2 were set to 0.9 and 0.98, respectively, while ϵ was set to 10^{-9}. These settings facilitated efficient convergence and improved model performance during training.

$$l_{rate} = d_{(-0.5)}^{model} \min(step_{num}^{0.5} \times step_{num} \times warmup_{steps}^{1.5}) \quad (1)$$

This corresponds to increasing the learning rate linearly for the first warm-up training steps and decreasing it thereafter proportionally to the inverse square root of the step number. With setting $warmup_steps = 4000$ and a learning rate of 0.01 we achieved the best result of $BLEU_1 = 78.9$ and $BLEU_4 = 87.0$. In the end, as a final contribution, we utilized an end-to-end transformer training approach using a vision transformer and Roberta. Where the encoder shares the encoded features with the multihead attention module of the decoder. The decoder takes the encoded features and input tokens as embeddings and generates captions based on cross-entropy loss. The model learns to improve the caption generation process using gradient descent. For training this seq2seq transformer model, a max length of 512 was set, batch size of 16, and beam size 4 with repetition penalty and trained for 12 epochs. Early stopping was used to avoid over-fitting and data split was 80:10:10 for train-validate-test. The results are reported in the form of BLEU scores and their variants. We found that our generated captions were contextually correct.

4.3 Comparison to state-of-the-art models

We conducted a series of experiments, and the results are presented in a tabular format. In evaluating the image captioning task, we employed the BLEU score as the evaluation metric. For both our proposed models and the baseline models, we reported the BLEU score. To provide a visual representation, Fig. 3 showcases input images along with their corresponding ground truth captions. Furthermore, we summarized the generated captions for the three models in Table 7. This comprehensive presentation allows for a clear comparison and analysis of the model's performance.

In Table 7, we can see that our proposed model ViT+Roberta outperforms other models, constituting two transformers one for the image to extract features and the other for text to decode extracted features into text.

5 Discussion and conclusions

We proposed and compared multiple architectures and training approaches to improve image captioning performance and the experimental results are given in Tables 5, 6 and 7. Specifically, we explored and compared various encoder-decoder models, including CNN+LSTMs, CNN+AttenLSTM, CNN+AttenLSTM with Urdu Word-2Vec embeddings, CNN+Baseline Transformer, and Vit+Roberta. By leveraging a wide range of evaluation metrics such as BLEU-1, BLEU-2, BLEU-3, BLEU-4, BERT-FScore, and LASER, we comprehensively assessed the performance of our models against the baseline. In addition to architecture comparison, we conducted experiments to investigate the impact of different training methods, such as early stopping, varying batch sizes, and custom rate scheduling. These experiments allowed us to gain insights into the optimal training strategies for our models (Fig. 4).

Based on our extensive experimentation and thorough analysis, we conclude that the transformer-based encoder-decoder model outperforms other architectures in terms of both quantitative and qualitative evaluations. Our first model inspired by the standard CNN+AttenLSTM model was tweaked to improve performance based on a trained URDU word2vec embedding. However, adding this layer didn't show any significant improvements. Then we experimented with current state-of-the-art language models and image classification models. We used a transformer as a replacement for our language model. Transformers can easily model long contexts and don't have vanishing gradient problems. One reason for opting for a transformer was the simplicity of the inner workings of the transformer. Our proposed model consistently generates Urdu captions that are superior in terms of syntax, context, and semantics. These findings demonstrate the effectiveness and potential of transformers in enhancing the quality of image captions in the Urdu language.

The resultant captions were produced by using different CNN models and changing the hyper-parameters of our transformer. We found out that Adam works well for our problem as an optimizer. We also found out that training transformers are efficient in terms of performance and compute cost as compared to RNNs/LSTMs.

Our results show how changes in architecture from experiment to experiment impact the results by measuring the BLEU metric, BERT-F1, and LASER metrics. We started with a baseline model with which we compared our results. Then the decoder with the base transformer model (Hodosh et al. 2013) kept the encoder with the same CNN-based model. This change improved the BLEU4 score from 31.6 to 37.6. But using a transformer also improved performance, reduced inference time for the model, and took less

Image	Model	Prediction	BLEU1
Figure 3 (a)	W2V	ایک کتا رکاوٹ کے اوپر اچمل رہا ہے	6.3
Figure 3 (a)	CNN+ Transformer	ایک کتا فریسبی کو پکڑنے کے لئے ہوا میں کود رہا ہے	50.79
Figure 3 (a)	ViT+ Roberta	ایک سیاہ سفید کتا ہوا میں فریسبی پکڑنے کے لئے چھلانگ لگا رہا ہے	82
Figure 3 (b)	W2V	ایک آدمی ایک کتے کے ساتھ کھیل رہا ہے	6.13
Figure 3 (b)	CNN+ Transformer	یک آدمی اور ایک کتا ساحل پر بیں	12
Figure 3 (b)	ViT+Roberta	ایک کتا اور ایک آدمی بندوق کے ساتہ ہے	100
Figure 3 (c)	W2V	ایک سیاہ کتا گھاس میں چھلانگ لگا رہا ہے	38
Figure 3 (c)	CNN+ Transformer	ایک سیاہ کتا ایک لکڑی کے اوپر سے چھلانگ لگا دیتا ہے	55
Figure 3 (c)	ViT+Roberta	ایک سیاہ کتا درخت کے تنے سے چھلانگ لگا رہا ہے	70
Figure 3 (d)	W2V	ایک بہورا کتا اپنے منہ میں چھڑی لئے ہوئے ہے	7.24
Figure 3 (d)	CNN+ Transformer	ایک کتا پانی میں کھڑا ہو رہا ہے	60.4
Figure 3 (d)	ViT+Roberta	ایک بہورا سفید کتا پانی میں کھڑا ہو رہا ہے	75.1
Figure 3 (e)	W2V	ایک عورت اپنے کتے کے ساتھ کھیل رہی ہے	49.2
Figure 3 (e)	CNN+ Transformer	ایک عورت اپنے سامنے کتے کے ساتھ بینچ کے نیچے بھاگ رہی ہے	53.4
Figure 3 (e)	ViT+Roberta	ایک عورت اپنے کتے کے ساتھ بینچ پر بیٹھ رہی ہے	73
Figure 3 (f)	W2V	ایک سیاہ سفید کتا گھاس میں بھاگ رہا ہے	49.2
Figure 3 (f)	CNN+ Transformer	ایک چھوٹا کتا اپنے منہ میں گیند پکڑے گھاس میں بھاگ رہا ہے	92.34
Figure 3 (f)	ViT+Roberta	ایک چھوٹا سیاہ سفید کتا اپنے منہ میں گیند پکڑے گھاس میں بھاگ رہا ہے	100
Figure 3 (g)	W2V	ایک بہورا کتا اپنے منہ میں چھڑی لئے ہوئے ہے	7.24
Figure 3 (g)	CNN+ Transformer	ایک کتا رکاوٹ کے راستے پر کود رہا ہے	33.33
Figure 3 (g)	ViT+Roberta	ایک بہورا کتا ایک رکاوٹ عبور کر رہا ہے	55
Figure 3 (h)	W2V	دو کتے پانی میں کھیل رہے ہیں	79.1
Figure 3 (h)	CNN+ Transformer	20 دو کتے ساحل سمندر پر کھیل رہے ہیں	90
Figure 3 (h)	ViT+Roberta	دو کتے ساحل سمندر پر پانی میں کھیل رہے ہیں	100

Table 7 Comparison of model caption predictions on various test images

time to train the model. The reason for that was, how transformers handle the embeddings and can process the whole sequence of tokens in a parallel manner with the help of positional encodings as compared to traditional sequential models that process tokens one by one in the order they are received.

Lastly, we noticed that the language model was performing well. However, the feature maps generated by the CNN model were not relevant and attention maps weren't

focusing on relevant areas for the token predicted by the decoder. Thus, we replaced the CNN with a vision transformer. Vision transformers show significant results on the ImageNet dataset. Thus, by using a transformer both for encoding as well as decoding we saw how it brought significant improvements to our results and improved model performance by a significant margin. Vision transformers were able to improve the image features extracted by an encoder. Hence, significantly improving the input to the decoder and resulting in much better, longer, and more contextual captions. By training Roberta on URDU corpus. We also demonstrated the effectiveness of pre-training approaches making it easy to train models for downstream tasks with less data and limited data resources.

References

Afzal MK, Shardlow M, Tuarob S et al (2023) Generative image captioning in Urdu using deep learning. J Ambient Intell Humaniz Comput 14(6):7719–31

Aneja J, Deshpande A, Schwing AG (2018) Convolutional image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5561–5570

Antol S, Agrawal A, Lu J et al (2015) Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp 2425–2433

Bakar A, Sarwar R, Hassan SU et al (2023) Extracting algorithmic complexity in scientific literature for advance searching. J Comput Appl Linguist 1:39–65

Bouchard C, Omhover Jf, Mougenot C, et al (2008) Trends: a content-based information retrieval system for designers. In: Design Computing and Cognition'08: Proceedings of the Third International Conference on Design Computing and Cognition. Springer, pp 593–611

Chen X, Lawrence Zitnick C (2015) Mind's eye: A recurrent visual representation for image caption generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2422–2431

Chen W, Lucchi A, Hofmann T (2016) A semi-supervised framework for image captioning. arXiv preprint arXiv:1611.05321

Cornia M, Stefanini M, Baraldi L, et al (2020) Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10578–10587

Dai B, Fidler S, Urtasun R, et al (2017) Towards diverse and natural image descriptions via a conditional gan. In: Proceedings of the IEEE international conference on computer vision, pp 2970–2979

Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), IEEE, pp 886–893

Del Chiaro R, Twardowski B, Bagdanov A et al (2020) Ratt: Recurrent attention to transient tasks for continual image captioning. Adv Neural Inf Process Syst 33:16736–16748

Dosovitskiy A, Beyer L, Kolesnikov A et al (2021) An image is worth 16 x 16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations

Dubey S, Olimov F, Rafique MA et al (2023) Label-attention transformer with geometrically coherent objects for image captioning. Inf Sci 623:812–831

Goodrum AA (2000) Image information retrieval: an overview of current research. Inf Sci 3:63

Hassan MU, Alaliyat S, Sarwar R et al (2023) Leveraging deep learning and big data to enhance computing curriculum for industry-relevant skills: a Norwegian case study. Heliyon 9(4):e15407

Hassan SU, Aljohani NR, Tarar UI et al (2023) Exploiting tweet sentiments in altmetrics large-scale data. J Inf Sci 49(5):1229–1245

He S, Liao W, Tavakoli HR et al (2020) Image captioning through image transformer. In: Proceedings of the Asian conference on computer vision

Herdade S, Kappeler A, Boakye K et al (2019) Image captioning: transforming objects into words. Advances in neural information processing systems 32

Hodosh M, Young P, Hockenmaier J (2013) Framing image description as a ranking task: data, models and evaluation metrics. J Artif Intell Res 47:853–899

Ilahi I, Zia HMA, Ahsan MA et al (2020) Efficient urdu caption generation using attention based lstm. arXiv preprint arXiv:2008.01663

Jawaid B, Kamran A, Bojar O (2014) A tagged corpus and a tagger for Urdu. In: LREC. pp 2938–2943

Karpathy A, Joulin A, Fei-Fei LF (2014) Deep fragment embeddings for bidirectional image sentence mapping. Advances in neural information processing systems 27

Khan MU, Abbas A, Rehman A et al (2020) Hateclassify: a service framework for hate speech identification on social media. IEEE Internet Comput 25(1):40–49

Kiros R, Salakhutdinov R, Zemel RS (2014) Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539

Li LH, Yatskar M, Yin D et al (2019) Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557

Li X, Yin X, Li C, et al (2020) Oscar: Object-semantics aligned pre-training for vision-language tasks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Springer, pp 121–137

Li J, Li D, Xiong C et al (2022) Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning, PMLR, pp 12888–12900

Limkonchotiwat P, Phatthiyaphaibun W, Sarwar R et al (2020) Domain adaptation of Thai word segmentation models using stacked ensemble. Association for Computational Linguistics

Lin TY, Maire M, Belongie S et al (2014) Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13, Springer, pp 740–755

Lowe DG (1999) Object recognition from local scale-invariant features. In: Proceedings of the seventh IEEE international conference on computer vision, IEEE, pp 1150–1157

Luo J, Li Y, Pan Y et al (2023) Semantic-conditional diffusion networks for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 23359–23368

Mao J, Wei X, Yang Y et al (2015) Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In: Proceedings of the IEEE international conference on computer vision, pp 2533–2541

Mohamed E, Sarwar R (2022) Linguistic features evaluation for hadith authenticity through automatic machine learning. Digit Scholarsh Humanit 37(3):830–843

Mohamed E, Sarwar R, Mostafa S (2023) Translator attribution for Arabic using machine learning. Digit Scholarsh Humanit 38(2):658–666

Mohammad S, Khan MU, Ali M, et al (2019) Bot detection using a single post on social media. In: 2019 third world conference on smart trends in systems security and sustainability (WorldS4), IEEE, pp 215–220

Ojala T, Pietikäinen M, Mäenpää T (2000) Gray scale and rotation invariant texture classification with local binary patterns. In: Computer Vision-ECCV 2000: 6th European Conference on Computer Vision Dublin, Ireland, June 26–July 1, 2000 Proceedings, Part I 6, Springer, pp 404–420

Ordonez V, Kulkarni G, Berg T (2011) Im2text: Describing images using 1 million captioned photographs. Advances in neural information processing systems 24

Ramos R, Martins B, Elliott D et al (2023) Smallcap: lightweight image captioning prompted with retrieval augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2840–2849

Saadany H, Mohamed E, Sarwar R (2023) Towards a better understanding of tarajem: Creating topological networks for Arabic biographical dictionaries. J Data Min Digit Humanit. https://doi.org/10.46298/jdmdh.8990

Sabah F, Chen Y, Yang Z et al (2023) Model optimization techniques in personalized federated learning: a survey. Expert Syst Appl 243:122874

Safder I, Mahmood Z, Sarwar R et al (2021) Sentiment analysis for Urdu online reviews using deep learning models. Expert Syst 38(8):e12751

Sap M, Shwartz V, Bosselut A, et al (2020) Commonsense reasoning for natural language processing. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pp 27–33

Sarwar R (2022) Author gender identification for urdu articles. In: International Conference on Computational and Corpus-Based Phraseology, Springer, pp 221–235

Sarwar R, Hassan SU (2021) Urduai: Writeprints for Urdu authorship identification. Trans Asian Low-Resour Lang Inf Process 21(2):1–18

Sharma P, Ding N, Goodman S, et al (2018) Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 2556–2565

Shetty R, Rohrbach M, Anne Hendricks L et al (2017) Speaking the same language: Matching machine to human captions by adversarial training. In: Proceedings of the IEEE International Conference on Computer Vision, pp 4135–4144

Silva K, Can B, Blain F et al (2023) Authorship attribution of late 19th century novels using gan-bert. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pp 310–320

Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. Advances in neural information processing systems 30

Wang P, Yang A, Men R et al (2022) Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: International Conference on Machine Learning, PMLR, pp 23318–23340

Xu K, Ba J, Kiros R et al (2015) Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning, PMLR, pp 2048–2057

Yan S, Xie Y, Wu F et al (2020) Image captioning via hierarchical attention mechanism and policy gradient optimization. Signal Process 167:107329

Yang X, Zhang H, Jin D et al (2020) Fashion captioning: Towards generating accurate descriptions with semantic rewards. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16, Springer, pp 1–17

Yao BZ, Yang X, Lin L et al (2010) I2t: Image parsing to text description. Proc IEEE 98(8):1485–1508

Zaman F, Shardlow M, Hassan SU et al (2020) Htss: A novel hybrid text summarisation and simplification architecture. Inf Process Manag 57(6):102351

Zhong Y, Wang L, Chen J et al (2020) Comprehensive image captioning via scene graph decomposition. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, Springer, pp 211–229