


TRCaptionNet: A novel and accurate deep Turkish image captioning model with vision transformer based image encoders and deep linguistic text decoders

Serdar YILDIZ^{1,3*}, Abbas MEMİŞ¹, Songül VARLI^{1,2}

¹Department of Computer Engineering, Faculty of Electrical and Electronics Engineering,

Yıldız Technical University, İstanbul, Türkiye

²Health Institutes of Türkiye, İstanbul, Türkiye

³BİLGEM, TÜBİTAK, Kocaeli, Türkiye

Received: 29.06.2023

Accepted/Published Online: 17.10.2023

Final Version: 27.10.2023

Abstract: Image captioning is known as a fundamental computer vision task aiming to figure out and describe what is happening in an image or image region. Through an image captioning process, it is ensured to describe and define the actions and the relations of the objects within the images. In this manner, the contents of the images can be understood and interpreted automatically by visual computing systems. In this paper, we proposed the TRCaptionNet a novel deep learning-based Turkish image captioning (TIC) model for the automatic generation of Turkish captions. The model we propose essentially consists of a basic image encoder, a feature projection module based on vision transformers, and a text decoder. In the first stage, the system encodes the input images via the CLIP (contrastive language–image pretraining) image encoder. The CLIP image features are then passed through a vision transformer and the final image features to be linked with the textual features are obtained. In the last stage, a deep text decoder exploiting a BERT (bidirectional encoder representations from transformers) based model is used to generate the image captions. Furthermore, unlike the related works, a natural language-based linguistic model called NLLB (No Language Left Behind) was employed to produce Turkish captions from the original English captions. Extensive performance evaluation studies were carried out and widely known image captioning quantification metrics such as BLEU, METEOR, ROUGE-L, and CIDEr were measured for the proposed model. Within the scope of the experiments, quite successful results were observed on MS COCO and Flickr30K datasets, two known and prominent datasets in this field. As a result of the comparative performance analysis by taking the existing reports in the current literature on TIC into consideration, it was witnessed that the proposed model has superior performance and outperforms the related works on TIC so far. Project details and demo links of TRCaptionNet will also be available on the project’s GitHub page (<https://github.com/serdaryildiz/TRCaptionNet>).

Key words: Image captioning, image understanding, Turkish image captioning, contrastive language–image pretraining, bidirectional encoder representations from transformers, image and natural language processing

1. Introduction

Computer vision, as it is widely known, basically aims to bring human vision to computer systems through cameras and image-sensing devices. In this context, a wide variety of computational image operations such as object detection, object localization, object classification, object segmentation, image classification, image enhancement and restoration, feature matching, key-point detection, three-dimensional reconstruction, image

*Correspondence: serdar.yildiz@std.yildiz.edu.tr

transformation, scene, and content analysis have been addressed and studied in the literature of computer vision. One of the essential topics studied in computer vision and image processing is image captioning which is simply defined as the generation of the text depicting an image, that is, the conversion of image content into words. In this manner, it is aimed to bring vision-based systems the ability to describe and interpret the content of pictures and video images automatically. As stated in [1], automatic image captioning is a more challenging computer vision task since the models need to deeply focus on the object scenes and object relationships to generate text descriptions that are correct both syntactically and semantically. Automatic image captioning has applications in a wide variety of research fields [2] including medical image captioning and interpretation [3, 4], assistive smart technologies for visually impaired people [5, 6], automated analysis of remote sensing images [7, 8], visual storytelling [9] and tourism [10]. It can also be in the other types of intelligent systems such as autonomous vehicles [11], robotic systems [12], and traffic scene understanding for intelligent transportation [13], as it can provide assistive data for the decision mechanisms [14].

The methodological solutions proposed for the image captioning problem are mainly classified into two categories [15]: (1) Traditional techniques including the retrieval-based [16] and template-based methods [17], and (2) the methods using deep learning-based approaches [18–20]. In retrieval-based image captioning, the candidate images are fetched and the captions of the query images are produced from the preexisting captions of the matching images [15]. Although the retrieval-based models can successfully generate image captions that are informative and grammatically correct, these models often fail to generate more descriptive and diverse captions for the new query images due to the limitations regarding the repository capacity [21]. On the other hand, template-based type of image captioning models usually produce image descriptions by using the syntactic rules defined previously [15] and discovering some visual image features such as objects and relationships [22]. Since these kinds of systems cannot represent visual content accurately, they cannot be able to produce meaningful image descriptions [15]. In the earlier image captioning studies, both retrieval-based and template-based were adopted to make the computers automatically generate the image descriptions. However, no satisfactory and no more promising performances could be achieved in such systems [22]. The rising and unprecedented success of deep-learning in computer vision has led to significant advances in image captioning in the last decade, as in many research fields. The main reason why deep learning-based models are so successful is that they simulate the neural network structure of the human brain to extract high-level and more complex features. In the current literature of deep learning-based image captioning, various approaches such as Convolutional Neural Networks (CNNs) [18], attention mechanisms [23], vision transformers [20], visual encoders [24], text decoders [25] and unsupervised learning strategies [19] have been analyzed. More detailed information on deep learning-based methods, datasets and evaluation metrics for image captioning can be followed from the extensive review and survey publications [26, 27] previously reported.

In this paper, we proposed a novel and accurate deep Turkish Image Captioning (TIC) model for the automatic generation of Turkish caption texts by leveraging the unprecedented success of deep learning. The model we proposed is essentially constructed based on the vision transformer-based image encoders and the deep linguistic text decoders. In the initial level, the model encodes the input images via the CLIP image encoder. Then, the CLIP image features are passed through a vision transformer. Finally, a text decoder module is used to generate the image captions. The main contributions of the paper are stated as follows:

- A novel deep image captioning model, named TRCaptionNet, is proposed for automatic Turkish caption generation.

- We also tried to ensure that different pretrained models such as BERTurk can be used in the text decoder module of TRCaptionNet (as a transfer learning strategy) by using a projection block located between the image encoder and text decoder structures.
- The third contribution of the article is that a natural language-based linguistic model called NLLB (No Language Left Behind) was employed to produce Turkish captions from the original English captions.
- Fourthly, as an ablation study, we analyzed the performances of multiple CLIP image encoder types in the image encoding module of the proposed model.
- As the final contribution, we carried out a transfer learning-based performance evaluation by using the weights of the pretrained BERTurk model in the text decoder module of our proposed caption generator model.

The caption model proposed in this study can not only be used for generic Turkish image captioning tasks but it can also be used as a submodule in some types of smart systems such as intelligent assistive technologies for visually impaired people, and autonomous systems requiring automatic scene understanding/analysis. In addition, it can be integrated into the other types of intelligent systems that need visual descriptions. Furthermore, the descriptions produced by such a caption model can also be categorized and they can be considered in automatic decision-making and classification mechanisms. The rest of the article is organized as follows: In Section 2, the related works in the current literature of TIC are discussed and described briefly. The image captioning datasets used to evaluate the proposed model are introduced in Section 3. In Section 4, the methodological structure of the proposed model is delineated and all the procedures handled in the implementation of the system are explained. Experimental results and discussions are stated in Section 5. Finally, the conclusions are emphasized in Section 6.

2. Related works

As is known, the literature of image captioning is quite deep. However, deep learning-based image captioning applications have increased considerably in recent years. In these studies, not only a wide variety of deep caption generation models were proposed, but also high success rates were achieved compared to the traditional image captioning approaches. In this context, we have discussed some deep learning-based prominent and featured research studies in the last few years in this section.

In a recent article [29], Ma et al. reported an image captioning study focusing on local visual analysis. They proposed a novel Locality-Sensitive Transformer Network (LSTNet) which enhances the ability of local perception. The LSTNet also includes two novel network designs named Locality-Sensitive Attention (LSA) and Locality-Sensitive Fusion (LSF) to improve local and semantic understanding. In the related article [29], LSTNet is reported to have superior performance on MS-COCO compared to the SOTA image captioning models. It is also reported that the related model was evaluated and verified on Flickr dataset variants. In another recent article [30] addressing the task of image captioning, Hu et al. proposed a Multi-head Association Attention Enhancement Network (MAENet) which achieves competitive success rates on the MS COCO dataset. In [31], Wang et al. presented a prompt-based image captioning model which optimizes prompt embeddings to exploit stylized data. Their model is also capable of generating diverse image captions by providing distinct prompts. They achieved superior performance on the MS COCO benchmark dataset with their prompt-based image captioning approach. In [32], Hu et al. presented a Triple-Stream Feature Fusion Network (TSFNet) in which a novel Dual-level Attention (DA) mechanism is proposed. They aimed to take advantage of the grid,

region, and scene graph triple-stream visual representations in TSFNet for the task of image captioning. In performance evaluation tests on the MS COCO dataset, they observed quite successful results that outperform several SOTA image captioning approaches.

In [33], Jiang et al. proposed a novel Hybrid Attention Network (HAN) that combines the prevalent machine attention procedures with human captioning attention. In experimental analyses performed on Flickr and MS COCO datasets, they observed high success rates with their HAN model. In [34], Hu et al. proposed a Bi-Positional Attention (BPA) module that combines the absolute and relative position encodings. With the incorporation of these encodings, the object relations and geometric information in an image are intended to be found. They also constructed a Position-Guided Transformer (PGT) network which is able to learn more exhaustive positional representations. It is reported in the related article [34] that competitive performance results were achieved on the MS COCO dataset. In [35], Wang et al. proposed an improved Geometry Attention Transformer (GAT) framework. They also designed two new geometry-aware structures to achieve geometric representation ability: i) a geometry gate-controlled self-attention refiner (in encoder), and ii) a group of position-LSTMs (in decoder). It is reported that the GAT framework provides quite successful rates on the MS COCO and Flickr image caption datasets. In [36], Wei et al. presented an Outside-in Attention. By using this approach, they aimed for the model to learn the dependencies within the image regions and dependencies between the image regions. In the related paper [36], it is stated that the proposed approach which was integrated into a Sequential Transformer Framework (S-Transformer) achieves successful results on the MS COCO dataset. In [37], Wang et al. introduced a contextual and selective attention network (CoSA-Net) that memorizes contextual attention and finds out the primary components of each attention. In the extensive performance evaluation tests, they observed that the CoSA-Net has superior performance on the MS COCO dataset. In [38], Wang et al. presented a new fully attentive network and they also proposed contrastive learning for image captioning, which takes the word-level and sentence-level semantics into consideration. In experimental tests performed on the MS COCO dataset, they observed superior success rates with their model and contrastive learning approach.

In [39], Ji et al. presented an image captioning model named Multi-branch Distance-sensitive Self-Attention Network (MD-SAN) and introduced Distance-sensitive Self-Attention (DSA) and Multi-branch Self-Attention (MSA) for the distance insensitivity and low-rank bottleneck issues of the existing self-attention based networks. They validated the efficacy of DSA and MSA on the MS COCO dataset. In [40], Du et al. introduced a new approach to balance accuracy and diversity in image captioning. Their model incorporates saliency information and relative position information of the objects. It is reported in [40] that the related model is successful in the generation of diverse or accurate captions. In [41], Wang and Gu proposed the Joint Relationship Attention Network (JRAN) to improve the feature relationship representations in image captioning. Their model finds out the feature relationships and it is also capable of fully learning both visual relationships and visual-semantic relationships. They observed successful captioning statistics on MS COCO and Flickr datasets. In addition to the related works mentioned above, there are also many image captioning studies [42–46] conducted in the last few years on MS COCO and Flickr datasets.

Since the challenge of image captioning we addressed in this paper is specific to TIC, we considered that it would be more appropriate to include the research studies related to the TIC in this section. In this context, we analyzed and discussed the reports in the current TIC literature. In an early work [47] proposed by Unal et al., a new dataset that can be employed as a benchmark is presented to enable the generation of Turkish

image descriptions for the task of TIC. The related TIC dataset called “TasvirEt” was created by collecting Turkish captions for the images in the Flickr8K dataset. Unal et al. also tried two data-driven approaches in this study [47] to solve the problem of caption generation in Turkish. Since the numbers and sizes of the TIC datasets are quite limited and also the creation process of a new image captioning dataset is time-consuming and quite laborious, Samet et al. [48] examined whether a training set obtained via an automatic translation tool could be used or not in the automatic caption generation process and they observed successful results in the generation of the Turkish caption texts. In TIC, one of the specific challenges is the suffixes that may change the meanings of the words completely due to the nature of the language. In [49], Kuyu et al. proposed a study to address the related challenge in generation of the Turkish caption texts. They proposed a Long Short Term Memory (LSTM) based deep learning model using the subword units and stated that the related model provides more successful results than the word-based model. In [50] proposed for TIC, Yilmaz et al. used a deep learning-based model composed of a CNN-based encoding module (for the extraction of the input image features) and an RNN-based (Recurrent Neural Network) decoding module (for the generation of descriptions as the image captions). In the related study [50], MS COCO, a widely known dataset, is used and the Turkish captions obtained by translating the original captions in English via Yandex Translation API were evaluated. In another study [51] similar to [50], Yıldız et al. also employed a deep model that uses an encoder-decoder architecture, in which a CNN structure encodes the input images and an LSTM module generates the image captions. The related model was evaluated on the MS COCO dataset and Turkish captions were automatically translated via Yandex translation API. In [51], it is stated that the performance evaluation results for TIC are quite satisfactory.

In a recent and thematic study [52] on TIC, Atıcı and Omurca automatically generated titles for the product images by using deep learning-based image captioning methods. It was reported in [52] that fair successful results were observed in the description of product images as a result of the experimental studies performed on approximately 1.8 million images and titles. In another more recent study [53], Ani and Amasyalı proposed a Turkish CLIP (Contrastive Language-Image Pre-Training) model. The related model called TrCLIP was trained with 2.5M images and captions that were translated into Turkish with Google Translate. The performance of the TrCLIP was evaluated on e-commerce data and a vast domain-independent dataset, and it is stated that there is no need for any extra fine-tuning for the model to work in Turkish. In [54], Golech et al. introduced the Turkish MS COCO dataset. The related dataset was built by translating the original MS COCO captions into Turkish with a translation API, Google Translate. In [54], Golech et al. also observed successful results with Meshed Memory Transformers.

3. Materials

To perform the automatic caption generation in Turkish and evaluate the system performance we used two widely-known public image datasets in this study: MS COCO [55] and Flickr30K [56]. MS COCO is a large-scale benchmark image dataset, and its use is common and popular in fundamental computer vision tasks of object detection, segmentation, and captioning. Caption set MS COCO [57] contains over 1,500,000 captions describing over 330,000 images and 5 captions are provided for each image. Flickr30K dataset consists of 31,783 images of everyday activities, events, and scenes, and a total of 158,915 descriptive captions (5 captions for each image) exist in the Flickr dataset. In particular, we evaluated our proposed model on MS COCO and Flickr30K subsets defined by Karpathy’s split [63]. The whole MS-COCO and Flickr30K datasets contain 123,287 and



Figure 1. Image samples from MS COCO (1st row) and Flickr30K (2nd row) datasets.

31,014 images, respectively. In the MS COCO dataset, 113,287 ($\approx 92\%$) of 123,287 images were separated for training, 5000 ($\approx 4\%$) for validation, and 5000 ($\approx 4\%$) for testing. On the other hand, 29,000 ($\approx 94\%$) of 31,014 images were separated for training, 1014 ($\approx 3\%$) for validation, and 1000 ($\approx 3\%$) for testing in Flickr30K dataset. In Figure 1, sample images from MS COCO and Flickr datasets are presented. In our study, we automatically translated the original MS COCO and Flickr captions in English into Turkish language by using a natural language-based deep linguistic model called NLLB (No Language Left Behind) [28]. The details of this process are stated in the next section in detail.

4. Methods

In this section, we introduce our novel Turkish image captioning model named TRCaptionNet and discuss other methodological details. We detailed TRCaptionNet in Section 4.1. In Section 4.2, the NLLB deep linguistic model, which is used in the automatic translation of the original MS COCO and Flickr captions in English into Turkish language, and the machine translation process are briefly explained. We additionally reviewed the most commonly used metrics in the image captioning literature in Section 4.3, and detailed the environmental setup in Section 4.4.

4.1. TRCaptionNet: A novel deep learning-based Turkish image captioning model

Deep image captioning models traditionally consist of two key components: an image encoder and a text decoder. The role of the image encoder is to convert the image into an embedding vector, while the text decoder transforms this vector into text. It is essential to train with an extensive dataset to create a robust, general-purpose image encoder and text decoder. However, this presents significant difficulty for languages with limited data, such as Turkish, and complicates the high-level generalization performance. To address this issue, we proposed a novel deep learning-based image captioning model, named TRCaptionNet, for Turkish. TRCaptionNet is essentially constructed by using a vision transformer-based image encoder, an image projection block and a deep linguistic text decoder. In our model, we propose the use of CLIP (contrastive language-image pretraining) [58], a language-independent image encoder that is known for its exceptional generalization abilities, to generate the image embedding vector. For the text decoding process, we employ a classical text decoder using the pretrained weights of the BERTurk [59] which is a BERT (bidirectional encoder representations from transformers) [60] based deep linguistic model for Turkish. We also propose employing the projection blocks to ensure compatibility between the two distinct image and text domains. The overall architecture is depicted in Figure 2.

Initially, in the image encoder block, the models introduced in the CLIP project [58] form the basis for obtaining embedding vectors. The CLIP project proposes using the Vision Transformer (ViT) [61] and ResNet

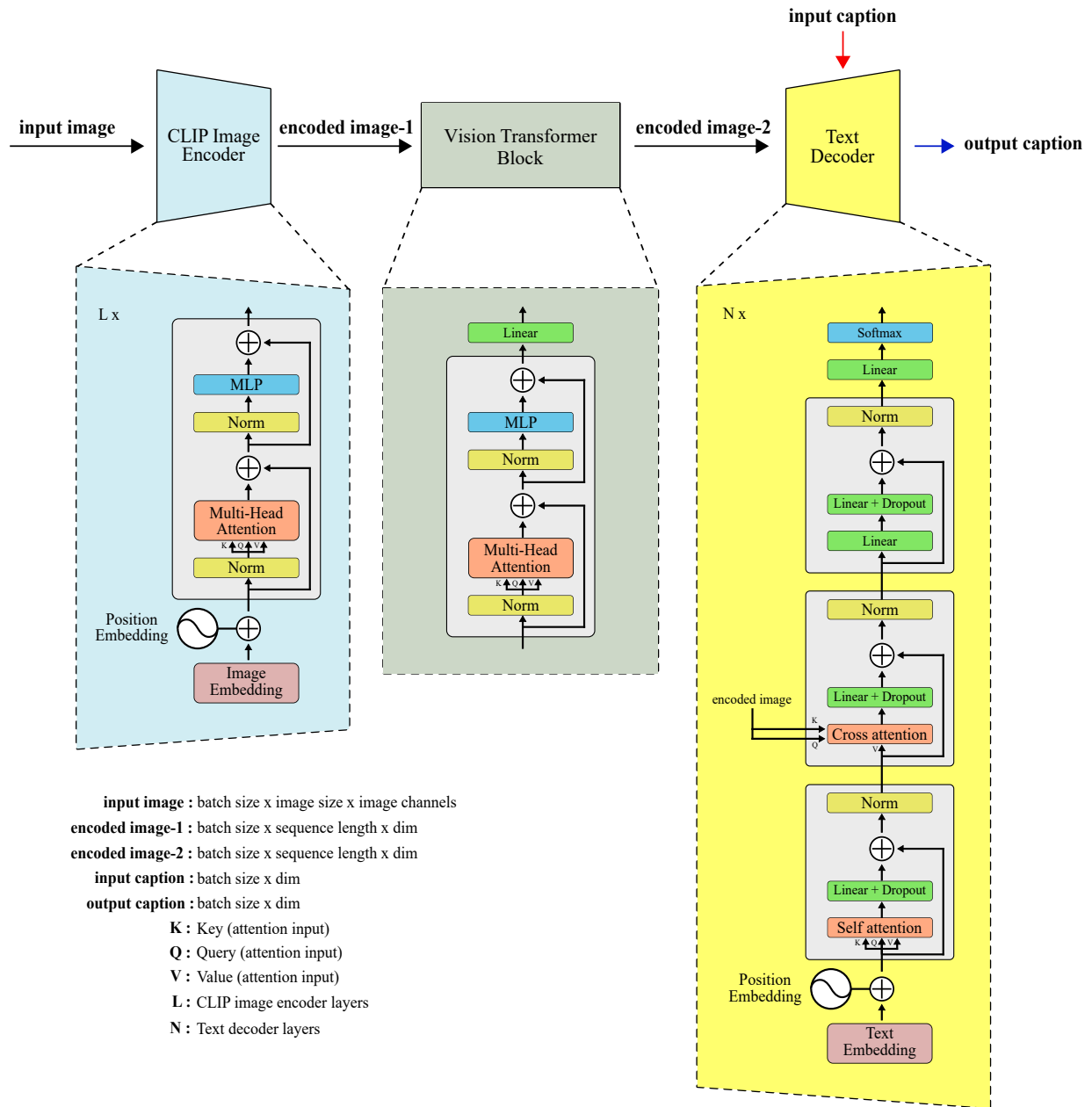


Figure 2. Visual scheme of the proposed deep Turkish image captioning model.

[62] models as image encoders. In our study, we adopted the ViT architectures and modified them to extract image patch embedding vectors. In the standard ViT architecture, an embedding vector is generated for each image patch, which is then combined via a multi-layer perceptron (MLP) head to generate the final image embedding vector. In our modified approach, we discard the combination process in the final layer of the vision encoder block and instead directly use the patch-derived embedding vectors. This modification aims to allow the decoder module to capture inter-patch relationships and be influenced by specific image details beyond the general context.

In practice, ViT takes the image $x \in \mathbb{R}^{3 \times h \times w}$ as input and produce image embedding $\mathbf{E}_{\text{clip}} \in \mathbb{R}^{N_{\text{patch}} \times d_{\text{vision}}}$ where N_{patch} denotes the number of patches and d_{vision} denotes the embedding dimension. The image encoder block can be represented as follows:

$$\mathbf{E}_{\text{clip}} = \text{CLIP}(x) \quad (1)$$

where $\text{CLIP}(\cdot)$ is the modified ViT encoder block. The projection block architecture, which follows the encoder block and consists of a vision transformer block with an MLP (multi-layer perceptron) layer, is crucial in our approach. Its primary function is to take the \mathbf{E}_{clip} embedding vector as an input and transform it into the $\mathbf{E}_{\text{text}} \in \mathbb{R}^{N_{\text{patch}} \times d_{\text{times}}}$ embedding vector. The process of projection of embedding vectors is as follows:

$$\mathbf{E}_{\text{text}} = \text{MLP}(\text{ViTBlock}(\mathbf{E}_{\text{clip}})) \quad (2)$$

where $\text{ViTBlock}(\cdot)$ denotes the vision transformer block, and $\text{MLP}(\cdot)$ represents a multi-layer perceptron. The projection block serves a critical role in the model, as it maps vision embedding vectors into a meaningful space for the text decoder and modifies the dimension of the embedding vector to match the requirements of the text decoder. Importantly, it also eliminates the need for training the image encoder block, enabling the use of pretrained models. This strategy reduces training costs and facilitates achieving high generalization performance, even with limited data.

The text decoder which takes the \mathbf{E}_{text} embedding vector generates corresponding text, token by token, in an autoregressive manner. It functions by using the embedding vectors derived from the image patches while also considering the associated features of the language. This approach ensures the generation of contextually accurate text. The text decoder module can be summarized as in the following equation:

$$T_{i+1} = \text{TextDecoder}(\mathbf{E}_{\text{text}}, T_1, T_2, \dots, T_i) \quad \text{for } i \in 1, 2, \dots, N - 1 \quad (3)$$

where T_1 refers to the initial token, T_N corresponds to the end of sequence token, and N represents the length of the sequence. To effectively learn the relationships between tokens, the text decoder cannot rely solely on captioning datasets. To overcome this limitation, we initialize the text decoder module by using the BERTurk deep linguistic model which includes the same text decoder architecture. The initialization process aims to transfer the features of the Turkish language to the model and enhance the performance of the model in caption text generation.

To simplify our approach, the process begins with the image encoder module which is used to generate embedding vectors from each image patch. These vectors are then transformed into appropriate embedding vectors for the text decoder in the projection module. Subsequently, captions for the image are produced by the text decoder module utilizing these embedding vectors. Notably, in our proposed model, the image encoder block is not involved in the training process, as it employs the pretrained CLIP model, and the projection block directly uses the generated patch embedding vectors.

4.2. Machine translation of captions

Due to the scarcity of the datasets proposed for the task of Turkish image captioning, various caption datasets that are originally constructed in English are usually converted into Turkish by using widely known language translation APIs such as Google Translate and Yandex Translate, and then employed. Contrary to the API-based language translation tools, we employed a deep machine translation model in this study. Specifically, caption annotations in English in COCO and Flickr30K datasets were translated into Turkish using a natural

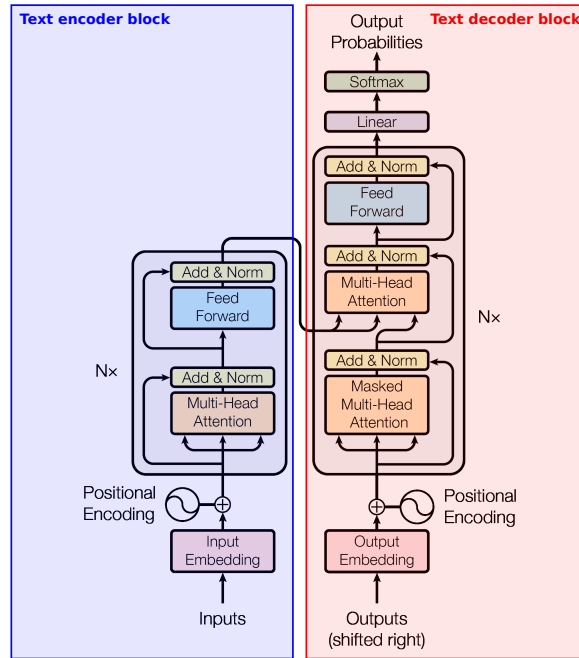


Figure 3. The transformer encoder-decoder architecture [63].

language-based deep linguistic model called NLLB (No Language Left Behind) [28]. The NLLB model is based on a transformer structure [63] and it supports a wide range of languages to translate texts. The transformer structure of the NLLB model consists of two main parts, a text encoding block and a text decoding block as illustrated in Figure 3. With the text encoder block, the source text in any language is encoded basically. Here, the source language sentence is tokenized and given as input to the encoder block consisting of self-attention and feed-forward layers. The encoder block finally converts the input sequence to embedding vectors. Then, the features of the input text are transferred to the text decoder block and the translation process into the target language is performed. The decoder block autoregressively generates the target sentence using the text embedding vectors. In the production of the sentence in the target language, beam search is also used. In the NLLB project, an extensive translation model supporting 200 unique languages is introduced. According to the reports [28], the assembled dataset encompasses more than one million samples of Turkish texts. Therefore, given its extensive Turkish data, it is considered appropriate for translation tasks from English to Turkish. Using the NLLB model, caption sets originally defined (5 caption texts for each image) in the COCO and Flickr30K datasets were translated into Turkish. In Figure 4, original English image captions for two sample images in MS COCO and Flickr30K datasets and corresponding Turkish captions that are automatically translated with the NLLB deep language translation model are presented.

4.3. Performance evaluation metrics

Measuring the success of image captioning models is more difficult than the other computer vision tasks. This is because the success of the predicted caption produced by the model must be matched both syntactically and semantically to reference captions. Although some metrics are available to calculate the similarity of the prediction to the reference caption in semantic space for English, there are no such recommended metrics for Turkish image captioning. For this reason, we employed four standard image captioning metrics [64] in this

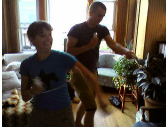

| | | |
|---|---|---|
|  | <p>MS COCO actual captions in English</p> <ol style="list-style-type: none"> 1) Two smiling young people playing a game of Wii. 2) A man and woman playing a video game together. 3) A man and woman in a living room play against each other on a Wii. 4) Two people are playing video games in a living room. 5) Two people play a game on a Wii. | <p>MS COCO translated captions in Turkish</p> <ol style="list-style-type: none"> 1) İki gülümseyen genç, bir oyun oynarken. 2) Bir erkek ve bir kadın birlikte bir video oyunu oynar. 3) Bir oturma odasında bir erkek ve bir kadın bir wii üzerinde birbirleriyle oynar. 4) İki kişi oturma odasında video oyun oynamaktadır. 5) İki kişi bir Wii'de bir oyun oynuyor. |
|  | <p>Flickr30k actual captions in English</p> <ol style="list-style-type: none"> 1) A man wearing a helmet, floating in the water. 2) A man wearing a white helmet is in the water. 3) Mountain climber safely lands in the water. 4) The spelunker finds water during his trek. 5) A man in a helmet treads water. | <p>Flickr30k translated captions in Turkish</p> <ol style="list-style-type: none"> 1) Suda yüzen bir kask giyen bir adam. 2) Beyaz bir kask giyen bir adam su altında. 3) Dağ tırmanıcısı suya güvenli bir şekilde iniyor. 4) Spelunker yürüyüş sırasında su bulur. 5) Bir kasklı adam suya basar. |

Figure 4. Original English image captions for two sample images in MS COCO and Flickr30K datasets and corresponding Turkish captions that are automatically translated with NLLB deep language translation model.

study as the scores of BLEU, METEOR, ROUGE-L, and CIDEr.

One of the most widely used quantification metrics in the literature of image captioning is the BLEU (bilingual evaluation understudy) [65] score. It was introduced first to measure machine translation accuracy. The BLEU score is calculated by taking the geometric mean of the precision scores produced on an n-gram basis for the predicted captions. It also multiplies the precision scores with a brevity penalty coefficient to penalize the captions that are too short. In addition, the BLEU score is named as the BLEU-n (BLEU-1, BLEU-2, BLEU-3, BLEU-4) according to the value of n-gram. Another n-gram-based caption quantification metric is the METEOR (Metric for Evaluation of Translation with Explicit ORDERing) [66]. But, unlike the BLEU score, it also examines the WordNet-based similarity and word root matching. In addition, the metric of METEOR considers the recall statistic. For these reasons, it is more successful in capturing the semantic similarity between the ground truth and predicted captions compared to BLUE. However, the computational cost is much higher than BLUE.

In our study, we also measured the ROUGE-L [67] and CIDEr [68] scores. The ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) is an LCS-based (longest common subsequence) caption quantification metric. In its calculation, it is assumed that the longest matching word sequence between the ground truth and predicted captions will express the similarity between the two caption texts. ROUGE-L basically measures the F1-score according to the sequence length and gives the similarity rate between the two texts. The CIDEr (Consensus-based Image Description Evaluation) score was developed specifically for the image captioning task, unlike the other quantification metrics. Similarly, the CIDEr metric is based on n-grams, but unlike the other metrics, it weights every n-gram using the TF-IDF (term frequency-inverse document frequency) method. In this way, it increases the weight of n-grams that are commonly found in the reference caption but less frequently in the entire dataset, and highlights this kind of words. This is an important metric of whether the model captions the prominent objects in an image.

4.4. Environmental setup

In the implementation of the experimental studies, various software tools and packages were utilized. In the construction of the proposed model, the pre-train CLIP (in image encoding process) and BERTurk (in text decoding process) projects were used. The caption annotations in English in COCO and Flickr30K datasets were translated into Turkish using the NLLB deep linguistic model. In the Beam-search algorithm (in the machine translation process of captions), the minimum number of tokens is set to 12, the maximum number of

Table 1. Parameter sets for 4 different CLIP image encoders.

| image encoder | input image resolution | image patch size | image grid size | sequence length | embedding dimension | encoder layers | attention block heads | parameter size |
|---------------------|------------------------|------------------|-----------------|-----------------|---------------------|----------------|-----------------------|----------------|
| CLIP ViT-B/16 | 224 x 224 | 16 | 14 x 14 | 197 | 768 | 12 | 12 | 149,620,737 |
| CLIP ViT-B/32 | 224 x 224 | 32 | 7 x 7 | 50 | 768 | 12 | 12 | 151,277,313 |
| CLIP ViT-L/14 | 224 x 224 | 14 | 16 x 16 | 257 | 1024 | 24 | 16 | 427,616,513 |
| CLIP ViT-L/14@336px | 336 x 336 | 14 | 24 x 24 | 577 | 1024 | 24 | 16 | 427,944,193 |

tokens to 35, and the beam size to 3. In the training of the TRCaptionNet, the learning rate parameter is set to $5 \cdot 10^{-4}$, beta values to $\beta = \{0.9, 0.999\}$ and weight decay to 0.01. The training process was continued for 50,000 iterations with a batch size of 64. In addition, AdamW was used as the optimizer algorithm. Performance analyses were carried out on a computer system with an Intel i9-12900K CPU, an NVIDIA GeForce RTX 3090 Ti graphics card, 32GB of memory, and running on Ubuntu-Linux operating system. In software-based implementation performed with Python (v3.8) programming language, Pytorch deep learning library (v2.0) was also used. The average time consumed in model training is approximately 24 h.

5. Results and discussions

As stated in the previous sections, our Turkish image captioning model consists of three principal components: An encoder, a vision transformer, and a text decoder. Since each of these components has a parametric structure, we trained different models by taking the predefined parameters into account and compared the performances of these models on two different captioning datasets, MS COCO, and Flickr30K. In the image encoding process, we defined 4 distinct models with 4 different parameter sets for the CLIP image encoder. These CLIP parameter sets belonging to 4 distinct models are presented in Table 1. As it can be noticed from Table 1, the parameters that constitute the differences between the CLIP encoder models are the resolution of the input images and the image patch size, which indicates the number of the patches into which the vision transformer structure in CLIP divides the input image. In Table 1, the column of “image grid size” represents the dimensions of the patches, the “sequence length” is calculated by adding the value of 1 to the multiplication of dimensions of patches, “encoder layers” indicates the repetition of layers (the parameter of L in Figure 2) and the “attention block heads” shows the total number of the attention blocks. The total number of the parameters for each CLIP image encoder is also given in Table 1.

The encoded image features obtained as the CLIP encoder output are transferred directly to the vision transformer block as delineated in Figure 2. The embedding dimension of the vision transformer block depends on the encoder embedding length which can be also followed from Table 1. The vision transformer block transforms encoded images by employing 1 encoder layer with 16 attention block heads and produces a projected feature vector of length 768. Then, the projected feature vector representing the input image is given to the text decoder which is also fed by the input captions. In this stage, we analyzed two distinct types of text decoder: i) A pretrained text decoder initialized with the BERTurk model, and ii) A normal no-pretrained text decoder. Both of the related text decoders have the same parameter set of {“embedding dimension”:768, “decoders layers”:12}. The parameter of “decoders layers” also represents the N in the text decoder block in Figure 2. In addition, the total parameter size of each text decoder is about 140M (139,016,192).

In the experimental studies, we trained a total of 8 distinct models on MS COCO and Flickr30K datasets by combining 4 image encoders, 1 vision transformer block, and 2 text decoders. In the first stage, we trained all 8 models by using only the MS COCO training set. Then, the MS COCO test images and Flickr30K test

Table 2. Performance statistics for the captions of COCO and Flickr test sets on all the 8 deep models trained only with MS COCO training data.

| Model | COCO test statistics | | | | | | | Flickr test statistics | | | | | | |
|---|----------------------|--------|--------|--------|--------|---------|--------|------------------------|--------|--------|--------|--------|---------|--------|
| | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L | CIDEr | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L | CIDEr |
| image encoder + (text decoder pretrain) | | | | | | | | | | | | | | |
| CLIP ViT-B/16 + (no pretrain) | 0.4658 | 0.3174 | 0.2057 | 0.1320 | 0.2172 | 0.4119 | 0.4851 | 0.4354 | 0.2651 | 0.1552 | 0.0853 | 0.1796 | 0.3579 | 0.2112 |
| CLIP ViT-B/32 + (no pretrain) | 0.4956 | 0.3369 | 0.2157 | 0.1397 | 0.2193 | 0.4133 | 0.4572 | 0.4315 | 0.2668 | 0.1602 | 0.0872 | 0.1743 | 0.3523 | 0.1996 |
| CLIP ViT-L/14 + (no pretrain) | 0.5304 | 0.3672 | 0.2379 | 0.1534 | 0.2306 | 0.4322 | 0.5205 | 0.4576 | 0.2881 | 0.1702 | 0.0922 | 0.1847 | 0.3718 | 0.2502 |
| CLIP ViT-L/14@336px + (no pretrain) | 0.5038 | 0.3465 | 0.2244 | 0.1444 | 0.2309 | 0.4298 | 0.4446 | 0.2849 | 0.1729 | 0.0975 | 0.1894 | 0.3827 | 0.2622 | |
| CLIP ViT-B/16 + (BERTurk) | 0.5612 | 0.4007 | 0.2724 | 0.1863 | 0.2476 | 0.4547 | 0.6227 | 0.4905 | 0.3236 | 0.2052 | 0.1240 | 0.2003 | 0.3928 | 0.3302 |
| CLIP ViT-B/32 + (BERTurk) | 0.5393 | 0.3777 | 0.2545 | 0.1721 | 0.2392 | 0.4408 | 0.5738 | 0.4667 | 0.3000 | 0.1862 | 0.1120 | 0.1909 | 0.3789 | 0.2851 |
| CLIP ViT-L/14 + (BERTurk) | 0.5694 | 0.4063 | 0.2772 | 0.1896 | 0.2512 | 0.4579 | 0.6392 | 0.5218 | 0.3484 | 0.2207 | 0.1315 | 0.2091 | 0.4113 | 0.3726 |
| CLIP ViT-L/14@336px + (BERTurk) | 0.5773 | 0.4146 | 0.2839 | 0.1954 | 0.2546 | 0.4652 | 0.6552 | 0.5311 | 0.3587 | 0.2293 | 0.1395 | 0.2146 | 0.4192 | 0.3798 |

Table 3. Performance statistics for the captions of COCO and Flickr test sets on all the 8 deep models trained only with Flickr30K training data.

| Model | COCO test statistics | | | | | | | Flickr test statistics | | | | | | |
|---|----------------------|--------|--------|--------|--------|---------|--------|------------------------|--------|--------|--------|--------|---------|--------|
| | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L | CIDEr | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L | CIDEr |
| image encoder + (text decoder pretrain) | | | | | | | | | | | | | | |
| CLIP ViT-B/16 + (no pretrain) | 0.3680 | 0.2128 | 0.1186 | 0.0667 | 0.1709 | 0.3299 | 0.2352 | 0.4826 | 0.3221 | 0.2103 | 0.1320 | 0.2052 | 0.3997 | 0.3546 |
| CLIP ViT-B/32 + (no pretrain) | 0.3670 | 0.2057 | 0.1142 | 0.0647 | 0.1647 | 0.3202 | 0.2055 | 0.4780 | 0.3152 | 0.2047 | 0.1273 | 0.1993 | 0.3904 | 0.3267 |
| CLIP ViT-L/14 + (no pretrain) | 0.3986 | 0.2397 | 0.1373 | 0.0777 | 0.1808 | 0.3484 | 0.2816 | 0.5286 | 0.3621 | 0.2434 | 0.1575 | 0.2178 | 0.4247 | 0.4324 |
| CLIP ViT-L/14@336px + (no pretrain) | 0.4190 | 0.2563 | 0.1483 | 0.0844 | 0.1874 | 0.3601 | 0.2966 | 0.5294 | 0.3608 | 0.2389 | 0.1530 | 0.2224 | 0.4319 | 0.4348 |
| CLIP ViT-B/16 + (BERTurk) | 0.3927 | 0.2340 | 0.1346 | 0.0801 | 0.1836 | 0.3462 | 0.2595 | 0.5163 | 0.3469 | 0.2275 | 0.1445 | 0.2150 | 0.4181 | 0.3898 |
| CLIP ViT-B/32 + (BERTurk) | 0.3748 | 0.2147 | 0.1172 | 0.0649 | 0.1722 | 0.3304 | 0.2314 | 0.4962 | 0.3269 | 0.2132 | 0.1358 | 0.2065 | 0.4022 | 0.3791 |
| CLIP ViT-L/14 + (BERTurk) | 0.4018 | 0.2454 | 0.1447 | 0.0871 | 0.1898 | 0.3547 | 0.2972 | 0.5314 | 0.3665 | 0.2452 | 0.1592 | 0.2241 | 0.4334 | 0.4519 |
| CLIP ViT-L/14@336px + (BERTurk) | 0.4213 | 0.2576 | 0.1518 | 0.0900 | 0.1933 | 0.3619 | 0.3145 | 0.5477 | 0.3771 | 0.2535 | 0.1671 | 0.2271 | 0.4393 | 0.4668 |

Table 4. Performance statistics for the captions of COCO and Flickr test sets on all the 8 deep models trained with both MS COCO and Flickr30K training data (MS COCO + Flickr30K).

| Model | COCO test statistics | | | | | | | Flickr test statistics | | | | | | |
|---|----------------------|--------|--------|--------|--------|---------|--------|------------------------|--------|--------|--------|--------|---------|--------|
| | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L | CIDEr | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L | CIDEr |
| image encoder + (text decoder pretrain) | | | | | | | | | | | | | | |
| CLIP ViT-B/16 + (no pretrain) | 0.5069 | 0.3438 | 0.2190 | 0.1416 | 0.2221 | 0.4127 | 0.4934 | 0.4754 | 0.2980 | 0.1801 | 0.1046 | 0.1902 | 0.3732 | 0.2907 |
| CLIP ViT-B/32 + (no pretrain) | 0.4795 | 0.3220 | 0.2056 | 0.1328 | 0.2157 | 0.4065 | 0.4512 | 0.4581 | 0.2866 | 0.1742 | 0.1014 | 0.1855 | 0.3754 | 0.2659 |
| CLIP ViT-L/14 + (no pretrain) | 0.5262 | 0.3643 | 0.2367 | 0.1534 | 0.2290 | 0.4296 | 0.5209 | 0.5186 | 0.3407 | 0.2184 | 0.1346 | 0.2045 | 0.4058 | 0.3507 |
| CLIP ViT-L/14@336px + (no pretrain) | 0.5325 | 0.3693 | 0.2376 | 0.1528 | 0.2338 | 0.4387 | 0.5288 | 0.5259 | 0.3525 | 0.2249 | 0.1334 | 0.2157 | 0.4237 | 0.3808 |
| CLIP ViT-B/16 + (BERTurk) | 0.5572 | 0.3945 | 0.2670 | 0.1814 | 0.2459 | 0.4499 | 0.6146 | 0.5400 | 0.3742 | 0.2533 | 0.1677 | 0.2232 | 0.4324 | 0.4636 |
| CLIP ViT-B/32 + (BERTurk) | 0.5412 | 0.3802 | 0.2555 | 0.1715 | 0.2387 | 0.4419 | 0.5848 | 0.5182 | 0.3523 | 0.2348 | 0.1532 | 0.2105 | 0.4079 | 0.4010 |
| CLIP ViT-L/14 + (BERTurk) | 0.5761 | 0.4124 | 0.2803 | 0.1905 | 0.2523 | 0.4609 | 0.6437 | 0.5713 | 0.4056 | 0.2789 | 0.1843 | 0.2330 | 0.4491 | 0.5154 |
| CLIP ViT-L/14@336px + (BERTurk) | 0.4639 | 0.3198 | 0.2077 | 0.1346 | 0.2276 | 0.4190 | 0.4971 | 0.4548 | 0.3039 | 0.1937 | 0.1179 | 0.2056 | 0.3966 | 0.3550 |

images were evaluated separately on these models. The performance statistics observed for the COCO and Flickr test captions are given in Table 2. As seen in Table 2, the model with ViT-L/14@336px CLIP encoder and BERTurk pretrained text decoder provides the best success rates on both test sets for all the performance evaluation metrics. In the second stage of the experimental analysis, we trained all 8 models by using only the Flickr30K training set. Then, the MS COCO test images and Flickr30K test images were evaluated separately on these models. The performance statistics observed for the COCO and Flickr test captions are given in Table 3. It can be observed from Table 3 that the model of CLIP ViT-L/14@336px + BERTurk has superior performance on the test data of both data sets when only Flickr30K is used in training of the models.

In the third stage, a single training set including 142,287 images (113,287 images from COCO + 29,000 images from Flickr30K) and corresponding captions was created by combining the training sets of both datasets. Then, all models were trained by using this combined set of data. The performance statistics observed for the COCO and Flickr test captions in the models that are trained with combined data are given in Table 4. Contrary to the case in Table 2 and Table 3, the CLIP ViT-L/14@336px + BERTurk model has a poor performance in accordance with the other models. In the performance analysis carried out on the combined training set, the model of CLIP ViT-L/14 + BERTurk has achieved the best success rates. Besides the success rates in Table 2, Table 3 and Table 4, some sample caption predictions quantified as successful and unsuccessful according to the BLEU scores from both datasets are presented in Figure 5 and Figure 6.





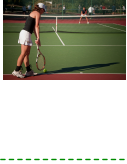

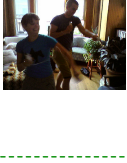



| | | | |
|---|---|--|---|
|  <p>reference captions Çimenli bir alanda iki adam frisbee oynuyordu. Mavi gömlekli bir adam beyaz bir frisbee tutuyor. Bir adamın bir frisbee yakaladığı bulanık bir fotoğraf. Bir kişi havaya atlayıp bir frisbee yakalıyor ve bir kişi onu kovalıyor. Bir frisbee yakalamak için havaya atlayan bir adam.</p> <p>predicted caption bir adam havaya atlayıp bir frisbee yakalıyor</p> | <p>BLEU-scores BLEU-1: 1.0000 BLEU-2: 0.9129 BLEU-3: 0.7937 BLEU-4: 0.7071</p> |  <p>reference captions Boş ve açık bir buzdolabı ve bir şişeler rafı olan bir oda. duvarın yanında bir buzdolabı bir buzdolabı bir mutfakta kapısı açık. içinde açık bir buzdolabı olan bir oda. Açık bir buzdolabı ve üstündeki kavanozlu rafı bir oda.</p> <p>predicted caption buzdolabı ve kurutucu ile bir mutfak alanı, duvarda posterler ve kağıtlar var</p> | <p>BLEU-scores BLEU-1: 0.3333 BLEU-2: 0.1741 BLEU-3: 0.0000 BLEU-4: 0.0000</p> |
|  <p>reference captions Bir erkek ve bir kadın çikolata kalbi şeklinde bir kek kesiyorlar. Bir erkek ve bir kadın bir pastanı kesmek için bıçak paylaşıyorlar. İki kişi birlikte bir tabakta bir kek kesip duruyor. İki kişi bir bıçak tutarak birlikte bir kek keser. Özel bir olay için güzel bir kek kesen mutlu bir çift.</p> <p>predicted caption bir erkek ve bir kadın birlikte bir kek kesiyorlar</p> | <p>BLEU-scores BLEU-1: 1.0000 BLEU-2: 0.9354 BLEU-3: 0.8550 BLEU-4: 0.6756</p> |  <p>reference captions Çimenler arasında yürüyen filler sürüsü Çimenlerde duran küçük bir fil sürüsü. ağaçların önünde bir tarlada duran bir fil grubunu Çimenli bir yerde, arka planda ağaçlarla yürüyen bir fil sürüsü. Çimenlerde süren birkaç filin resmi.</p> <p>predicted caption bir grup fil, otlu bir tarla boyunca yürüyor ve arka planda ağaçlar var</p> | <p>BLEU-scores BLEU-1: 0.3846 BLEU-2: 0.1790 BLEU-3: 0.0000 BLEU-4: 0.0000</p> |
|  <p>reference captions siyah bir üstlülkü bir tenisçi kadın tenis oynar. tenis meydanında bir raketi tutan bir kadın. Bir kadın tenisçi topunu servis etmeye hazırlanıyor. Bir kadın bir tenis raketi ile bir meydanın üzerinde duruyor. Bir kadın yere tenis topu almak için hazırlanıyor.</p> <p>predicted caption tenis meydanında bir raketi tutan bir kadın, tenis topu üzerinde sallıyor</p> | <p>BLEU-scores BLEU-1: 0.8182 BLEU-2: 0.8090 BLEU-3: 0.7138 BLEU-4: 0.6530</p> |  <p>reference captions Bir kutu, bol bol yiyeceklerle dolu. Çok lezzetli bir yemek ve biraz sebze. ağşap masada beyaz bir tabakta sergilenen çeşitli gıda ürünleri. Brokoli quinoa siyah fasulyeler ve sebze içeren bir tabak dolu. Brokoli ve mısır içeren bir kase dolu yiyecek.</p> <p>predicted caption içinde yiyecek olan bir kase var ve üzerinde çok farklı yiyecek var</p> | <p>BLEU-scores BLEU-1: 0.4167 BLEU-2: 0.1946 BLEU-3: 0.0000 BLEU-4: 0.0000</p> |
|  <p>reference captions İki gülümseyen genç, bir oyun oynarken. Bir erkek ve bir kadın birlikte bir video oyunu oynar. Bir oturma odasında bir erkek ve bir kadın bir wii üzerinde birbirleriyle oynar. İki kişi oturma odasında video oyun oynamaktadır. İki kişi bir Wii'de bir oyun oynuyor.</p> <p>predicted caption bir erkek ve bir kadın oturma odasında bir wii ile oynamaktadır</p> | <p>BLEU-scores BLEU-1: 0.9091 BLEU-2: 0.7977 BLEU-3: 0.6564 BLEU-4: 0.5157</p> |  <p>reference captions Büyük bir otobüs ve bir kamyon sokakta. bir şehir otobüsüne bağlanmış büyük bir çekme kamyonu. Mavi ve beyaz bir çekme kamyonu arkasında mavi bir otobüs sürüyor. Büyük çekme kamyonu arka tarafta bir şehir otobüsü taşıyor. Bir kamyon, yoğun bir sokakta otobüsü çekiyor.</p> <p>predicted caption bir otobüs, bir konut caddesi üzerinde bir kamu taşımacılığı otobüsünü çekiyor</p> | <p>BLEU-scores BLEU-1: 0.3636 BLEU-2: 0.1907 BLEU-3: 0.0000 BLEU-4: 0.0000</p> |
|  <p>reference captions Bir adam şemsiye tutan bir adamla bir kaldırımda yürüyen birkaç kişi. gökkuşağı şemsiyesini taşıyan bir kişi. bir kişi büyük renkli bir şemsiye tutuyor Gökkuşağı renkli bir şemsiye taşıyan bir insan sokakta yürüyor Bir kare içinde gökkuşağı renkli bir şemsiye taşıyan bir kişi.</p> <p>predicted caption bir kişi gökkuşağı renkli bir şemsiye tutan bir sokakta yürüyor</p> | <p>BLEU-scores BLEU-1: 1.0000 BLEU-2: 0.8819 BLEU-3: 0.6632 BLEU-4: 0.4518</p> |  <p>reference captions büyük bir ayna önünde dans eden bir kadın. Bir kadın, ayna duvarının önünde dans pozunda duruyor. bir stüdyoda dans pistinde poz yapan bir kadın. Bir kadın insanların önünde poz yapıyor. Yere oturmuş, etrafı bir izleyiciyle birlikte performans sergiliyor.</p> <p>predicted caption siyah giysiler giyen bir kadın dans ediyor ve bir sopayı tutuyor</p> | <p>BLEU-scores BLEU-1: 0.3636 BLEU-2: 0.1907 BLEU-3: 0.0000 BLEU-4: 0.0000</p> |

Figure 5. Caption prediction samples quantified as successful (left column) and unsuccessful (right column) according to the BLEU scores from MS COCO dataset.

As it can be seen in Table 2, Table 3 and Table 4, the success rates we observed on MS COCO and Flickr30K datasets are quite promising for TIC. The reports in the current TIC literature were analyzed and discussed in the previous section of related works. Although the literature of image captioning in foreign languages is quite deep, the current literature on TIC is quite limited in terms of the number of papers and datasets. Therefore, it was very difficult to make a detailed comparison, but we have shared a comparison that we made based on the success rates observed in current studies in Table 5. As can be followed from the related table, the performance values observed in this study are superior to the performance values presented in many other previous reports.

In the current literature on image captioning, there are also many image captioning studies in English on MS COCO and Flickr datasets. Some of these studies, especially presented within the last few years, and the success rates observed in these works are given in Table 6. As can be followed from Table 5 and Table 6, image captioning performances in English are quite higher than the Turkish image captioning stats. We think that the main reason for the low performances in Turkish image captioning models is that the caption texts in English are translated into Turkish by using machine translation systems due to the lack of comprehensive datasets for Turkish image captioning. Therefore, the success of the translation systems is directly related to the success of the captioning model. To overcome such problems, comprehensive Turkish caption sets can be constructed to

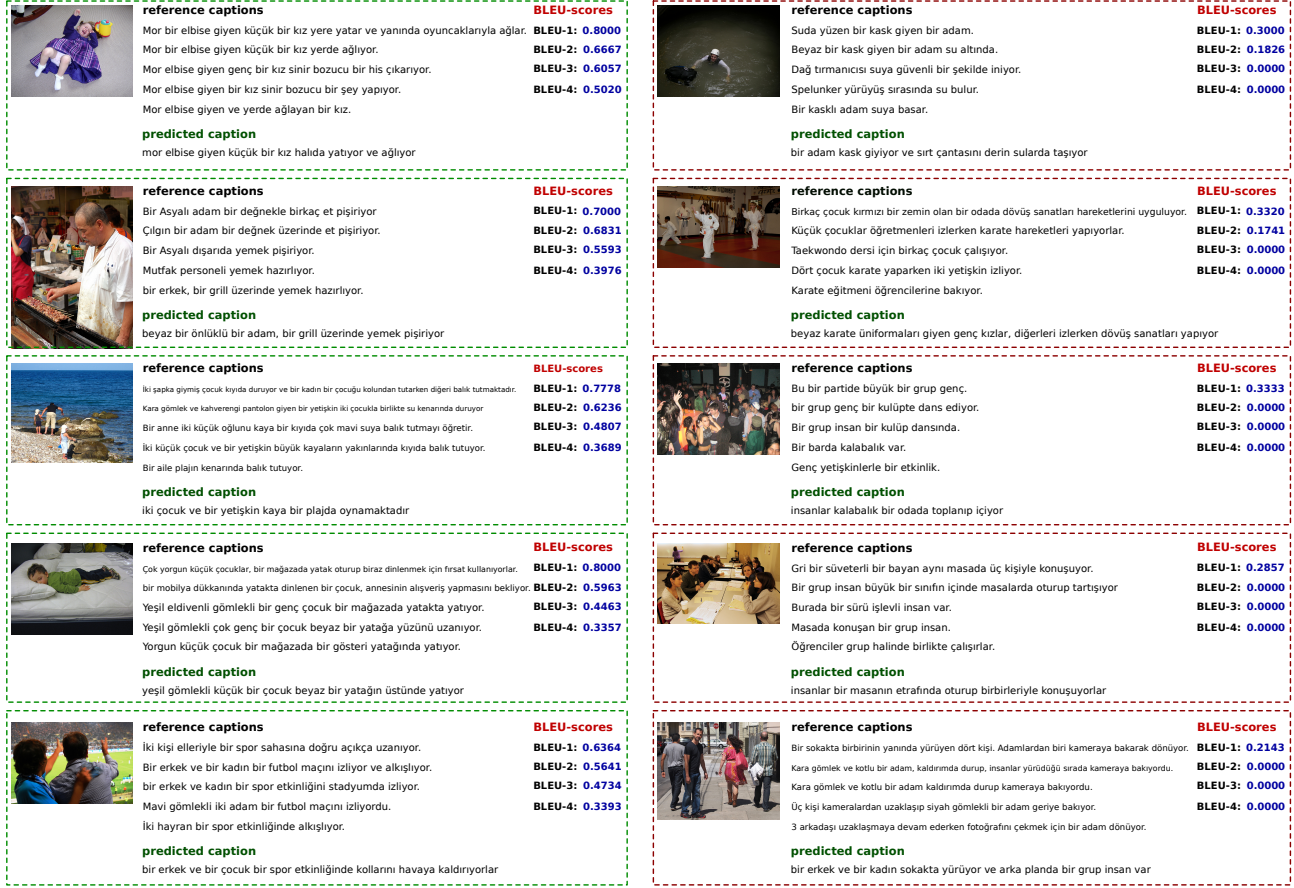


Figure 6. Caption prediction samples quantified as successful (left column) and unsuccessful (right column) according to the BLEU scores from Flickr30K dataset.

Table 5. A performance-based comparison of the related studies on Turkish image captioning.

| Related work | Training dataset | Test dataset | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L | CIDER |
|---------------------------------|-----------------------------|-------------------|--------|--------|--------|--------|--------|---------|--------|
| [48] by Samet et al. (M1) | MS COCO (80,000) | MS COCO (500) | 0.2820 | 0.1500 | 0.0730 | 0.0350 | 0.1410 | 0.2880 | 0.3600 |
| [48] by Samet et al. (M2) | MS COCO (80,000) | MS COCO (500) | 0.3130 | 0.1730 | 0.0850 | 0.0440 | 0.1540 | 0.3080 | 0.4500 |
| [48] by Samet et al. (M3) | MS COCO (80,000) | MS COCO (500) | 0.3420 | 0.1810 | 0.0850 | 0.0390 | 0.1570 | 0.3220 | 0.4610 |
| [48] by Samet et al. (F2) | Flickr30K (29,783) | MS COCO (500) | 0.1950 | 0.0770 | 0.0240 | 0.0110 | 0.0990 | 0.2070 | 0.0940 |
| [48] by Samet et al. (F2) | Flickr30K (29,783) | Flickr30K (1,000) | 0.3410 | 0.1920 | 0.1070 | 0.0530 | 0.1320 | 0.3040 | 0.2130 |
| [49] by Kuyu et al. (Word) | MS COCO (80,000) | MS COCO (500) | 0.2740 | 0.1480 | 0.0690 | 0.0330 | 0.1470 | 0.2920 | 0.4850 |
| [49] by Kuyu et al. (Subword) | MS COCO (80,000) | MS COCO (500) | 0.2930 | 0.1650 | 0.0880 | 0.0530 | 0.1470 | 0.3020 | 0.5670 |
| [49] by Kuyu et al. (Word) | Flickr30K (30,000) | MS COCO (500) | 0.1800 | 0.0750 | 0.0260 | 0.0120 | 0.0890 | 0.1900 | 0.0840 |
| [49] by Kuyu et al. (Subword) | Flickr30K (29,783) | MS COCO (500) | 0.2150 | 0.0890 | 0.0360 | 0.0190 | 0.1040 | 0.2200 | 0.1480 |
| [50] by Yılmaz et al. | MS COCO (83K) | MS COCO (41K) | 0.2880 | 0.1550 | 0.0710 | 0.0300 | 0.1250 | 0.2660 | 0.4790 |
| [51] by Yıldız et al. (Model 1) | MS COCO (83K) | MS COCO (41K) | 0.2880 | 0.1550 | 0.0710 | 0.0300 | 0.1250 | 0.2660 | 0.4790 |
| [51] by Yıldız et al. (Model 2) | MS COCO (83K) | MS COCO (41K) | 0.2970 | 0.1640 | 0.0760 | 0.0350 | 0.1290 | 0.2720 | 0.5280 |
| [54] by Golech et al. (M2) | MS COCO (82,783) | MS COCO (5,000) | 0.4990 | 0.4240 | 0.3130 | 0.2370 | 0.2500 | 0.4630 | 1.2570 |
| [54] by Golech et al. (M3) | MS COCO (82,783) | MS COCO (5,000) | 0.5620 | 0.4010 | 0.2990 | 0.2270 | 0.1930 | 0.4480 | 0.9800 |
| [54] by Golech et al. (M4) | MS COCO (82,783) | MS COCO (5,000) | 0.4370 | 0.3540 | 0.2750 | 0.2140 | 0.1700 | 0.4050 | 0.8990 |
| [54] by Golech et al. (M5) | MS COCO (82,783) | MS COCO (5,000) | 0.7230 | 0.5620 | 0.4130 | 0.2960 | 0.2550 | 0.5270 | 0.9350 |
| Proposed system (Best config) | MS COCO (113,287) | MS COCO (5,000) | 0.5773 | 0.4146 | 0.2839 | 0.1954 | 0.2546 | 0.4652 | 0.6552 |
| Proposed system (Best config) | MS COCO (113,287) | Flickr30K (1,000) | 0.5311 | 0.3587 | 0.2293 | 0.1395 | 0.2146 | 0.4192 | 0.3798 |
| Proposed system (Best config) | Flickr30K (29,000) | MS COCO (5,000) | 0.4213 | 0.2576 | 0.1518 | 0.0900 | 0.1933 | 0.3619 | 0.3145 |
| Proposed system (Best config) | Flickr30K (29,000) | Flickr30K (1,000) | 0.5477 | 0.3771 | 0.2535 | 0.1671 | 0.2271 | 0.4393 | 0.4668 |
| Proposed system (Best config) | MS COCO+Flickr30K (142,287) | MS COCO (5,000) | 0.5761 | 0.4124 | 0.2803 | 0.1905 | 0.2523 | 0.4609 | 0.6437 |
| Proposed system (Best config) | MS COCO+Flickr30K (142,287) | Flickr30K (1,000) | 0.5713 | 0.4056 | 0.2789 | 0.1843 | 0.2330 | 0.4491 | 0.5154 |

Table 6. A performance-based comparison of the related works on English image captioning.

| Related work | Dataset | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L | CIDEr |
|--|-----------|--------|--------|--------|--------|--------|---------|--------|
| [29] Locality-sensitive transformer network (LSTNet) | MS COCO | 0.8150 | - | - | 0.4030 | 0.2960 | 0.5940 | 1.3480 |
| [30] Multi-head association attention enhancement network (MAENet) | MS COCO | 0.7830 | 0.6220 | 0.4850 | 0.3830 | 0.2880 | 0.5800 | 1.2020 |
| [31] Prompt-based image captioning | MS COCO | - | - | - | 0.4050 | 0.3090 | - | 1.3370 |
| [32] Triple-steam feature fusion network (TSFNet) | MS COCO | 0.8170 | - | - | 0.4030 | 0.2980 | 0.5960 | 1.3350 |
| [32] Triple-steam feature fusion network (TSF-T) | MS COCO | 0.8190 | - | - | 0.4040 | 0.2980 | 0.5980 | 1.3360 |
| [33] Hybrid attention network (HAN) | MS COCO | - | - | - | 0.3740 | 0.2820 | 0.5810 | 1.2430 |
| [34] Position-guided transformer network (PGT) | MS COCO | 0.8130 | - | - | 0.3990 | 0.2950 | 0.5910 | 1.3420 |
| [35] Geometry attention transformer framework (GAT) | MS COCO | 0.8080 | - | - | 0.3970 | 0.2910 | 0.5900 | 1.3050 |
| [36] Sequential transformer framework (S-Transformer) | MS COCO | 0.8050 | 0.6540 | 0.5080 | 0.3890 | 0.2900 | 0.5880 | 1.2880 |
| [37] Contextual and selective attention network (CoSA-Net) | MS COCO | - | - | - | 0.3900 | 0.2900 | 0.5870 | 1.2950 |
| [38] Attention-reinforced transformer with contrastive learning (ArCo) | MS COCO | 0.8280 | - | - | 0.4140 | 0.3040 | 0.6040 | 1.3970 |
| [39] Multi-branch distance-sensitive self-attention network (MD-SAN) | MS COCO | 0.8150 | - | - | 0.3980 | 0.2960 | 0.5910 | 1.3510 |
| [40] Object semantic analysis for image captioning | MS COCO | - | - | - | - | 0.2990 | - | 1.1910 |
| [41] Joint relationship attention network (JRAN) | MS COCO | 0.8130 | 0.6470 | 0.5000 | 0.3860 | 0.2840 | 0.5840 | 1.2860 |
| [42] Scene graphs with transformer (SGT) | MS COCO | 0.8140 | - | - | 0.3980 | 0.2960 | 0.5920 | 1.3290 |
| [43] Hadamard product perceptron attention (HPPA) | MS COCO | 0.8090 | 0.6550 | 0.5130 | 0.3930 | 0.2910 | 0.5890 | 1.3050 |
| [44] Attentional long short term memory (ALSTM) | MS COCO | 0.7880 | 0.6380 | 0.5040 | 0.3950 | 0.2880 | 0.5840 | 1.2110 |
| [45] Semantic-conditional diffusion networks (SCD-Net) | MS COCO | 0.8130 | 0.6610 | 0.5150 | 0.3940 | 0.2920 | 0.5910 | 1.3160 |
| [46] Hierarchical aggregation of augmented views (HAAV) | MS COCO | - | - | - | 0.4100 | 0.3020 | - | 1.4150 |
| [29] Locality-sensitive transformer network (LSTNet) | Flickr30K | 0.6710 | - | - | 0.2330 | 0.2040 | 0.4430 | 0.6450 |
| [33] Hybrid attention network (HAN) | Flickr30K | - | - | - | 0.2990 | 0.2270 | 0.5030 | 0.6500 |
| [35] Geometry attention transformer framework (GAT) | Flickr30K | 0.7440 | 0.5670 | 0.4180 | 0.3080 | 0.2340 | - | 0.6800 |
| [40] Object semantic analysis for image captioning | Flickr30K | - | - | - | 0.3345 | 0.2470 | 0.5349 | 0.7757 |
| [41] Joint relationship attention network (JRAN) | Flickr30K | 0.7130 | 0.5350 | 0.3850 | 0.2830 | 0.2530 | 0.5350 | 0.5820 |
| [42] Scene graphs with transformer (SGT) | Flickr30K | 0.7760 | 0.6050 | 0.4630 | 0.3580 | 0.2510 | 0.5380 | 0.7850 |
| [43] Hadamard product perceptron attention (HPPA) | Flickr30K | 0.7240 | 0.5530 | 0.4150 | 0.3100 | 0.2290 | 0.5050 | 0.6610 |
| [44] Attentional long short term memory (ALSTM) | Flickr30K | 0.6830 | 0.4990 | 0.3550 | 0.2530 | 0.2080 | 0.4770 | 0.5560 |
| [46] Hierarchical aggregation of augmented views (HAAV) | Flickr30K | - | - | - | 0.3430 | 0.2510 | - | 0.8560 |

use in the task of Turkish image captioning. However, creating such datasets is quite labour-intensive and time-consuming. Instead of this, various operations, which can be defined as text postprocessing, can be performed on machine-translated caption sets. In this way, translation-based linguistic errors (such as grammatical errors and semantic transference) in Turkish image captions can be avoided. Therefore, it is thought that the accuracy level of the captions in Turkish generated by the TIC models will increase and the models will produce captions more successfully.

6. Conclusions

In this paper, we proposed a novel and accurate deep Turkish image captioning model, named TRCaptionNet, to address the challenge of automatic Turkish caption text generation. The TRCaptionNet essentially consists of a basic image encoder, a feature projection module based on vision transformers, and a text decoder. The model encodes the input images via the CLIP image encoder. Then, the model passes the CLIP image features to a vision transformer block which basically performs a feature projection operation. Finally, the text decoder generates the captions by using the image and caption features. We evaluated the performance of our model on two widely known image-caption datasets, MS COCO and Flickr30K. In addition, 8 distinct models (TRCaptionNet variants) were built by combining 4 image encoders, 1 vision transformer block, and 2 text decoders. Furthermore, we used a deep machine translation model, called NLLB, in this study to translate the image captions in English into Turkish, unlike the other related works on Turkish image captioning. Within the scope of the experiments, quite successful results were observed on MS COCO and Flickr30K image-caption datasets. As a result of the comparative performance analysis by taking the existing reports in the current literature on Turkish image captioning into consideration, it was observed that the proposed model outperforms most of the related works on Turkish image captioning. 0.5773 BLEU-1, 0.4146 BLEU-2, 0.2839 BLEU-3,

0.1954 BLEU-4, 0.2546 METEOR, 0.4652 ROUGE-L and 0.6552 CIDEr rates were observed for the MS COCO test split. In addition, 0.5477 BLEU-1, 0.3771 BLEU-2, 0.2535 BLEU-3, 0.1671 BLEU-4, 0.2271 METEOR, 0.4393 ROUGE-L, and 0.4668 CIDEr rates were measured for the Flickr30K test split. In our future works, we aim to improve the performance of the proposed system by employing further text-processing operations on machine-translated Turkish captions. Moreover, we intend to analyze the performance of TRCaptionNet on some Turkish caption subsets.

Acknowledgment

The authors declare that they have no conflict of interest. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. S.Y. gave the idea and did the experiments, S.Y. and A.M. conceptualized the experiments and wrote the manuscript, S.Y. and A.M. and S.V. interpreted the results, A.M. and S.V. supervised the study, A.M. reviewed and edited the manuscript.

References

- [1] Chen F, Li X, Tang J, Li S, Wang T. A Survey on Recent Advances in Image Captioning. *Journal of Physics: Conference Series* 2021; 1914 (1): 012053. <https://doi.org/10.1088/1742-6596/1914/1/012053>
- [2] Ghandi T, Pourreza H, Mahyar H. Deep Learning Approaches on Image Captioning: A Review. *arXiv preprint arXiv:2201.12944* 2022. <https://doi.org/10.48550/arXiv.2201.12944>
- [3] Ayesha H, Iqbal S, Tariq M, Abrar M, Sanaullah M et al. Automatic medical image interpretation: State of the art and future directions. *Pattern Recognition* 2021; 114: 107856. <https://doi.org/10.1016/j.patcog.2021.107856>
- [4] Pavlopoulos J, Kougia V, Androutsopoulos I, Papamichail D. Diagnostic captioning: A survey. *Knowledge and Information Systems* 2022; 64 (7): 1691-1722. <https://doi.org/10.1007/s10115-022-01684-7>
- [5] Makav B, Kılıç V. A new image captioning approach for visually impaired people. In: 2019 11th International Conference on Electrical and Electronics Engineering (ELECO); Bursa, Turkey 2019;45-949. <https://doi.org/10.23919/ELECO47770.2019.8990630>
- [6] Dognin P, Melnyk I, Mroueh Y, Padhi I, Rigotti M et al. Image captioning as an assistive technology: Lessons learned from VizWiz 2020 challenge. *Journal of Artificial Intelligence Research* 2022; 73: 437-459. <https://doi.org/10.1613/jair.1.13113>
- [7] Zhao B. A systematic survey of remote sensing image captioning. *IEEE Access* 2021; 9: 154086-154111. <https://doi.org/10.1109/ACCESS.2021.3128140>
- [8] Li Y, Fang S, Jiao L, Liu R, Shang R. A multi-level attention model for remote sensing image captions. *Remote Sensing* 2020; 12 (6): 939. <https://doi.org/10.3390/rs12060939>
- [9] Huang TH, Ferraro F, Mostafazadeh N, Misra I, Agrawal A et al. Visual storytelling. In: *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*; San Diego, California, USA 2016;233-1239. <http://dx.doi.org/10.18653/v1/N16-1147>
- [10] Fudholi DH, Windiatmoko Y, Afrianto N, Susanto PE, Suyuti M et al. Image captioning with attention for smart local tourism using EfficientNet. *IOP Conference Series: Materials Science and Engineering* 2021; 1077 (1): 012038. <https://doi.org/10.1088/1757-899X/1077/1/012038>
- [11] Mori Y, Hirakawa T, Yamashita T, Fujiyoshi H. Image captioning for near-future events from vehicle camera images and motion information. In: *2021 IEEE Intelligent Vehicles Symposium (IV)*; Nagoya, Japan 2021;378-1384. <https://doi.org/10.1109/IV48863.2021.9575562>

- [12] Zhang B, Zhou L, Song S, Chen L, Jiang Z et al. Image captioning in Chinese and its application for children with autism spectrum disorder. In: Proceedings of the 2020 12th International Conference on Machine Learning and Computing; Shenzhen, China; 2020. pp. 426-432. <https://doi.org/10.1145/3383972.3384072>
- [13] Li W, Qu Z, Song H, Wang P, Xue B. The traffic scene understanding and prediction based on image captioning. *IEEE Access* 2020; 9: 1420-1427. <https://doi.org/10.1109/ACCESS.2020.3047091>
- [14] Sathe S, Shinde S, Chorge S, Thakare S, Kulkarni L. Overview of Image Caption Generators and Its Applications. In: Bhalla S, Bedekar M, Phalnikar R, Sirsikar S (editors). *Proceeding of International Conference on Computational Science and Applications, Algorithms for Intelligent Systems*, Springer, Singapore 2021;105-110. https://doi.org/10.1007/978-981-19-0863-7_8
- [15] Sharma D, Dhiman C, Kumar D. Evolution of visual data captioning methods, datasets, and evaluation metrics: A comprehensive survey. *Expert Systems with Applications* 2023; 221: 119773. <https://doi.org/10.1016/j.eswa.2023.119773>
- [16] Farhadi A, Hejrati M, Sadeghi MA, Young P, Rashtchian C et al. Every picture tells a story: Generating sentences from images. In: *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision*; Heraklion, Crete, Greece 2010;15-29. https://doi.org/10.1007/978-3-642-15561-1_2
- [17] Mitchell M, Dodge J, Goyal A, Yamaguchi K, Stratos K et al. Midge: Generating image descriptions from computer vision detections. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*; Avignon, France 2012; 747-756.
- [18] Aneja J, Deshpande A, Schwing AG. Convolutional image captioning. In: *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*; Salt Lake City, UT, USA 2018; 5561-5570. <https://doi.org/10.1109/CVPR.2018.00583>
- [19] Feng Y, Ma L, Liu W, Luo J. Unsupervised image captioning. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; Long Beach, CA, USA 2019; 4125-4134. <https://doi.org/10.1109/CVPR.2019.00425>
- [20] He S, Liao W, Tavakoli HR, Yang M, Rosenhahn B et al. Image captioning through image transformer. In: *Proceedings of the 15th Asian Conference on Computer Vision (ACCV)*; Kyoto, Japan 2020;153-169. https://doi.org/10.1007/978-3-030-69538-5_10
- [21] Yang M, Liu J, Shen Y, Zhao Z, Chen X et al. An ensemble of generation-and retrieval-based image captioning with dual generator generative adversarial network. *IEEE Transactions on Image Processing* 2020; 29: 9627-9640. <https://doi.org/10.1109/TIP.2020.3028651>
- [22] Luo G, Cheng L, Jing C, Zhao C, Song G. A thorough review of models, evaluation metrics, and datasets on image captioning. *IET Image Processing* 2022; 16 (2): 311-332. <https://doi.org/10.1049/ipr2.12367>
- [23] You Q, Jin H, Wang Z, Fang C, Luo J. Image captioning with semantic attention. In: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; Las Vegas, NV, USA 2016; 4651-4659. <https://doi.org/10.1109/CVPR.2016.503>
- [24] Yao T, Pan Y, Li Y, Mei T. Exploring visual relationship for image captioning. In: *Proceedings of the 15th European Conference on Computer Vision (ECCV)*; Munich, Germany 2018; 684-699. https://doi.org/10.1007/978-3-030-01264-9_42
- [25] Herdade S, Kappeler A, Boakye K, Soares J. Image captioning: Transforming objects into words. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NIPS)*; Vancouver, BC, Canada 2019; 11137-11147.
- [26] Xu L, Tang Q, Lv J, Zheng B, Zeng X et al. Deep image captioning: A review of methods, trends and future challenges. *Neurocomputing* 2023; 546: 126287. <https://doi.org/10.1016/j.neucom.2023.126287>

- [27] Stefanini M, Cornia M, Baraldi L, Cascianelli S, Fiameni G et al. From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2022; 45 (1): 539-559. <https://doi.org/10.1109/TPAMI.2022.3148210>
- [28] Costa-jussà MR, Cross J, Çelebi O, Elbayad M, Heafield K et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672* 2022. <https://doi.org/10.48550/arXiv.2207.04672>
- [29] Ma Y, Ji J, Sun X, Zhou Y, Ji R. Towards local visual modeling for image captioning. *Pattern Recognition* 2023; 138: 109420. <https://doi.org/10.1016/j.patcog.2023.109420>
- [30] Hu N, Fan C, Ming Y, Feng F. MAENet: A novel multi-head association attention enhancement network for completing intra-modal interaction in image captioning. *Neurocomputing* 2023; 519: 69-81. <https://doi.org/10.1016/j.neucom.2022.11.045>
- [31] Wang N, Xie J, Wu J, Jia M, Li L. Controllable image captioning via prompting. *Proceedings of the AAAI Conference on Artificial Intelligence* 2023; 37 (2): 2617-2625. <https://doi.org/10.1609/aaai.v37i2.25360>
- [32] Hu N, Ming Y, Fan C, Feng F, Lyu B. TSFNet: Triple-stream image captioning. *IEEE Transactions on Multimedia* 2022; 1-14. <https://doi.org/10.1109/TMM.2022.3215861>
- [33] Jiang W, Li Q, Zhan K, Fang Y, Shen F. Hybrid attention network for image captioning. *Displays* 2022; 73: 102238. <https://doi.org/10.1016/j.displa.2022.102238>
- [34] Hu J, Yang Y, Yao L, An Y, Pan L. Position-guided transformer for image captioning. *Image and Vision Computing* 2022; 128: 104575. <https://doi.org/10.1016/j.imavis.2022.104575>
- [35] Wang C, Shen Y, Ji L. Geometry attention transformer with position-aware LSTMs for image captioning. *Expert Systems with Applications* 2022; 201: 117174. <https://doi.org/10.1016/j.eswa.2022.117174>
- [36] Wei Y, Wu C, Li G, Shi H. Sequential transformer via an outside-in attention for image captioning. *Engineering Applications of Artificial Intelligence* 2022; 108: 104574. <https://doi.org/10.1016/j.engappai.2021.104574>
- [37] Wang J, Li Y, Pan Y, Yao T, Tang J et al. Contextual and selective attention networks for image captioning. *Science China Information Sciences* 2022; 65 (12): 222103. <https://doi.org/10.1007/s11432-020-3523-6>
- [38] Wang Z, Shi S, Zhai Z, Wu Y, Yang R. ArCo: Attention-reinforced transformer with contrastive learning for image captioning. *Image and Vision Computing* 2022; 128: 104570. <https://doi.org/10.1016/j.imavis.2022.104570>
- [39] Ji J, Huang X, Sun X, Zhou Y, Luo G et al. Multi-branch distance-sensitive self-attention network for image captioning. *IEEE Transactions on Multimedia* 2022. <https://doi.org/10.1109/TMM.2022.3169061>
- [40] Du S, Zhu H, Lin G, Wang D, Shi J et al. Object semantic analysis for image captioning. *Multimedia Tools and Applications* 2023. <https://doi.org/10.1007/s11042-023-14596-7>
- [41] Wang C, Gu X. Learning joint relationship attention network for image captioning. *Expert Systems with Applications* 2023; 211: 118474. <https://doi.org/10.1016/j.eswa.2022.118474>
- [42] Li Z, Wei J, Huang F, Ma H. Modeling graph-structured contexts for image captioning. *Image and Vision Computing* 2023; 129: 104591. <https://doi.org/10.1016/j.imavis.2022.104591>
- [43] Jiang W, Hu H. Hadamard product perceptron attention for image captioning. *Neural Processing Letters* 2023; 55: 2707-2724. <https://doi.org/10.1007/s11063-022-10980-w>
- [44] Xiao F, Xue W, Shen Y, Gao X. A new attention-based LSTM for image captioning. *Neural Processing Letters* 2022; 54 (4): 3157-3171. <https://doi.org/10.1007/s11063-022-10759-z>
- [45] Luo J, Li Y, Pan Y, Yao T, Feng J et al. Semantic-conditional diffusion networks for image captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; Vancouver, BC, Canada; 2023. pp. 23359-23368. <https://doi.org/10.1109/CVPR52729.2023.02237>

- [46] Kuo CW, Kira Z. HAAV: Hierarchical aggregation of augmented views for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Vancouver, BC, Canada 2023; 11039-11049. <https://doi.org/10.1109/CVPR52729.2023.01062>
- [47] Unal ME, Citamak B, Yagcioglu S, Erdem A, Erdem E et al. TasvirEt: A benchmark dataset for automatic Turkish description generation from images. In: 2016 24th Signal Processing and Communication Application Conference (SIU); Zonguldak, Turkey 2016; 1977-1980. <https://doi.org/10.1109/SIU.2016.7496155>
- [48] Samet N, Hiçsönmez S, Duygulu P, Akbaş E. Could we create a training set for image captioning using automatic translation?. In: 2017 25th Signal Processing and Communications Applications Conference (SIU); Antalya, Turkey 2017; 1-4. <https://doi.org/10.1109/SIU.2017.7960638>
- [49] Kuyu M, Erdem A, Erdem E. Image captioning in Turkish with subword units. In: 2018 26th Signal Processing and Communications Applications Conference (SIU); İzmir, Turkey 2018; 1-4. <https://doi.org/10.1109/SIU.2018.8404431>
- [50] Yılmaz BD, Demir AE, Sönmez EB, Yıldız T. Image Captioning in Turkish Language. In: 2019 Innovations in Intelligent Systems and Applications Conference (ASYU); İzmir, Turkey 2019; 1-5. <https://doi.org/10.1109/ASYU48272.2019.8946358>
- [51] Yıldız T, Sönmez EB, Yılmaz BD, Demir AE. Image captioning in Turkish language: Database and model. Journal of the Faculty of Engineering and Architecture of Gazi University 2020; 35 (4): 2089-2100. <https://doi.org/10.17341/gazimmfd.597089>
- [52] Atıcı B, İlhan Omurca S. Generating Classified Ad Product Image Titles with Image Captioning. In: Trends in Data Engineering Methods for Intelligent Systems: Proceedings of the International Conference on Artificial Intelligence and Applied Mathematics in Engineering (ICAIAME 2020); Antalya, Turkey 2021; 211-219. https://doi.org/10.1007/978-3-030-79357-9_21
- [53] Ani Y, Amasyali MF. A General Purpose Turkish CLIP Model (TrCLIP) for Image&Text Retrieval and its Application to E-Commerce. In: 2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA); Biarritz, France 2022; 1-6. <https://doi.org/10.1109/INISTA55318.2022.9894123>
- [54] Golech SB, Karacan SB, Sönmez EB, Ayrıl H. A complete human verified Turkish caption dataset for MS COCO and performance evaluation with well-known image caption models trained against it. In: 2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME); Maldives, Maldives 2022; 1-6. <https://doi.org/10.1109/ICECCME55909.2022.9988025>
- [55] Lin TY, Maire M, Belongie S, Hays J, Perona P et al. Microsoft COCO: Common objects in context. In: 13th European Conference Computer Vision - ECCV 2014; Zurich, Switzerland 2014;740-755. https://doi.org/10.1007/978-3-319-10602-1_48
- [56] Young P, Lai A, Hodosh M, Hockenmaier J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics 2014; 2: 67-78. https://doi.org/10.1162/tacl_a_00166
- [57] Chen X, Fang H, Lin TY, Vedantam R, Gupta S et al. Microsoft COCO captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 2015. <https://doi.org/10.48550/arXiv.1504.00325>
- [58] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G et al. Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020 2021. <https://doi.org/10.48550/arXiv.2103.00020>
- [59] Schweter S. BERTurk - BERT models for Turkish (1.0.0). Zenodo 2020. <https://doi.org/10.5281/zenodo.3770924>
- [60] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 2018. <https://doi.org/10.48550/arXiv.1810.04805>
- [61] Han K, Wang Y, Chen H, Chen X, Guo J et al. A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence 2022; 45 (1): 87-110. <https://doi.org/10.1109/TPAMI.2022.3152247>

- [62] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Las Vegas, Nevada, USA; 2016. pp. 770-778. <https://doi.org/10.48550/arXiv.1512.03385>
- [63] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17); Long Beach, California, USA; 2017. pp. 6000–6010.
- [64] Sharif N, White L, Bennamoun M, Shah SAA. NNEval: Neural network based evaluation metric for image captioning. In: Proceedings of the 15th European Conference on Computer Vision (ECCV); Munich, Germany; 2018. pp. 37-53. https://doi.org/10.1007/978-3-030-01237-3_3
- [65] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02); Philadelphia, Pennsylvania, USA; 2002. pp. 311–318. <https://doi.org/10.3115/1073083.1073135>
- [66] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In: Proceedings of the Second Workshop on Statistical Machine Translation (StatMT'07); Prague, Czech Republic; 2007. pp. 228–231.
- [67] Lin CY. ROUGE: A package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004); Barcelona, Spain; 2004. pp. 74-81.
- [68] Vedantam R, Zitnick CL, Parikh D. CIDer: Consensus-based image description evaluation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Boston, MA, USA; 2015. pp. 4566–4575. <https://doi.org/10.1109/CVPR.2015.7299087>