Attention Based Image Captioning In Turkish Language

1st Ege Sendogan
Department of Computer Engineering
Galatasaray University
Istanbul, Turkey
https://orcid.org/0000-0002-4808-347X

***Abstract*—The task of automatically produce descriptions for images and videos is becoming increasingly popular in deep learning. In the field of image captioning, which is a study field of both computer vision and natural language processing, various studies have been carried out for common languages, especially in English. Despite the increasing number of studies, this challenging task is still open to improvement. For image captioning, models with attention mechanism have been produced by taking inspiration from object detection and machine translation models. In this study, an attention based encoder-decoder architecture is used to automatically produce Turkish descriptions for images. The quality of generated captions are evaluated with bilingual evaluation understudy (BLEU) metric and a BLEU-1 score of 0.418 is achieved.**

I. Introduction

Image captioning is the process of generating textual description of an image. Because of the fact that the relationship between extracted features of an image has to be revealed with a descriptive language in the image captioning task, it should be addressed as both computer vision and natural language processing problems.

We can divide the existing image captioning methods into three basic groups. These are template based image captioning, retrieval based image captioning and deep learning based image captioning.[9]

Template based methods are one of the oldest image captioning methods. In this method, there are a number of predetermined sentence structures for image descriptions. Captions are obtained by filling this template with inferences from the pictures. [9,10] For this reason, template based methods suffer from lack of flexibility. [11]

Retrieval based methods are also an old-fashioned method like template based methods. In this method, candidate captions are fetched from the caption database, taking into account the visual similarity, and a new caption is generated with the help of these candidate descriptions. For this reason, it is not a very creative method and objects in pictures can be misidentified.[9,10]

Among the image captioning methods, deep learning based methods are the most recent approaches and they generate the most successful captions [9]. The main ones of these methods are encoder-decoder architectures, and attention based models which are more advanced versions of encoder-decoder architectures. Encoder-decoder models consist of convolutional neural networks that extract visual features from input images and recurrent neural networks that generate captions using these features. [9,11]

With the powerful deep neural network architectures developed in recent years, many tasks in the field of computer vision such as image classification, face verification, face recognition, and object detection which is based on labelling present objects with predetermined words, handled successfully. Despite all these improvements, image captioning is still an intricate task for computer vision.

Besides that, image captioning applications can be useful for annotating large data sets and explaining a video frame by frame. Additionally, image captioning can be used in the tasks that textual and visual representations have to be combined such as multimodal classification, exploring hate speech detection in multimodal publications, multimodal author profiling and so on. Thus, this topic has attracted the attention of many researchers and several articles have been published on this topic in recent years.

Finally, although many data sets have been created and many studies have been done in this area so far, most of them have been made in English. Even in some of the most spoken languages in the world such as Hindu, there are few studies related to image captioning. [12] Unfortunately, there is a small number of studies in the Turkish language also [13]. Image captioning has many uses, as mentioned above. Therefore, the increase in Turkish language studies and the development of successful image captioning applications will bring diverse benefits.

II. Related Works

One of the pioneering work in this field is the study of Kiros et al. [14]. In this work they proposed two modality biased log-bilinear models as well as they showed that word representations can be learned along with image features by the virtue of convolutional neural networks. The language model that they proposed in this paper was a feed forward model.

Similar to the work of Kiros et al., Mao et al. proposed a model that learns image features and word representations together and generates words biased by these image features.

But in this work, they employed a recurrent neural language model instead of a feed forward language model.[15]

Donahue et al. developed an architecture which is applicable for video recognition, image description and retrieval tasks. In this architecture that they called Long-term Recurrent Convolutional Networks (LRCNs), they utilized convolutional neural networks in order to extract visual features from images, and a stack of long-short term memories (LSTM) which are fed with these features and then produce a description.[16]

A considerable number of different deep neural network architectures have been introduced for image captioning so far. Many of the models that have the best performances have a similar basic network designing: a convolutional neural network that generates features from a given image and a recurrent neural network that uses those generated features as the initial hidden state.

Deep Visual-Semantic Alignments for Generating Image Descriptions (Karpathy and Li [2014]) is one of the works based on this architecture. In this work, they firstly used a model in order to generate annotations for visual regions on an image. Then, they employed a multimodal recurrent network that produces descriptive captions from those annotated regions.

Besides that, Fang et al. proposed a new image captioning approach that consists of three phases. In first phase, they reveal a number of words from input images by the virtue of convolutional neural networks. Then, they have a language model that produces candidate sentences according to the words extracted in the first step. In the last phase, they re-rank these best candidate sentences in order to generate final description.[17]

In another project, K. Xu and Benagio (2014) employed a method that they denoted as "RNN with attention" after the convolutional feature extraction phase. The main purpose of the attention mechanism is to eliminate the problem of learning a single vector representation for each sentence that traditional encoder-decoder models suffer from.

Besides, they chose Bahdanau attention, as the attention mechanism in their model. With this technique, they weighted every location on the image according to degree of attention of those locations, thus, instead of treating every image as static representations, they made remarkable features in an image to come to the forefront.

Similarly, P. Anderson (2017) produced a model with the principles of attention mechanism in Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering paper. The model presented in this paper shows one of the best performance in image captioning task while it has a simpler architecture than aforementioned attention model.

As a final example of image captioning models that combines convolutional neural networks and recurrent neural networks, Sharma and Tan (2019) developed an experimental model, called "Specimen-Model". In this model they transferred input images into a pre-trained object detection model where they add an untrained fully connected layer in place of the last layer of that model. Prior to this CNN model, they employed a RNN model that takes the output of the fully connected layer as well as the embedding for the token as input/hidden states.

Most of the work in the field of image captioning has been done in English. The number of Turkish studies in this field is few. The number of data sets prepared for image captioning in Turkish language is even less. To solve this problem Unal et al. presented a benchmark dataset that they called "TasvirEt" for image captioning in Turkish language. This data set was created by collecting Turkish captions with crowdsourcing method to 8000 images in the Flickr8K data set. It contains two captions for each image (a total of 16000 captions). In this study, two different image captioning methods are also presented. In the first method, a caption is fetched from the caption database, taking visual similarity into account. The second method is based on finding the caption based on the root similarities of Turkish words.[18]

<start> Çiftlik ortamında genç beyaz kuzu yetişkin koyun. <end>
<start> Büyük bir kuzu bebek kuzusu yanında duruyor. <end>
<start> Çiftlikte bir bebek koyun yanında duran yetişkin bir koyun. <end>
<start> Çitlerin yanındaki çimenli bir alanda bir koyun ve kuzu. <end>
<start> Bir tarlada annesinin yanında duran bir kuzu. <end>

Fig. 1. An example image and its captions from the dataset created by Samet et al.

Similarly, Samet et al. created a benchmark dataset for image captioning in Turkish language. Instead of creating Turkish captions for the images of COCO dataset, they translated English captions of more than 160K images automatically by using Google's Translate API. Because of the fact that COCO is one of the largest image captioning datasets and has five captions for each image, the dataset created in this project is much bigger than TasvirEt dataset.

Besides, they trained and tested an Encoder-Decoder image captioning model with this dataset and obtained better results

compared to the results of the methods presented in TasvirEt project.[19]

One of the Turkish image captioning studies using an encoder-decoder based long-short term memories (LSTM) model is the work of Kuyu et al. In addition to the LSTM model, they also used a byte pair encoding (BPE) model in order to parse the words in the training set into sub words.[20]

Recently, Yıldız et al. proposed an encoder-decoder model that consist of convolutional neural networks and recurrent neural networks for image captioning in Turkish language. Besides, they also created a brand new dataset for Turkish image captioning by translating 616,767 captions in the COCO dataset with Yandex Translation API in approximately seven days.[13]

III. Methodology

A. Dataset

The number of data sets that can be used in Turkish image captioning studies is limited. According to current information, there are two publicly available image captioning datasets that can be used for this purpose. One of them is the dataset prepared by Unal et al. This dataset includes 8000 images from the Flickr8K dataset and two Turkish captions for each image (16000 captions in total). Turkish captions were prepared manually, not by any translator application. However, the size of the dataset is not enough for the successful operation of deep learning models and therefore it was not used in this project. [19]

Another publicly available dataset is created by Samet et al. Encouraged by the increasing success of Google Translate, which is an automatic translation application, they have translated all capions in MSCOCO and Flickr data sets, which are quite large data sets, into Turkish with Google Translate. As a result, they created a much bigger dataset for Turkish image captioning than TasvirEt dataset.Thus, this dataset is used also in this paper. An example image and its captions are shown in Figure1.

Besides that, images that belong to Microsoft Common Objects in Context (MS COCO) dataset were used in this project. MS COCO is a large-scale object detection, segmentation, and captioning dataset that contains everyday scenes with ordinary objects. This dataset contains 300K images, more than 200K of which are tagged. In addition, this dataset contains photos of 80 object, 91 stuff classes as well as it has 5 captions per image.[21]

B. Model

The architecture used in this project consists of three components:

- A convolutional neural networks based encoder
- A bahdanau attention mechanism
- A recurrent neural networks as the decoder unit

Each component has its own specific functions. Details regarding these functions and the implementation of components will be explained in this section.

1) The Encoder: Convolutional neural networks have been used frequently in image captioning studies since the emergence of the idea that word representations and image features can be learned together.

By the virtue of transfer learning, the power of InceptionV3 architecture which is pretrained with ImageNet dataset was used while creating the encoder unit in this project. Transfer learning is a deep learning technique that allows to improve a new learning task by transmitting the learned parameters from another model which is trained with a bigger dataset for a similar purpose. The reason for choosing InceptionV3 architecture is that it extracts features from images faster because it has fewer parameters compared to models such as VGG-16. In addition to its speed, it has also quite high accuracy.[22]

In order to use InceptionV3 as a feature extractor, the last convolutional layer of this architecture which has an output shape of 8x8x2048 is used. The last fully-connected layer is removed because, it serves for a different task (object classification) and has no contribution to the feature extraction.

Finally, a fully connected layer fed with feature vectors that has a shape of (64,2048) was used to complete the encoder unit. The shape (64,2048) is obtained by squeezing a feature vector of shape (8,8,2048) that InceptionV3 outputs.

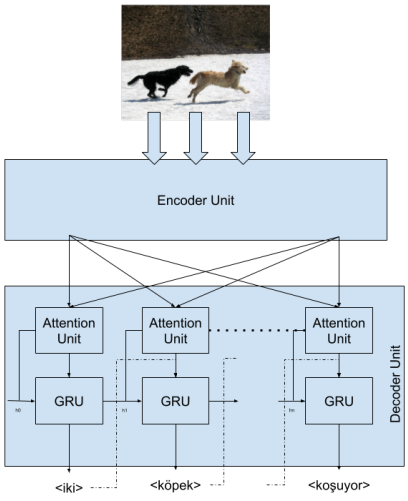

Fig. 3. Encoder-Decoder Model with Attention Mechanism

2) The Attention Mechanism: The classic image captioning models represent images as a single feature vector. Thus, the words generated by the language model describes only a portion of the image and the caption created doesn't capture the entire image's concept.[2] In order to being able to generate more meaningful captions by generating distinct words for various parts of the picture, attention mechanisms are used and better results are obtained than classical methods.

Attention is a very general process used for tasks such as machine translation, object recognition and captioning of images. An attention mechanism computes a weight for each location of the feature vector that the encoder unit outputs

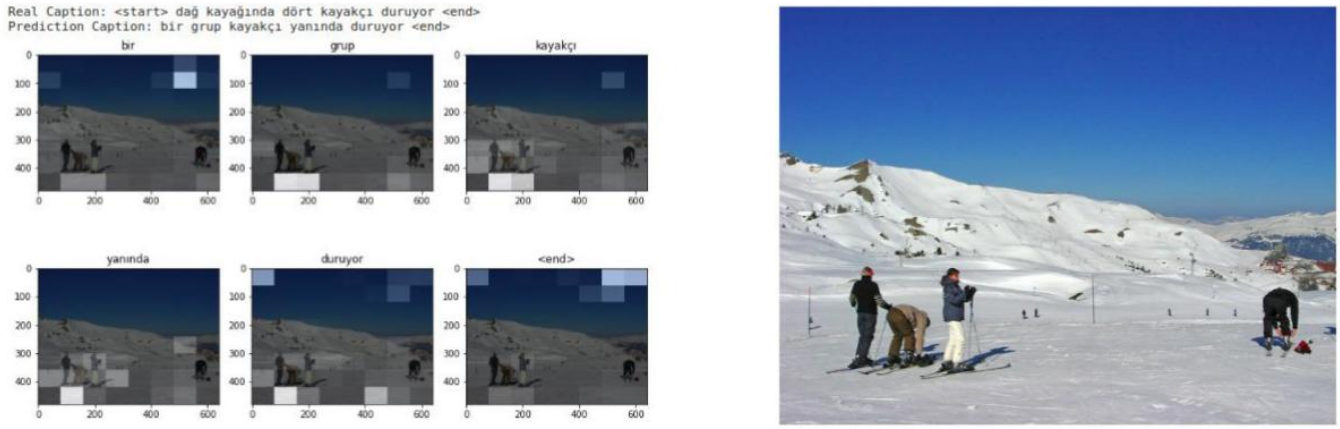

Fig. 2. A real caption and a generated caption for an image

which are also represent a real part of the image, and the decoder model can produce new words for the caption by focusing on only the relevant parts which are determined by these weights.

There are different types of attention mechanism, but the Bahdanau attention mechanism that is also known as "local attention" is used. The differences between Bahdanau and other attention mechanism are beyond the scope of this paper.

The attention unit in this project is a feed forward neural network that produces a context vector. In order to create a context vector, firstly it takes the hidden state of the encoder model and the previous hidden state of the decoder model and applies some linear and non-linear transformations on them.

Then, it computes the attention weights by applying a Softmax function on the unnormalized scores calculated after transformation processes. After the attention weights are obtained, it computes the context vector by using attention weights and the features extracted with the encoder. To be more precise, the context vector is the weighted sum of these features.

$$e_{ti} = f_{att}(s_{t-1}, h_j) \tag{1}$$

$$f_{att} = v_a^T * tanh(W_1 * h_j + W_2 * s_t) \tag{2}$$

$$\alpha_{jt} = \frac{e^{e_{jt}}}{\sum_{k=1}^{T_x} e^{e^{kt}}} \tag{3}$$

$$C_t = \sum_{j=1}^{T} \alpha_{jt} * h_j \tag{4}$$

Where s_t is the state of the decoder, h_j is the state of the encoder, α_{jt} is the vector of attention weights and C_t is the context vector.

3) The Decoder: In this project, the role of the decoder unit is to generate a word in each time step by using its previous hidden state, previously generated words which are represented with word embeddings, and the context vector which is the output of Bahdanau attention model. An example of a predicted caption for an input image is given in Figure 2.

In machine translation, image captioning and similar tasks, recurrent neural networks (RNNs) are used for producing texts. Because of the fact that RNNs allow the outputs they produce in each timestep to be used in future timesteps; they are different from feed forward neural networks.

Besides that, RNN units have hidden states that they can pass through other units and so, they pass the "information" through future time steps. Thus, recurrent neural networks become suitable for natural language processing tasks.

In order to generate image captions, a gated recurrent unit (GRU) network which is a variant of RNN is employed as the decoder model in this project.

The encoder, attention and decoder models are implemented with Keras framework. Keras is an open source neural network framework written in Python that uses the Tensorflow backend. Models are trained until 25 epochs while batch size was equal to 64. Training was stopped in this epoch as validation loss did not decrease after the 25th epoch. The change of the value of loss function can be shown in Figure 4.

Besides that, Adam algorithm which is an extension of stochastic gradient descent (SGD) is chosen as the optimization algorithm while learning rate was equal to 0,001. Fur-

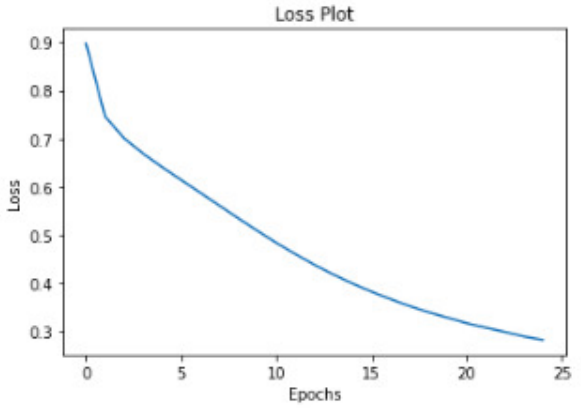

Fig. 4. The Change of the Value of Loss Function

thermore, sparse categorical cross entropy function of Keras framework that computes the cross entropy loss between the labels and predictions is employed as the loss function.

IV. Results

In order to measure the quality of captions generated by the decoder model, the Bilingual Evaluation Understudy (BLUE) method which is proposed by Papineni et al. in 2002 is used. The BLUE method computes a score between 0 and 1 for a generated sentence by comparing it with reference sentences. A BLEU score equal to 1 indicates that generated sentence completely matches with the reference sentences.

In this project, the implementation of the BLEU metric is provided using the Natural Language Processing Toolkit (NLTK) and individual 1-gram, 2-gram, 3-gram and 4-gram scores are calculated.

The main reasons for using BLEU evaluation metric in this project are that it has proven success and is frequently used in similar studies. Besides that, it is fast and simple to calculate, and most importantly it is language independent, so it can be used to evaluate produced image captions in Turkish language. The evaluation results are represented in Table 1:

TABLE I
Results

Metric	Dataset	Result
BLEU-1	MS COCO	0.418312
BLEU-2	MS COCO	0.238581
BLEU-3	MS COCO	0.172171
BLEU-4	MS COCO	0.084343

As seen in the table, the highest score is the BLEU-1 score, which is equal to 0.418. 3-gram and 4-gram BLEU scores are very low compared to 1 and 2-gram BLEU scores.

Unal et al., who carried out one of the first studies for Turkish image captioning, trained 2 different models using the images in the Flickr8K data set and their Turkish translations. They evaluated their models with BLEU-1, BLEU-2 and BLEU-3 scores and these scores are equal to 0.260, 0.102 and 0.034. Thus, it can be said that the attention based model presented in this paper outperformed the model of Unal et. Al.

In another study for Turkish image captioning, Samet et al. have developed various models and realised a number of experimental setup that they combined different models, and datasets such as MS COCO and Flickr. They obtained the best BLEU-1 score by training a model they call ”CNN + Kök Alma” with MS COCO dataset. This score was equal to 0.342. Similarly, they obtained best BLEU-2 and BLEU-3 scores with this experimental setup. These scores were equal to 0.181 and 0.085

In the most recent Turkish image captioning study, Yıldız et al. used two different encoder-decoder based models. Then, they trained them with the images of MS COCO dataset, and their Turkish captions that they obtained by translating original captions in Turkish by the virtue of Yandex Translation API. They the best BLEU-1, BLEU-2 and BLEU-3 scores they obtained were equal to 0.297, 0.164 and 0.076.

The result obtained in Turkish image captioning studies are much less than the result obtained in English studies. There are two important reasons for this. First of all, although the state-of-the-art models in image captioning employs attention mechanisms, in the studies realised for Turkish language, more traditional methods preferred so far. Recently, new approaches for image captioning that uses object detection models is emerged and these models obtained best results. In the future, these approaches should be also used in Turkish image captioning studies.

Besides that, because of the fact that to prepare a large dataset for image captioning is very troublesome and time consuming, all of the large Turkish datasets are prepared by using translator applications. Hence, those datasets are full of grammatical errors, inconsistent and inverted sentences This situation considerably reduces the success of the studies and prevents the production of sensible image descriptions.

V. Conclusion and Future Works

Image Captioning task is the target of both computer vision and natural language processing and is one of the most important study fields in deep learning. Because of the fact that it has a lot of applications and benefits, it has attracted the attention of researchers in recent years. Unfortunately, the number of image captioning studies in Turkish is quite low. Thus, the main purpose of this project is the contribute the Turkish image captioning studies.

In this project, a visual attention based encoder-decoder was used for Turkish image captioning, unlike existing studies. These models were created with convolutional neural networks and recurrent neural networks. The results obtained showed that an encoder-decoder model based on visual attention is more successful than other models offered for Turkish image captioning.

The Turkish captions in the data set were created by translating the captions in the MS COCO dataset into Turkish with the help of Google Translate in another study. For this reason, some captions contain meaningless or grammatical errors. Thus, the BLEU scores of some of the captions produced by

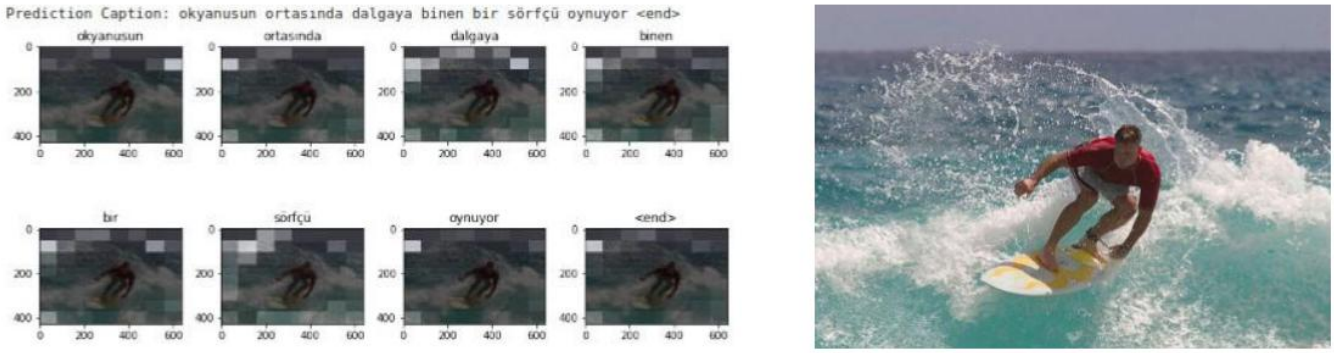

Fig. 5. A real caption and a generated caption for an image

the generated model were not bad, but these captions did not make much sense in terms of meaning.

For this reason, this project can be extended by creating a better dataset for turkish image captioning. This can be achieved in two ways:

- As in the previous study, by automatically translating the English captions to Turkish using a Translator API.
- By manually translating the English captions into Turkish

There were inconsistencies in the Turkish captions in the data set used in this project, but this data set was prepared 4 years ago. Within 4 years, google translate has improved greatly and now makes translations much more consistent. For this reason, a better data set can be obtained by converting the English captions in the MS COCO data set back to Turkish with the help of google translate again.

Another idea in order to extend this project, is the change the essence of image regions that the attention model performs on. To achieve this, a different visual feature extractor approach must be applied in the encoder unit. Each feature vector produced by InceptionV3 architecture, which is used as feature extractor in this project, represents a different local region of the image. At this stage, each image can be thought of as transformed into a grid of equal sized squares. Thus, the encoder unit employed in this project prevents the attention unit computes attention weights by focusing especially the precise location of objects and other important details on the image. To overcome this problem, an object detection architecture can be used as presented in "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering" paper of Anderson et al.

Object detection is a computer vision technique for detecting objects that belong to a certain class, and their exact positions on an image or video. There are several object detection architectures such as R-CNN, Fast R-CNN, Faster R-CNN, SSD, YOLO etc. proposed so far. Anderson et al. chose Faster R-CNN in their work and obtained great results which make their method state-of-the-art in image captioning task.

Hence, this project can be extended with a bottom-up and top-down mechanism that contains an object detection architecture such as YOLO or Faster R-CNN to generate Turkish captions

In addition, due to the limited time and the limited computational capabilities of the device used, less images were used when training models compared to other studies. In the future, you may achieve better results with a larger data set.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. CoRR,abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385

[2] R. K. K. C. A. C. C. R. S. R. S. Z. K. Xu, J. Ba and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. arXiv: 1502.03044, Apr. 2014. URL https://arxiv.org/pdf/1502.03044.pdf

[3] A. Karpathy and F. Li. Deep visual-semantic alignments for generating image descriptions. arXiv:1412.2306, Dec. 2014. URL https://cs.stanford.edu/people/karpathy/cvpr2015.pdf

[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In European Conference on Computer Vision, pages 740–755. Springer, 2014.

[5] C. B. D. T. M. J. S. G. L. Z. P. Anderson, X. He. Bottom-up and top-down attention for image captioning and visual question answering. arXiv: 1707.07998, Aug. 2017. URL https://arxiv.org/pdf/1707.07998.pdf

[6] Elania Tan, Lakshay Sharma: "Neural Image Captioning", 2019; [http://arxiv.org/abs/1907.02065].

[7] Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Joost van de Weijer: "Does Multimodality Help Human and Machine for Translation and Image Captioning?", 2016; arXiv:1605.09186. DOI: 10.18653/ v1/W16-2358

[8] • Caglayan, O., Madhyastha, P., Specia, L., amp; Barrault, L. (2019). Probing the Need for Visual Context in Multimodal Machine Translation. Proceedings of the 2019 Conference of the North. doi:10.18653/v1/n19-1422

[9] Hossain, M. Z., Sohel, F., Shiratuddin, M. F., amp; Laga, H. (2019). A Comprehensive Survey of Deep Learning for Image Captioning. ACM Computing Surveys, 51(6), 1-36. doi:10.1145/3295748

[10] Guan, Z., Liu, K., Ma, Y., Qian, X., amp; Ji, T. (2018). Sequential Dual Attention: Coarse-to-Fine-Grained Hierarchical Generation for Image Captioning. Symmetry, 10(11), 626. doi:10.3390/sym10110626

[11] Liu, S., Bai, L., Hu, Y., amp; Wang, H. (2018). Image Captioning Based on Deep Neural Networks. MATEC Web of Conferences, 232, 01052. doi:10.1051/matecconf/201823201052

[12] Dhir, R., Mishra, S. K., Saha, S., amp; Bhattacharyya, P. (2019). A Deep Attention based Framework for Image Caption Generation in Hindi Language. Computación Y Sistemas, 23(3). doi:10.13053/cys-23-3-3269

[13] Sonmez, E. B., Yildiz, T., Yilmaz, B. D., amp; Demir, A. E. (2020). Türkçe dilinde görüntü altyazısı: Veritabanı ve model. Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi. doi:10.17341/gazimmfd.597089

[14] Ryan Kiros, Ruslan Salakhutdinov, and Richard Zemel. 2014. Multimodal neural language models. In Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (ICML'14). JMLR.org, II–595–II–603.

[15] Mao, Junhua Xu, Wei Yang, Yi Wang, Jiang Yuille, Alan. (2014). Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN).

[16] Donahue, Jeff Hendricks, Lisa Guadarrama, Sergio Rohrbach, Marcus Venugopalan, Subhashini Darrell, Trevor Saenko, Kate. (2015). Long-term recurrent convolutional networks for visual recognition and description. 2625-2634. 10.1109/CVPR.2015.7298878.

[17] Fang, Hao Gupta, Saurabh Iandola, Forrest Srivastava, Rupesh Deng, li Dollar, Piotr Gao, Jianfeng He, Xiaodong Mitchell, Margaret Platt, John Zitnick, C. Zweig, Geoffrey. (2015). From captions to visual concepts and back. 1473-1482. 10.1109/CVPR.2015.7298754.

[18] Unal, Mesut Citamak, Begum Yagcioglu, Semih Erdem, Aykut Erdem, Erkut Ikizler, Nazli Cakici, Ruket. (2016). TasvirEt: A benchmark dataset for automatic Turkish description generation from images. 1977-1980. 10.1109/SIU.2016.7496155.

[19] N. Samet, S. Hiçsönmez, P. Duygulu and E. Akbaş, "Could we create a training set for image captioning using automatic translation?," 2017 25th Signal Processing and Communications Applications Conference (SIU), Antalya, 2017, pp. 1-4, doi: 10.1109/SIU.2017.7960638.

[20] M. Kuyu Et Al. , "Altsözcük Öğeleri ile Türkçe Görüntü Altyazılama (Image Captioning in Turkish with Subword Units)," 26. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SIU 2018) , İzmir, Turkey, pp.1-4, 2018

[21] Lin TY. et al. (2014) Microsoft COCO: Common Objects in Context. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, Cham. https://doi.org/10.1007/978-3-319-10602-1_48

[22] Katariya, Y. (n.d.). Image Captioning using InceptionV3 and Beam Search. Retrieved January 09, 2021, from https://yashk2810.github.io/Image-Captioning-using-InceptionV3-and-Beam-Search/

[23] Sarkar, S. (2020, March 07). Image Captioning using Attention Mechanism. Retrieved January 09, 2021, from https://medium.com/swlh/image-captioning-using-attention-mechanism-f3d7fc96eb0e