REMOTE SENSING IMAGE CAPTIONING

Ömer Talha BAYSAN, Tolga SAĞLAM
Computer Science and Engineering
Yıldız Technical University, 34220 Istanbul, Türkiye
{talha.baysan, tolga.saglam}@std.yildiz.edu.tr

***Özetçe* —Görüntü özetleme dijital bir görüntünün özelliklerini bularak o görüntü için bir açıklama üreten derin ögrenme yöntemidir. Yapılan çalışmalardan farklı olarak daha doğru ve anlamlı açıklamalar üretmek için geliştirilmiştir. Derin öğrenme yöntemlerinin ve dikkat mekanizmasının bir kombinasyonu, daha tutarlı ve kapsamlı açıklamalar sağlamak için görüntülerin özelliklerini daha iyi anlamak ve yorumlamak için kullanılmaktadır.**

Anahtar Kelimeler—görüntü özetleme, derin öğrenme, açıklama üretme, dikkat mekanizması.

***Abstract*—Remote Sensing Image Captioning is a deep learning method that finds features of a digital image and produces a description for that image. Unlike studies, it has been developed to produce more accurate and meaningful explanations. A combination of deep learning methods and attention mechanism is used to better understand and interpret the features of images to provide more coherent and comprehensive explanations.**

Keywords—image captioning, deep learning, explanation generation, attention mechanism.

I. INTRODUCTION

It uses a combination of technologies from different fields such as image captioning, natural language processing and artificial neural networks, which is a method of extracting subtitles by looking at the characteristics of a digital image at hand. Remote sensing image captioning is the processing of images containing geographic information obtained by vehicles such as drones, airplanes or helicopters. The information obtained from such images has great potential in many fields such as search and rescue, reconnaissance, traffic control.

They are images taken from above, called "God's perspective". In the general working logic, firstly, the properties of the photographs in the data set are determined. Different applications such as Res-Net, InceptionV3, MobileNet can be used in this process. Each contains multiple layers of artificial nerves that are different and interconnected. By specifying the properties, this information is given to the decoder (eg LSTM, RNN) and "start token" is given as the first word. Next, the attention mechanism determines what the next word will be based on the previous word and its attributes [1].

II. RELATED WORK

Remote sensing images are images acquired by remote sensing systems such as satellites or aircraft. These images are usually large in size and high resolution. Systems that can automatically annotate remote sensing images are being developed to facilitate the analysis of these images and to help people better understand the data.

Some work on annotating remote sensing images includes:

- "DeepCaption: A Deep Learning-Based Remote Sensing Image Captioning Framework" (2018): In this study, a framework is proposed for automatic annotation of remote sensing images using deep learning methods. Convolutional neural networks and recurrent neural networks were used to match images and text descriptions[2].
- "Remote Sensing Image Captioning with Hierarchical Attention Network" (2019): In this study, a method for describing remote sensing images using a hierarchical attention network is proposed. Attention mechanisms are used to highlight different features and parts of the images[3].
- "A Review on Remote Sensing Image Captioning: Techniques, Datasets, and Challenges" (2020): This study examines the techniques, datasets, and challenges encountered in annotating remote sensing images. The study discusses current methods and future research directions in this field[4].

These studies are considered as an important step in the field of automatic disclosure of remote sensing images. However, research on the subject is developing rapidly, and it is important to remember that newer and more advanced methods are emerging.

III. THE METHOD

Within the scope of the project, a CNN-RNN model was developed to describe remote sensing images. VGG-16 and ResNET architectures were used as the CNN component and the UCM Captions dataset was adapted to these architectures. Standard RNN, LSTM, BiLSTM and GRU architectures are used for the RNN component because these architectures are efficient in processing serial data such as text. The Adam algorithm was used for optimization and the Accuracy metric was used to measure the success of the model. In addition, the overall success of the model was evaluated using more advanced metrics such as BLEU and METEOR.

A. Data Set Preparation

Within the scope of the project, the UCM-Captions[5] dataset was used. Since remote sensing images are used, other general data sets are not suitable for our project. In the UCM-Captions dataset, there are 2100 remote sensing images and 5 abstracts of each image. Although the dataset is not large in size, all images were used.

B. CNN

CNN (Convolutional Neural Network) is a type of artificial neural network widely used in the field of deep learning. It has shown great success in the analysis of visual data such as image processing, pattern recognition and object detection.

CNNs have a special architecture used to learn the properties of the data. Unlike traditional neural networks, the layers of CNNs are optimized to better reflect the structure and properties of the image data[6].

The basic building blocks of a CNN usually consist of successive layers.

- Input Layer: Provides input of data to CNN. It can be in the form of a tensor represented by pixel values for image data.
- Convolutional Layers: A certain filtering process is applied on the image. These filters are used to highlight features in the image. Each convolution layer can contain many filters (kernels) and each is trained to capture different features.
- Activation Layers: An activation function is applied to the feature maps calculated after the convolution. Common activation functions such as Sigmoid or ReLU can be used.
- Pooling Layers: Help preserve important features while reducing the size of feature maps. With operations such as max pooling or average pooling, features are summarized and reduced in size.
- Fully Connected Layers: Feature maps are fed into traditional neural network layers used for classification or output. These layers do classification or regression using attributes.

CNNs use back propagation algorithm to train their weights and when trained on large datasets they can perform well in complex image processing tasks. It is widely used in areas such as image classification, object detection, face recognition, car driving.

C. RNN

RNN (Recurrent Neural Network) is a type of artificial neural network designed for processing serial data. RNNs work effectively on data types with features such as time dependency and sequential computing. They are widely used for analysis and prediction of serial data such as text, music, language.

The basic idea of RNNs is to incorporate information from previous steps into the current processing step. This gives RNNs the ability to remember past links and predict future steps, thanks to their memory.

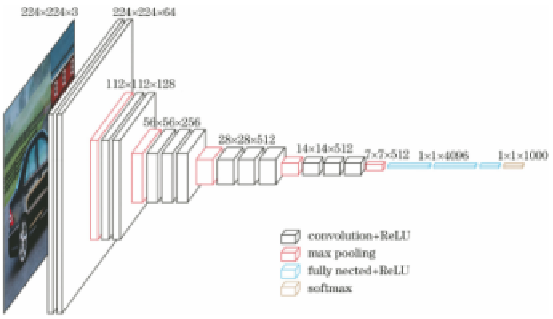

Figure 1 VGG_16 Architecture

The architecture of RNNs is different from traditional neural networks. Each step has a hidden state vector, this state vector carries and updates information from previous steps. This update is accomplished using an activation function and feedback links. These feedback links allow the network to access historical information and use the information from the current step on future steps.

There are several different types of RNNs, the most common being:

- Standard RNN (vanilla RNN): It is the simplest form of RNN. At each step, the latent state vector is updated using the current input and previous state vector.
- LSTM (Long Short-Term Memory): It is a type of RNN designed to learn about long-term dependencies. It controls the flow of information using memory cells and control mechanisms.
- GRU (Gated Recurrent Unit): Similar to LSTM, it is a type of RNN designed to better handle long-term dependencies. GRU has a simpler structure and can run faster than LSTM.

RNNs are an effective tool for modeling and processing serial data. They are used in many application areas such as text creation, language prediction, translation, speech recognition.

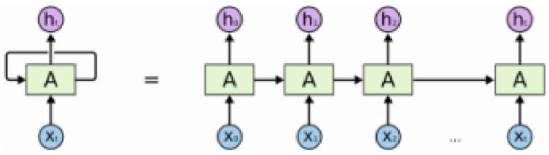

Figure 2 RNN Architecture

D. Attention Mechanism

The attention mechanism is a technique used in deep learning models and is particularly effective in processing sequential data. The attention mechanism allows the model to focus on specific areas and give more weight to important information.

The attention mechanism is used to identify relationships between different components or features within a data set.

For example, in a translation model used to translate a sentence, the attention mechanism may focus on the input words associated with the word being translated.

The attention mechanism is often used in encoder-decoder architectures. The encoder produces a vector representing the input data, while the solver produces the output using this vector. The attention mechanism allows the decoder to "pay attention" to the encoder output at each step and determine which encoder features are important.

The attention mechanism usually includes operations such as weighted sum and soft maximum. The weighted sum multiplies each encoder output of the solver by the weights to form a weighted sum vector. The soft maximum allows the decoder to focus on the encoder outputs with the highest weight.

The attention mechanism makes deep learning models more flexible and effective. It is widely used in areas such as natural language processing, text translation, image recognition and automatic captioning. The attention mechanism helps the model produce more accurate and meaningful outputs by directing its attention to the components it focuses on.

IV. Measuring System Success

A. BLEU Score

The BLEU (Bilingual Evaluation Understudy) score is a metric or benchmark used to evaluate the quality of machine translation systems. BLEU is based on comparing the translated text with human-generated reference translations.

The BLEU score uses the concepts of accuracy and n-gram similarity to measure how well a text translation is. The n-gram similarity between two texts measures how well the n-grams in the translated text overlap with the n-grams in the reference translation.

BLEU uses a series of n-gram precision values to measure the performance of the translation system. Typically, the BLEU score ranges from 1 to 5, with 1 representing the worst and 5 representing the best performance.

When calculating the BLEU score, the n-gram similarity between the texts produced by the translation system and the reference translations is taken into account. These similarities are then expressed as a fraction and a BLEU score is calculated by taking the geometric mean.

The BLEU score is widely used to compare the performance of translation systems, evaluate model improvements, and as a metric used in language translation research. However, the BLEU score also has some limitations and may not fully reflect the quality of a stand-alone translation system. Therefore, it is important to make a more comprehensive assessment by using the BLEU score in conjunction with other metrics.

B. METEOR Score

The METEOR (Metric for Evaluation of Translation with Explicit ORdering) score is a metric or metric used to evaluate the quality of text translation systems. METEOR also takes word order into account when comparing the outputs of translation systems with human reference translations.

The METEOR score matches the words in the translation system's output with the reference translation and calculates a score based on the similarity measure. The similarity measurement takes into account factors such as word overlap, word order, and word choice.

METEOR calculates a similarity score through a series of alignment and matching steps by comparing the text produced by the translation system with the reference translation. These scores are then obtained using a formula to get a METEOR score.

The METEOR score ranges from 0 to 1, similar to the BLEU score, with 1 representing the best performance. METEOR is considered a widely used metric in the field of language translation and is often used to evaluate the performance of translation systems.

METEOR provides a more comprehensive assessment of the translation, taking into account not only word level, but also word order and similarity. Therefore, METEOR score is used in conjunction with other metrics to assess the quality of translation systems, supporting more reliable results.

V. CONCLUSIONS

This project aimed to generate annotations using the properties of digital images using a deep learning method called image captioning. A combination of various methods has been used to provide significant improvements in the image identification field.

The project aimed to produce more accurate and meaningful explanations from an image with the combination of attention mechanism and other deep learning networks. The attention mechanism allowed the model to focus on specific features and highlight important details, helping to obtain more comprehensive explanations.

This study developed a method that performed better than previous studies. Evaluations using certain success metrics have shown that the developed method gives better results and produces more accurate explanations than other methods.

Image captioning is an exciting area of research at the intersection of computer vision and natural language processing, and this project has been a significant step in that area. The developed method helped us better understand the information contained in the images and produce more consistent explanations.

The successes of this project have led to new developments in the field of image captioning and inspired further research in the future. The automatic understanding and annotation of images offers great potential in many application areas, and this study has made significant contributions to the work done in this area.

References

[1] B. Zhao. (2021) A systematic survey of remote sensing image captioning. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9615088

[2] W. X. Junhua Mao. (2018) Deepcaption: A deep learning-based remote sensing image captioning framework. [Online]. Available: https://arxiv.org/abs/1412.6632

[3] M. Z. Haochen Li, Tian Wang and A. Zhu. (2019) Remote sensing image captioning with hierarchical attention network. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8997347

[4] B. M. Rita Ramos. (2022) A review on remote sensing image captioning: Techniques, datasets, and challenges. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9714367

[5] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195.

[6] A. Z. Karen Simonyan, "Very deep convolutional networks for large-scale image recognition," 2014.