Image Captioning via Vision and Language Transformers

Luke Davidson - [davidson.lu@northeastern.edu](mailto:davidson.lu@northeastern.edu)

Kishore Pagidi - [pagidi.k@northeastern.edu](mailto:pagidi.k@northeastern.edu)

Nand Dave - [dave.na@northeastern.e](mailto:dave.na@northeastern.edu)du

### *Abstract*

*With today’s rise in artificial intelligence applications, the relationship between computer vision and natural language processes has been an area of great interest. Image captioning is a task that combines aspects of computer vision and natural language processing to automatically generate descriptive captions of input images. Image captioning techniques are used in a wide range of applications, including visual media, virtual assistants, and accessibility domains.*

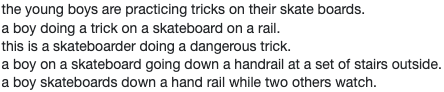
*In this project, we developed an end-to-end image captioning model using vision and language transformers capable of generating descriptive captions for a range of input images. We utilized transformer architecture based computer vision and natural language processing methods to build and train a model capable of generating these accurate captions in unobserved images. This model was constructed by implementing components such as image patch descriptors and encoders, decoders, multi-head attention layers and transformers to decompose and find the relationship between images and training captions. We were successfully able to develop a vision transformer and train our network to generate captions blah blah blah.*

### Introduction

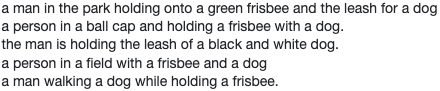
The relationship between computer vision and natural language processing applications, two major areas of deep learning, is one that is greatly increasing in popularity, research and implementation today due to its many potential use cases and benefits. Image captioning methods are a perfect combination of these two areas of deep learning, combining object detection and classification methods directly with sentence formation methods to create descriptive captions. As previously stated, one of the primary reasons for the increase in popularity of image captioning models is its many potential use cases and benefits. Automated image captioning methods are used in industries such as marketing, editing, and assistance services to complete tasks such as the automated captioning of social media posts, recommendation systems for image and video editing, and assisting visually impaired individuals with describing a visual of a live view or still image [1]. These use cases increase aspects of an individual's life, such as efficiency in their occupation or the general quality and simplicity of their everyday life. Image captioning methods and models have proven to be extremely efficient in these domains and are only becoming more accurate and robust [1].

In this study, we developed a transformer-based model to identify objects and corresponding object locations in images and match those objects and locations with descriptive captions. Transformer-based deep learning models have been widely used to solve natural language processing problems but have recently been more implemented in computer vision domains as well. A great strength of transformer-based deep learning models is their ability to encode feature location descriptions with detected features [1], whether that be the location of a word in a sentence or the relative location of an object in an image. This is an extremely important aspect when considering which model architecture to use to solve an image captioning problem, as overall scene understanding is a crucial aspect to generating an accurate, descriptive caption of an image. We chose a transformer-based model architecture primarily because of this benefit, implementing methods such as image patch extractors and descriptors, feature encoders and decoders, and multi-head attention layers, further explained below in the *Approach* section.

We used the Common Objects in Context (COCO) dataset [2] to train our transformer-based image captioning model. The COCO dataset was created for applications such as scene understanding and segmentation and aims to solve three core research problems seen in common image datasets: detecting non-iconic views (or non-canonical perspectives) of objects, contextual reasoning between objects, and the precise 2D localization of objects [3]. The dataset contains roughly 328,000 images of over 90 objects in their natural environment. Most importantly, the dataset also contains roughly 4-6 captions for each image, with over 2.5 million total captions describing key objects, object localizations and scene environments. Two examples of images and corresponding captions from the COCO dataset can be seen below in Figures 1 and 2.

*Figure 1: Example image and captions from COCO dataset*



*Figure 2: Example image and captions from COCO dataset*

### Background

As previously stated, there has been a large increase in the popularity of image captioning and image captioning like models that relate computer vision and natural language processing applications in recent years due to today’s advances in artificial intelligence. Transformer based architectures have been the most popular deep learning architecture to solve problems like this due to their ability to encode feature descriptors and efficiently work with large amounts of data.

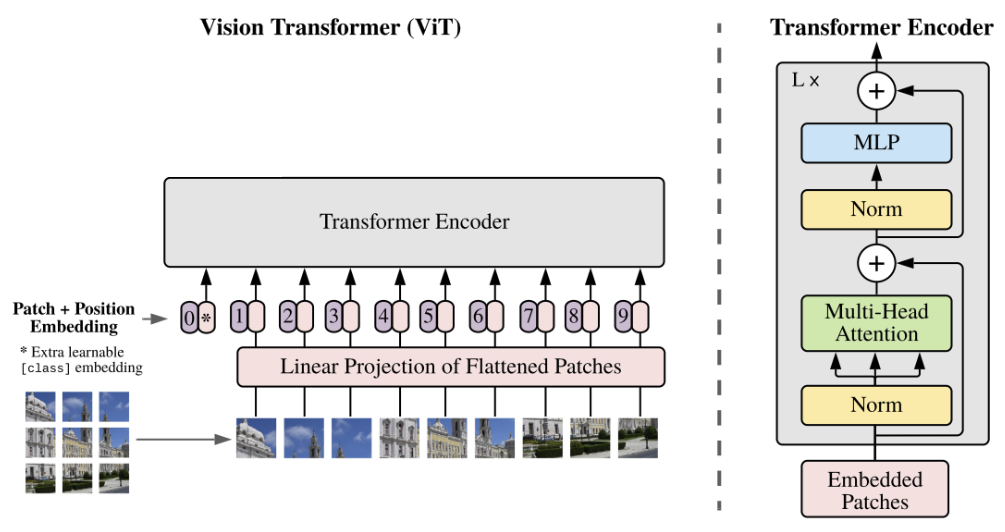
The popular paper *“An Image is Worth 16x16 Words: Transformers for Image Recognition At Scale”* [4] displays these strengths through comparing an implementation of a vision transformer with typical high performance neural networks such as ResNet and EfficientNet. The vision transformer, which uses 16x16 pixel subsections of images to pass to the transformer architecture, outperforms the neural networks when pre-trained with at least 100 million training images in just a fraction of the computational power, while maintaining the key benefit of being able to encode feature locations due to the use of the pixel patch extractors. Another extremely important aspect that makes a transformer-based architecture efficient in computer vision and natural language processing domains is the use of attention. An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key [4]. The paper *“Attention is All You Need”* [5] highlights the benefits of attention by proposing a new simple network architecture based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Yiyu Wang, Jungang Xi, and Yingfei Sun also introduce an alternative to the popular R-CNN backbone architecture used in many image captioning models in their paper *“End-to-End Transformer Based Model for Image Captioning”* [6]. Their approach replaces the R-CNN backbone encoder with a SwinTransformer to extract grid-level features. This method includes a refining encoder and mean pooling of grid features and addresses the limitations of pre-trained CNNs, object detectors in the encoder and LSTM in the decoder, using a Swin-Transformer encoder and MSA to determine the intra-relationship of image grid features. We have built off of these previous studies to implement our own variation of an end-to-end vision and caption transformer, further explained in the following *Approach* section.

### Approach

As previously stated, we chose to develop a transformer-based network architecture to create our image captioning model. Our implementation comprises four main components: a data pipeline, a vision transformer, a combined encoder and a combined decoder.

The data pipeline is a crucial first step to ensure the images and captions downloaded from the COCO dataset are in the correct format to optimize our model. Our data pipeline consists of four main steps: downloading the raw data, grouping captions to their corresponding images, tokenizing the captions, and transforming the images. The raw dataset of ~328,000 images and ~2.5 million captions is downloaded from the web. Each image is then paired with its corresponding list of ~4-6 captions. Next, each caption is tokenized and split into its key words, phrases and components. This step is crucial to the network performance as the individual caption features and locations are utilized in the encoder. Finally, the images are preprocessed by being resized to a certain dimension optimized for the vision transformer and normalized.

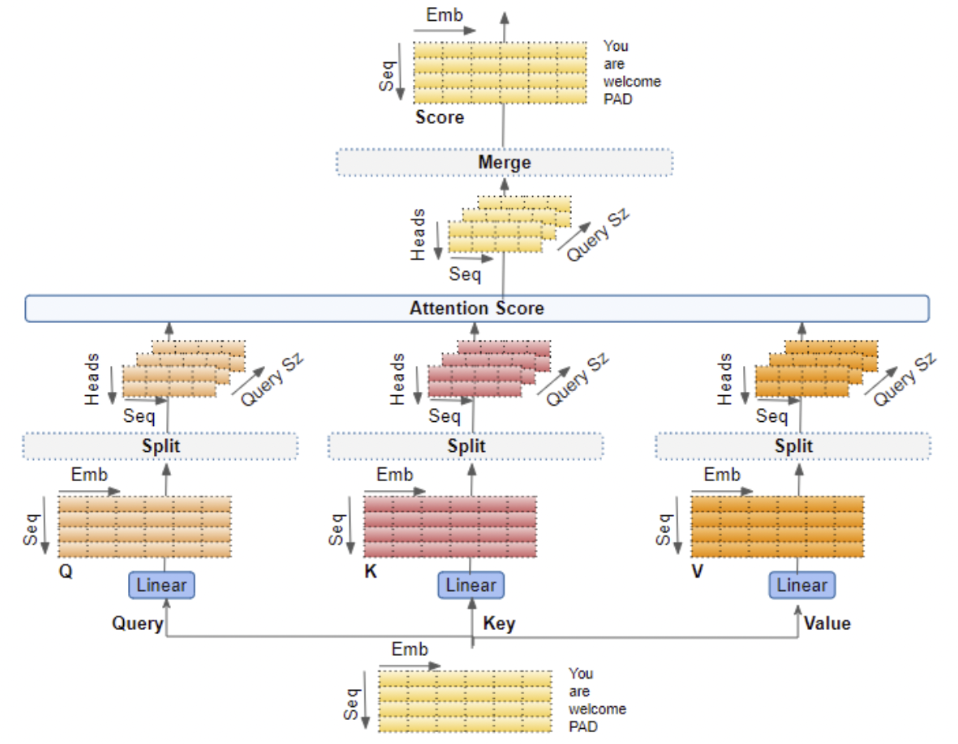
Once the data is prepared, the first step of the image network is implementing the vision transformer, or ViT. The general architecture of the vision transformer can be seen below in Figure 3:



*Figure 3: Vision Transformer architecture*

The image is first split into 196 16x16 pixel patches. Each patch is flattened into a tensor of size 1x768 (3x16x16) during the patch embedding process, and is concatenated with the positional embedding of the patch location in the image. These final embeddings are passed to the main ViT encoder to create a learned encoded representation of each patch embedding [7]. This ViT encoder contains an architecture of attention layers, normalization layers and MLP layers, all of which we experimented with and are further described in the following *Results* section. The final output of the ViT is passed to the main transformer encoder.

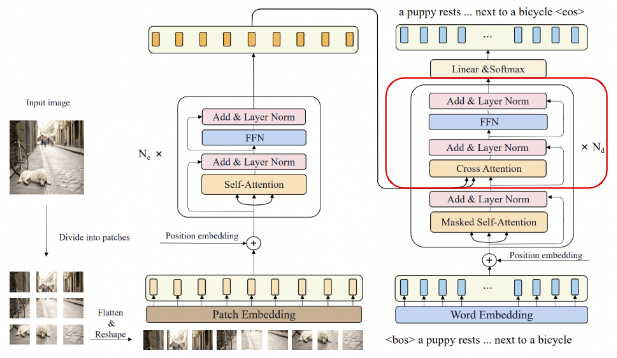
The second, and equally as important, aspect of the encoder is the caption encoder. This part of the encoder is carried out primarily using multi-head attention operation, shown in Figure 4:



*Figure 4: Graphical representation of Multi-Head Attention*

Each token of the caption is treated as an individual aspect of the caption and is concatenated with a positional encoding based on where it is in the sentence, similar to the ViT. These encodings are then fed the learnable parameters of the multi-head attention method, represented by Query, Key and Value in Figure 4, which then produces an encoded representation for each word in the input sequence, that now incorporates the attention scores for each word as well [7].

Once the encoded caption and image representations are calculated through the attention layers, they are passed to the decoder block, shown highlighted in red in Figure 5:



*Figure 5: Entire Image Captioning Transformer Architecture*

The most important method of the decoder is the Cross Attention layer of the network. This layer receives a representation of both the input image and target caption and produces a final representation with the attention scores for each generated caption token that also encapsulates the influence of the input image attention. This representation is then passed through another set of attention, normalization and MLP layers, and finally through a linear and softmax layer to create a final output.

### Results (~2-3 pages)

* Describe dataset specifics
* Experiments and performance evaluation
* Graphs, plots, images/figures needed to support results and conclusions
* Discussion
  + Conclusions from results
  + Recommended future directions

### Conclusion

Deep learning applications that combine aspects of computer vision and natural language processing methods, such as image captioning systems, will continue to increase in efficiency, accuracy and use in the near future due to today’s increase in artificial intelligence studies and use cases. Many individuals and organizations rely on the efficiency and accuracy of these models to complete tasks and increase factors such as job performance and overall quality of life. Transformer-based deep learning architectures have proven to be the leading methods for applications that combine these two industries. In this study, we have shown that a transformer-based deep learning model containing aspects such as a vision transformer, language encoder and decoder blocks can be extremely effective at tasks combining aspects of computer vision and natural language processing. Furthermore, altering the architecture of attention, normalization and MLP layers within the transformer architecture can greatly affect the overall output.

### 

### References

[1] https://www.ripublication.com/ijaer18/ijaerv13n9\_102.pdf

[2] https://cocodataset.org/#home

[3] https://arxiv.org/pdf/1405.0312.pdf

[4] https://arxiv.org/pdf/2010.11929.pdf

[5] https://arxiv.org/pdf/1706.03762.pdf

[6] https://arxiv.org/pdf/2203.15350.pdf

[7] https://towardsdatascience.com/transformers-explained-visually-part-3-multi-head-attention-deep-dive-1c1ff1024853#:~:text=Multiple%20Attention%20Heads,independently%20through%20a%20separate%20Head

### Contributions

Kishore Pagidi

Luke Davidson

* Developed main decoder architecture
* Contributed to report and presentation slides

Nand Dave

*All members contributed equally*