End-to-end Gated Multi-layer Perceptron for Image Captioning

Jay Ahn[1] and Tae Kim[2]

[1]Department of Computer Science & Software Engineering, California Polytechnic State University
[2]Department of Computer Engineering, California Polytechnic State University

June 11, 2022

Abstract

Recently, a gated multi-layer perceptron that performs as powerful as transformers in image classification and masked language modeling has been proposed. In this paper, we propose two end-to-end gMLP encoder-decoder architecture for image captioning, We train the gMLP-based image captioning models with attention and without attention separately with MSCOCO and Flickr8k and report their BLEU scores on their test-sets. Our proposed end-to-end gMLP models outperform a transformer-based model for both MSCOCO and Flickr8k datasets and demonstrate their strength in image captioning.

1 Introduction

Image captioning is a task that combines computer vision and natural language processing, where it aims to generate descriptive legends for images. It is a two fold process relying on accurate image understanding and correct language understanding both syntactically and semantically. Most image captioning systems use an encoder-decoder framework, the input image is given to the Convolutional Neural Network(CNN) to extract the features. Then the encoded output from the encoder process through the last state of the CNN which is connected to the decoder. The decoder process does language modelling up to the word level using Recurrent Neural Network(RNN). To simply put the input image is encoded into an intermediate representation of the information in the image and then decoded into a descriptive text sequence. Additionally, to train the model MSCOCO and flicker8k datasets were chosen since it was readily available as well as both datasets contained image and caption data. To analyze and compare the yields, two of the latest model was utilized which is

called Transformer, a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output. The Transformer architecture allows for significantly more parallelization and can reach new state of the art results in translation quality.

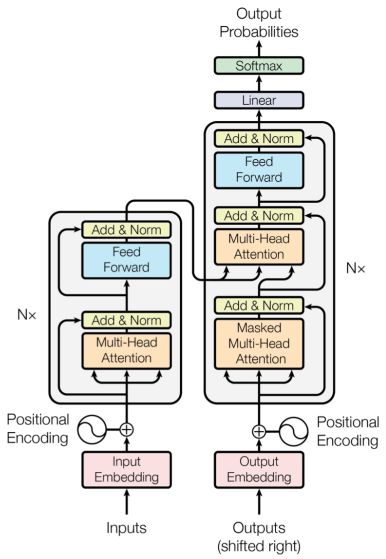

Figure 1: Transformer Architecture

In addition, Gated Multi-Layer Perceptron (gMLP) is a deep learning model that contains only basic multi-layer Perceptrons. Using fewer parameters, gMLP outperforms transformer models on natural-language processing (NLP) tasks and achieves comparable accuracy on computer vision (CV) tasks. The development of various deep neural network architectures such as sequence-to-sequence model [9] or transformers [10] has led to huge success in both computer vision and natural language processing. And, recently, gMLP [7] has been proposed as a novel architecture that performs as well as transformers on key nlp and vision problems without the need of self-attention that requires high memory usage.

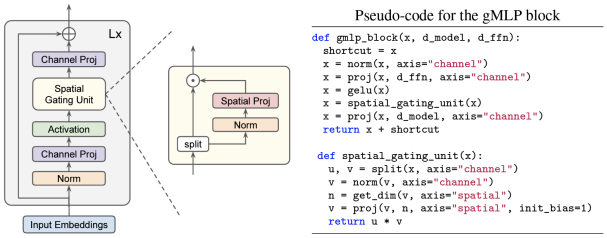

Figure 2: gMLP Architecture

2 Related Work

Proposed approaches for image captioning can be largely divided into two categories: compositional framework and end-to-end encoder-decoder model.

2.1 Compositional Framework

The compositional framework involves object detection, word-image alignment, and multi-modal similarity model to generate descriptions of images. Karpathy and Fei-Fei [5] have utilized combination of convolutional neural networks and bidirectional recurrent neural networks for word-image alignment model and adopted a multimodal recurrent neural network to generate descriptions of input image. Fang et al. [1] have included visual detectors that detect a set of words commonly found in image captions, generated descriptions using language model, and reranked the generated descriptions with Deep Multimodal Similarity Model (DMSM).

2.2 Encoder Decoder Architecture

An encoder-decoder architecture has been widely used for image captioning. Some people have utilized convolutional neural network for image encoder and recurrent neural networks for the decoder, while others have adopted transformers that consist of encoder and decoder as themselves. Also, a few researchers have proposed image captioning with a pre-trained image classification model such as VGG16 as a backbone to encode image.

2.2.1 Recurrent Neural Network (RNN)-based Image Captioning

Recurrent neural networks are popular approach to deal with sequential data. Therefore, it has been adopted for image captioning. Vinyals and et al. [11] have utilized a deep CNN to encode images and simple RNN to generate their captions. The success of simple recurrent neural network in generating sequential outputs led to more complex and powerful variants such as gated recurrent unit (GRU) and long short-term memory (LSTM) that resolve issues of diminishing gradients, especially with the long sequence of inputs. Accordingly, Patwari and Naik [8] have proposed an encoder-decoder model where they use CNN for encoder and GRU with attention for decoder. Also, Guan and Wang [2] have demonstrated inception-v3 or resnet for encoder and GRU for decoder. Furthermore, [12] have utilized a deep CNN network for encoder and a bidirectional LSTM for decoder. Our image captioning model utilizes gMLP for both encoder and decoder for image captioning as gMLP can be used to encode image as well as generate text data in an autoregressive fashion.

2.2.2 Transformer-based Image Captioning

Since recurrent neural networks receive sequential data one by one, it takes long time to train them. To fix this issue, transformers have been introduced

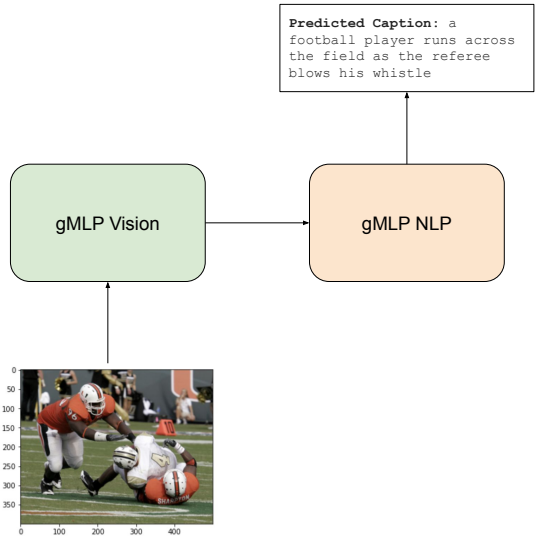

Figure 3: Overall end-to-end gMLP for Image Captioning.

with self-attention that is parallelizable and faster to train. Researchers around the world have adopted this powerful architecture to train image captioning model. Herdade et al. [4] have passed features extracted from a CNN-based object detector to transformer architectures to generate image captions. Li et al. [6] have proposed a way of using transformer for image captioning with their custom attention called EnTangled Attention. He et al. [3] have introduced image captioning using image transformer. Our image captioning model utilizes gMLP that is proposed to work comparably with transformers on nlp and vision tasks for image captioning.

3 System Development

We propose two end-to-end gMLP architectures for image captioning. These architectures follow high-level design of encoder decoder architecture where they use gMLP-vision for encoder and autoregressive gMLP-NLP for decoder as described in Figure 3. The first architecture just adds the image features from encoder to the text features from a gMLP block and generates captions, while the second architecture follows a transformer architecture in a sense that it passes the summed features to another gMLPBlock and generates captions. These proposed architectures are described in Figure 4, Figure 5.

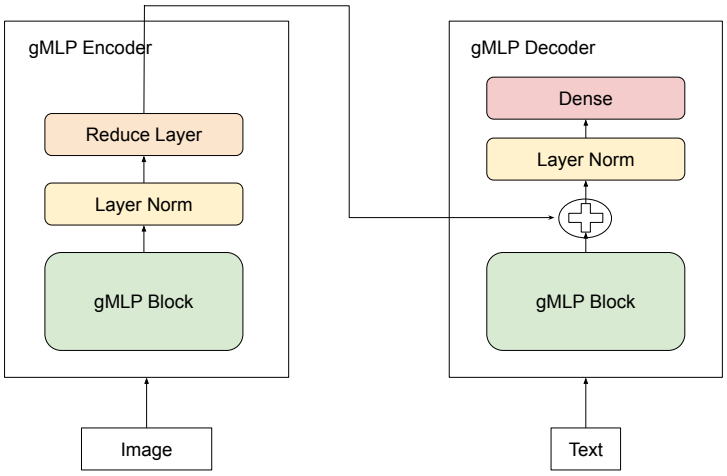

Figure 4: Simple end-to-end gMLP for Image Captioning.

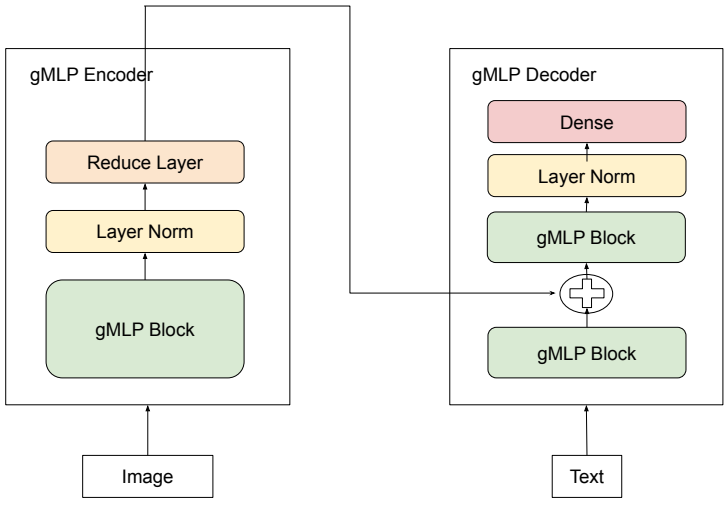

Figure 5: Transformer-like end-to-end gMLP for Image Captioning.

4 Experimental Design

To test the proposed architectures, we adopt MSCOCO and Flickr8k benchmark datasets. We use 60% of data for training, 20% for validation, and 20% for testing and compare their BLEU-4 scores with transformers'. We measure BLEU-4 scores with 100 images of the testing dataset due to the hardware limitation. Also, we add tiny attention proposed in original gMLP paper [7] to our architectures and measure their performance as well. Each model has a single decoder and encoder layer. For hyperparameters, we use a single head, d_{model} of 512, and d_{ffn} of 2048. For transformers, we adopt a pretrained InceptionV3 as a backbone, while we do not utilize pretrained encoder for gMLP models. We train each model for about 20 epochs until its validation accuracy gets plateaued.

5 Results

Model	Dataset	
	MSCOCO	Flickr8k
Transformer	47.903	43.602
gMLP-simple	50.029	42.948
gMLP-trans	**51.484**	43.347
aMLP-simple	50.606	44.711
aMLP-trans	50.477	**48.516**

Table 1: BLEU scores of different image captioning models on MSCOCO and flickr 8k. Bold text describes the best model for each dataset.

As demonstrated in Table 1, transformer-like gMLP model performs the best for MSCOCO and transformer-like aMLP model performs the best for Flickr8k. Also, gMLP models achieve comparable or better BLEU scores than transformer on each dataset for image captioning. The sample output captions for images from the datasets are further demonstrated in section 8.

6 Conclusion

Likewise, the proposed gMLP architectures for image captioning achieve better or similar performance with transformer on MSCOCO and Flickr8k even without backbone. Our experiment further supports that gMLP works as powerful as transformer for vision and NLP tasks with more efficient memory allocation. However, in the experiments, we noticed the models' overfitting on the training dataset which might have unintentionally affected the results. Also, in terms of training and inference time, gMLP took longer than transformers.

7 Future Work

In the future, we hope to integrate our gMLP decoder with pretrained image encoder such as VGG16 as a backbone for image captioning. Also, we would like to use techniques such as data augmentation to reduce overfitting and train each model. Furthermore, we would like to test our model on entire test dataset for accurate validation.

References

[1] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. *CoRR*, abs/1411.4952, 2014.

[2] Jinning Guan and Eric Wang. Repeated review based image captioning for image evidence review. *Image Commun.*, 63(C):141–148, apr 2018.

[3] Sen He, Wentong Liao, Hamed R. Tavakoli, Michael Yang, Bodo Rosenhahn, and Nicolas Pugeault. Image captioning through image transformer. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

[4] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *CoRR*, abs/1906.05963, 2019.

[5] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306, 2014.

[6] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[7] Hanxiao Liu, Zihang Dai, David R. So, and Quoc V. Le. Pay attention to mlps, 2021.

[8] Nikhil Patwari and Dinesh Naik. En-de-cap: An encoder decoder model for image captioning. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1192–1196, 2021.

[9] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[11] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014.

[12] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. Image captioning with deep bidirectional lstms. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, page 988–997, New York, NY, USA, 2016. Association for Computing Machinery.

8 Appendix

Figure 6: Transformer on mscoco

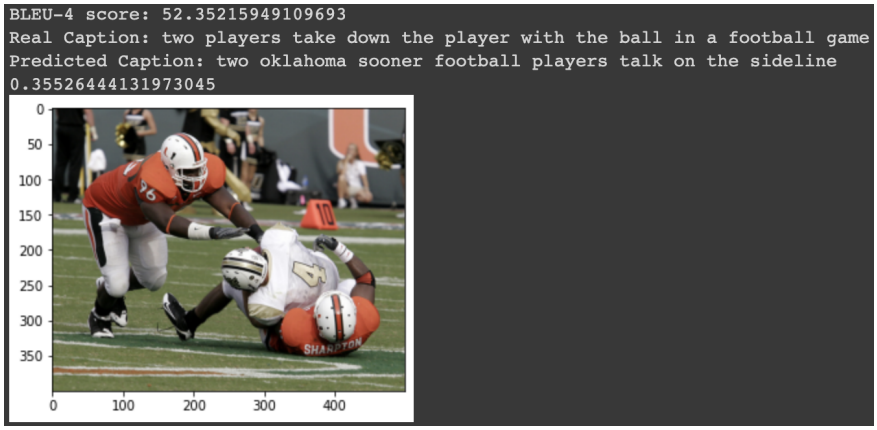

Figure 7: Transformer on flicker8k

```
BLEU-4 score: 68.65890479690393
Real Caption: a piece of luggage empty up against a wall.
Predicted Caption: a jockey is racing a horse through the water
0.6865890479690392
```

Figure 8: Transformer like gMLP on mscoco

```
BLEU-4 score: 48.54917717073234
Real Caption: the large wave is crashing down over a surfer
Predicted Caption: a surfer is falling into a big ocean wave
/usr/local/lib/python3.7/dist-packages/nltk/translate/bleu_score.py:490: UserWarning:
Corpus/Sentence contains 0 counts of 3-gram overlaps.
BLEU scores might be undesirable; use SmoothingFunction().
  warnings.warn(_msg)
0.5400734869541141
```

Figure 9: Transformer like gMLP on flicker8k

```
BLEU-4 score: 45.91497693322865
Real Caption: a boy that is jumping a skateboard on a ramp
Predicted Caption: a woman is sitting on a surfboard in the ocean.
0.4591497693322865
```

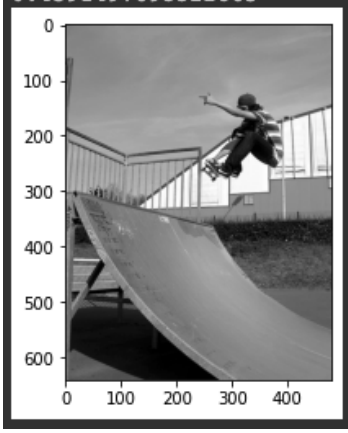

Figure 10: Simple gMLP on mscoco

```
BLEU-4 score: 39.2814650900513
Real Caption: dog with open mouth and a toy close by in the air
Predicted Caption: a brown and white dog is jumping up to catch a ball in the park
/usr/local/lib/python3.7/dist-packages/nltk/translate/bleu_score.py:490: UserWarning:
Corpus/Sentence contains 0 counts of 3-gram overlaps.
BLEU scores might be undesirable; use SmoothingFunction().
  warnings.warn(_msg)
0.28552132491142446
```

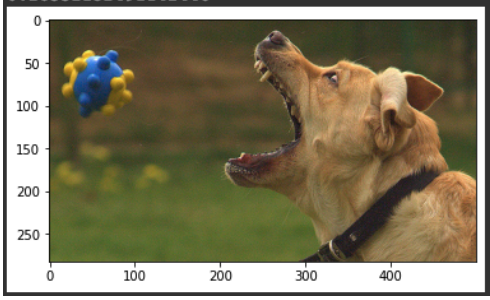

Figure 11: Simple gMLP on flicker8k

Figure 12: Transformer like aMLP on mscoco

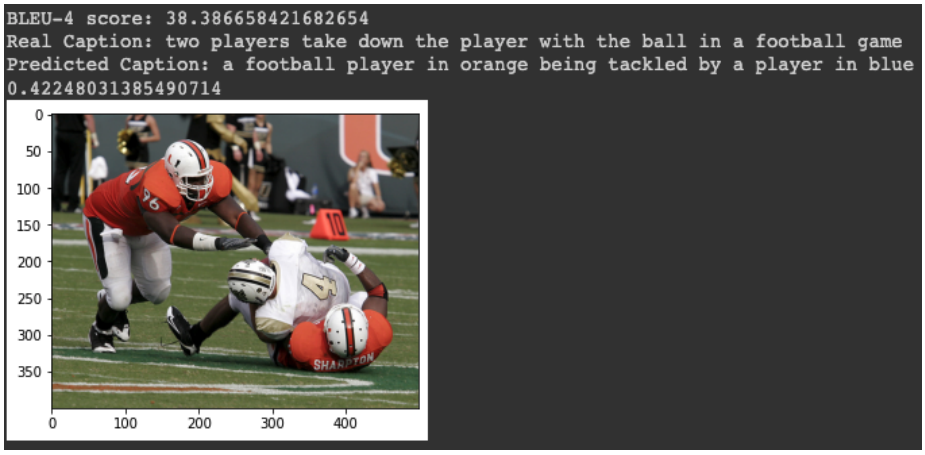

Figure 13: Transformer like aMLP on flickr8k

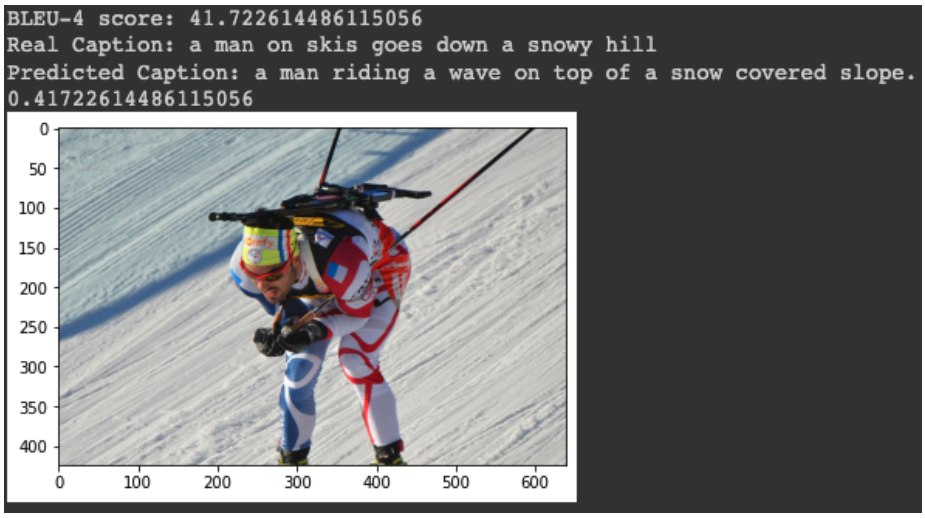

Figure 14: Simple aMLP on mscoco

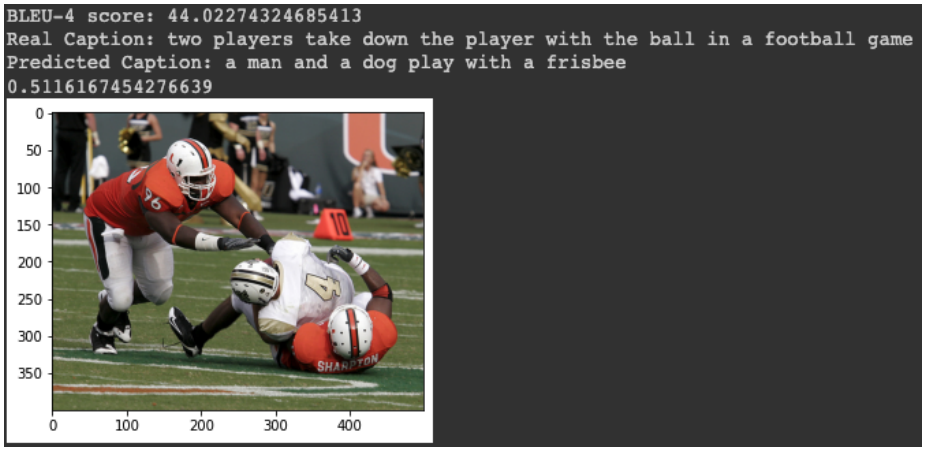

Figure 15: Simple aMLP on flickr8k