IET Research Journals

Research Article

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

Natural Language Processing and Object Detection with Transformer

Prithvi Raj Singh[1*]

[1] *Center for Advance Computer Studies, University of Louisiana at Lafayette, 70506, Lafayette, United States*
[2] *Department of Mechanical Engineering, University of Louisiana at Lafayette, Lafayette, US*
* *E-mail: prithvi.singh1@louisiana.edu*

Abstract: Still widely used as the state-of-art neural model for NLP work, RNN with sequence-to-sequence learning was replaced only recently by Transformer, an attention-based model. Transformers dominate the research market of NLP and now even for the computer vision task. As transformers tend to be computationally superior to RNN, transformer, and attention-based models will define the future research space in NLP. In this paper, we seek to implement Transformer primarily for the NLP task, focused on Sentiment Analysis, and Language Translation. We performed language translation for English-to-French, English-to-Spanish, Portuguese-to-English. Our general transformer does a decent performance in language translation, and we get acceptable BLEU score, accuracy, and F1 measure for all three language translations. Our sentiment analysis task focused on question answering (Q-A), text classification, and Named Entity Recognition (NER). For the Q-A we can very exact match score of 78%, we got consistent training accuracy of 40% for the NER task. The hugging face transformer API performs wildly better than any present models for NLP tasks. We discuss the recently developed Switch Transformer, MobileViT. Our implementation task for Switch Transformer and MobilViT wasn't successful as they have huge GPU power. We perform ViT tasks for image classification on CIFAR-10 with a test accuracy of 70% and top-5 accuracy of 97%.

1 Introduction

Natural Language Processing (NLP) and Understanding (NLU) have seen significant improvement over a few years and it still has many hot research topics. Several types of NLP tasks are being worked on by the research community like Machine translation, Sentiment Analysis, Question Answering, Language Modeling, Summarization, and Speech Recognition. In particular, many Tech giants have dedicated research teams on Sentiment Analysis, Speech Recognition, and Neural Machine Translation. NLP advancements have not only helped the general public but have also enabled the business to do target advertising and grow business. It is an extraordinary feet of scientific achievement that the machine can correctly predict the intention of a person by analyzing what they write or search for on the internet. It has to be noted that to gain this level of advancement in translation, and text classification (sentiment analysis), tech giants have collected, ethically or unethically, huge amounts of data from regular people. Companies like Apple, Google, Microsoft, and Meta have been successful in exploiting the benefits of NLP and made hundreds of billions out of it. However, ordinary people have also benefited from language translation, next-text prediction, and dictations. In this paper, we will be more focused on Sentiment analysis and Machine translation for NLP tasks.

Given a set of texts, the sentiment analysis model's objective is to determine the polarity of that text[4]. Sentiment analysis has several benefits to businesses and ordinary people. It allows targeted advertising by enabling the automatic detection and understanding of customer's feelings about the product or service. Political Campaigns also use sentiment analysis to find people's reactions to policy change and public opinion about certain candidates. Neural Machine Translation has become so good at translating languages that anybody can use their phone to communicate in different languages. The language translation model can also process a person's speech into a different language. Compared to traditional statistical machine translation (SMT), neural machine translation (NMT) is largely scale-able and powerful. However, NMT has its own challenges and problems including the need for a huge corpus to train on, and lower translation quality on very long sentences [3]. Despite a variety of approaches being proposed for NMT including the CNN model, the attention-based encoder-decoder gained traction and is most followed.

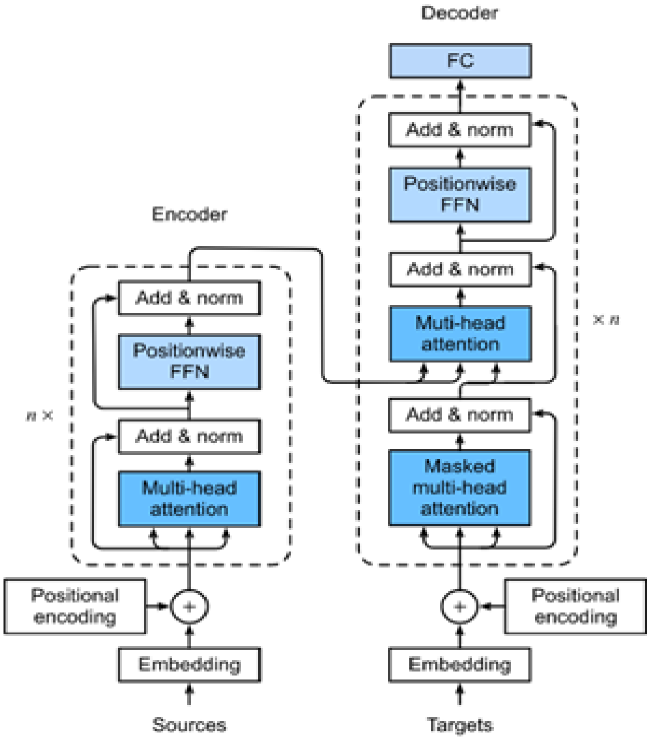


Fig 2: The Transformer Architecture.

NLP tasks have been boosted by the rediscovery of neural networks. Over the decade several neural networks have been introduced to work on NLP tasks. Recurrent Neural Networks (RNN) served as the state-of-art neural model for NLP tasks like text classification (sentiment analysis), language translation, etc. LSTM and

GRU are a modified version of widely used RNN models that overcome the shortcomings of traditional RNN models. RNNs are still popular models to build accurate and efficient language and speech recognition models. RNN recalls the past and its future selections are motivated by what it has learned in the past, a kind of auto-regressive property. However, the RNN models cannot manage long-range dependencies with ease and are unidirectional. To solve the issues mentioned above the Transformer model was introduced.

The transformer is a deep learning model that uses a self-attention mechanism to boost the training speed and efficiency of the model in making better predictions. Transformers, first introduced by Google via the "Attention is all you need" paper, rely solely on the attention mechanism without the use of any convolutional or recurrent layers [Vaswani, et.al.]. The transformer model is composed of an encoder and decoder each of which has its own multi-head self-attention as shown in Fig 1. The input and output sequence embeddings are added with positional encoding before being fed into the encoder and the decoder that stacks modules based on self-attention[8]. After the initial introduction of the transformer in 2017, researchers including Google have modified the Self-attention model to make it even more powerful hence the introduction of BERT [11].

Transformers have shown immense accuracy compared to conventional CNN models in object detection and image classification. Several modifications of the transformer have been introduced for computer vision and natural language processing tasks. Vision Transformer (ViT) and Compact Convolutional Transformer (CCT) are two transformer models for image classification problems. It seems that like BERT as the modern State of the Art (SOA) model for NLP tasks, the Transformer has become the SOA model for Computer Vision tasks. Vision Transformers are being widely explored for vision problems. Recently the introduction of a modified and scalable Vision Transformer for small devices by Apple [12] indicates Transformer models becoming the scientific trend in Computer Vision tasks.

In this paper, we will explore the use transformer model on NLP tasks like sentiment analysis and language translation. We will also investigate the use of the Vision transformer (ViT) for image classification tasks and recent advances in the field of ViT. We will some of the publicly available datasets and corpus for our training and testing purposes. We expect to learn about the transformer via experiments and literature reviews and see their efficiency. We will also learn about the shortcomings of the Transformer model and talk about how they can be tackled (as per research and literature review). We expect to better our understanding of the Transformer and how the ViT can be used for CV-based object trajectory tracking and prediction in 3D space. The main theme of this paper is centered around

- Exploring the power of Transformer for Neural Machine Translation. We will examine the performance of attention based on a model of language translation like English to French, English to Spanish, and other languages.
- Sentiment Analysis works like text classification where we will check the polarity of text. We will explain the workings of the transformer for sentiment analysis.
- We will also explore the rising influence of transformer models in computer vision tasks like image classification, and image captioning. Vision Transformer (ViT) is briefly explored.
- We will use large corpus like WikiQA corpus, Sentiment 140, and LibriSpeech to evaluate the training and testing efficiency.
- We will focus on the limitation of the sentence length for the Transformer model like BERT to understand the context of a sentence and predict missing words.
- We will also investigate the effect of transformer layer count for NLP tasks.

2 Preliminaries

In this section, we will introduce our audience to some of the NLP and Transformer related terminology and their definition. We will briefly explain a few most important concepts that the audience must know to understand the work in this paper. We will simply elaborate the Fig 1 in detail.

2.1 Encoder

The encoder layer processes the input iteratively one layer after another. The transformer encoder is a stack of residual attention blocks that maps the input sequence to a contextualized encoding sequence. Each encoder block consists of a bi-directional self-attention layer, followed by two feed-forward layers. The bi-directional self-attention layer puts each input vector into relation with all input vectors and transforms the input vector to a more refined contextual representation of itself. Each of the layers in the encoder has two sublayers. The first is a multi-head self-attention pooling and the second is a positionwise feed-forward network. The queries, keys, and values in the encoder self-attention are all from the outputs of the previous encoder layer. The transformer encoder outputs a d-dimensional vector representation for each position of the input sequence.

2.2 Decoder

The transformer decoder is also a stack of multiple identical layers with residual connections and layer normalizations [d2l]. The transformer-based decoder defines the conditional probability distribution of a target sequence given the contextualized encoding sequence [f]. The decoder consists of three sublayers with the third one being 'encoder-decoder attention' between the self-attention layer and the feed-forward network. In the encoder-decoder attention, queries are from the outputs of the previous decoder layer, and the keys and values are from the transformer encoder outputs [d2l]. Like Encoder, queries, keys, and values in decoder self-attention are from outputs of the previous decoder layer. The masked attention preserves the auto-regressive property of the decoder and predictions are based on previous output tokens.

2.3 Multi-Head Attention

Multi-head attention is a module for an attention mechanism that runs through an attention mechanism several times in parallel. The independent attention outputs are then concatenated and linearly transformed into the expected dimensions. Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions. With a single attention head, averaging inhibits this. Each multi-head attention block gets three inputs; Q(Query), K(Key), V(value) which are put through Dense layers before the multi-head attention function.

$$\text{MultiHead(Q, K, V)} = \text{Concat}(\text{head}_1, ..., head_h)W^O$$
$$\text{head}_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

2.4 Transformers for Vision

Vision Transformer (ViT): ViT is based on the transformer architecture originally designed for NLP tasks. The ViT uses an image pre-processing layer that partitions the image into a sequence of non-overlapping patches followed by linear projection [Srinadh et.al]. ViT was introduced as a competitor to the CNN model. ViT has seen major changes in the past year and the research trend points Transformer to be the next BERT-like model for Computer Vision tasks.

MobileViT: Recently published by Apple, MobileViT is a lightweight vision transformer. It combines the strength of CNNs (light-weight and easy to optimize for task-specific networks), and ViTs to build lightweight and low latency networks for vision tasks. MobileViT improves the hardware performance of the system since it requires less power per inference. MobileViT despite being claimed as better for performance, is very computation-heavy and will take a huge amount of resources to run training, especially custom training.

2.5 Switch Transformers

Switch Transformer [] is a switch feed-forward neural network (FFN) layer that replaces the standard FFN layer in the transformer architecture. Each switch layer contains multiple FFNs instead of a single FFN as shown in Figure 4. Like in network routers, the switch transformer routes the input signal via the model, activating only a subset of its parameter. Switch Transformers advocates training large models on relatively small amounts of data as an optimal computational approach. As a token passes through expert FFN, the number of floating point operations (FLOPS) per example stays constant even though, parameters might increase with a number of experts. The underlying principle of a switch transformer is that not all parameters need to be utilized for every token.

2.6 Named Entity Recognition

Named Entity was proposed at the Message Understanding Conference (MUC-6) to identify names of powerful, famous people and organizations and geographical locations in text, currency, and time [online]. Named Entity Recognition (NER) is an NLP technique to identify the mention of rigid designators from text belonging to particular semantic types. Pre-defined named entities are categorized as persons, organizations, locations, expressions of times, and money value. It helps in focused text extraction and information retrieval.

3 Methodology

To explore the efficiency of Transformers on NLP tasks we will run the transformer-based encoder-decoder model on different NLP tasks, primarily focusing on Sentiment Analysis and Language Translation. The primary way of research and learning is a literature review and active coding. We have run experiments with the available code and datasets, manipulating their hyperparameters to get better accuracy and BLEU score. For our image classification task, we focus more on the Vision Transformer (ViT) and recent advancements made in transformers for Computer Vision tasks. We will also discuss the Hugging Face Transformer API which has implemented all the attention models to date and has higher sentiment analysis accuracy. For our object detection task, we will focus on using the ImageNet or CIFAR-100 dataset. We will be working on Sentiment140, CoNLL2003.

4 Experiments

The experiments were conducted on the free version of Google Colab with variable runtime (GPU, CPU, TPU). Google Colab offers great computational power of GPU with Disk space of 78 GB, and RAM of nearly 13 GB. GPU runtime provides great efficiency when running the Deep Neural models. The experiments as mentioned were conducted on the following datasets: Eng-to-Spa, Eng-to-French, IMDB dataset, Sentiment140, LibriSpeech.

In all of the experiments, we conducted a hyperparameter search for every different method and reported the best-obtained results of each method. The important thing to mention is, that running more Epochs on the NLP tasks significantly more time than any Image Classification or object Detection task.

4.1 Eng-to-Spa Translation

We used the English to Spanish translation dataset and used the transformer model to train and predict. We performed a sentence-level translation instead of a token level. We see a decent performance of the transformer model on training and prediction. there is a slight overfitting of our model. We used the accuracy metrics and the BLEU score as performance metrics.

We can see that our model gives a training accuracy of nearly 72% and a loss of nearly 87% for 15 Epochs as we can see in the graph below. We will have to use the smoothing function for a sentence-level translation BLEU score []. We get a decent BLEU score of 0.39 for the English-to-spanish translation of ’I am playing a computer game’. The model shows decent performance.

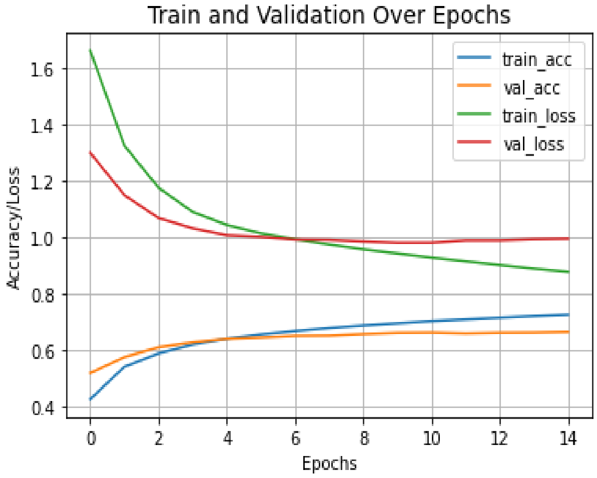


Fig 3: Eng2Spa Translation Performance chart.

If we choose to train our transformer model on all of the available text pairs in the dataset we get a decent performance improvement. The overfitting while it still exists, it isn’t as bad as the previous one

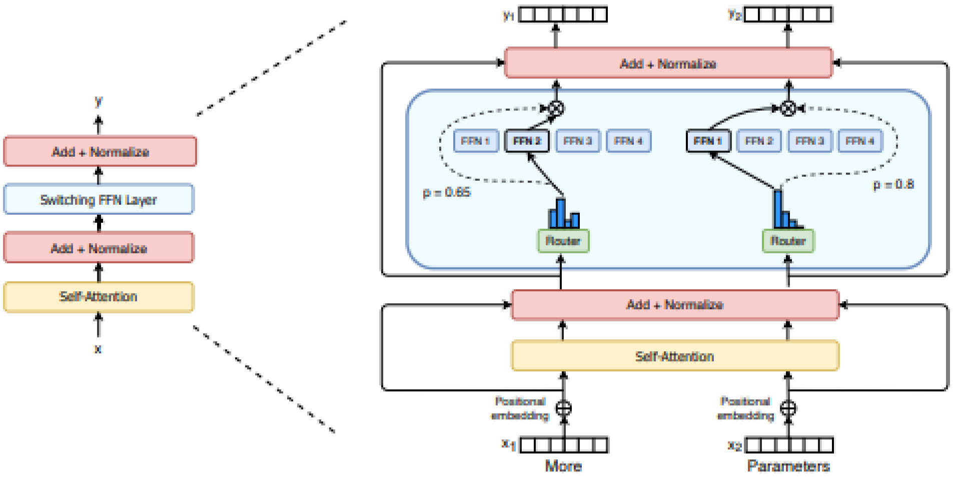


Fig. 1: : Switch Transformer Architecture

as we can see in the graph below.

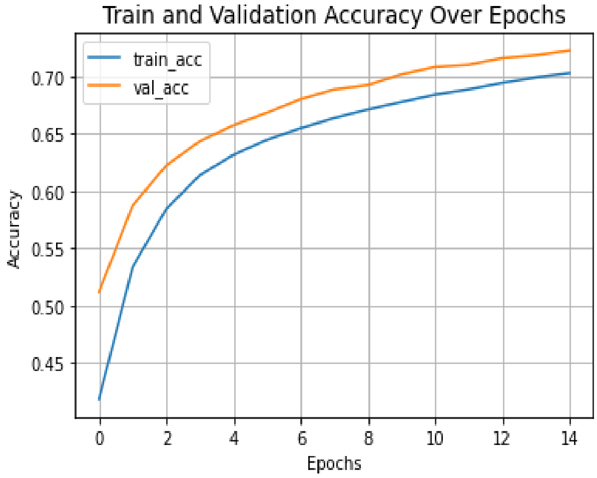


Fig 4: Eng2Spa Training w/ all dataset.

4.2 English to French Translation

The English to french language translation was done using the code available from the 'Dive into Deep Learning book'. The transformer model used performs well in the token-level translation but shows some problems when dealing with longer sentences resulting is a BLEU score of 0.00. The model has a gradual decrease in loss over 200 epochs showing decent training loss.

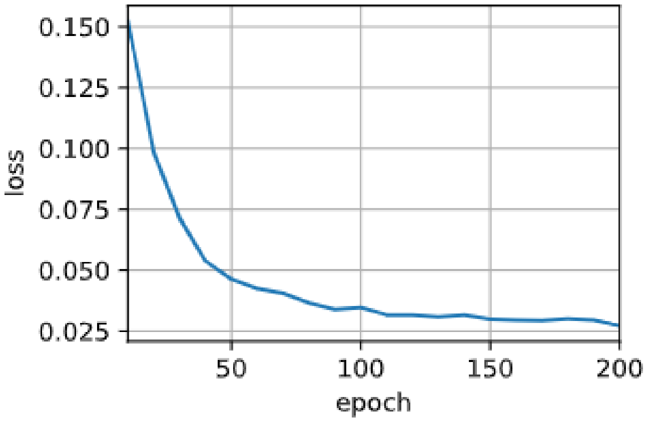


Fig 5: Training loss on Eng2Fra translation.

We shall add the attention heatmap and describe it. The running time of NLP should be described too.

4.3 Portuguese to English Translation

We used the Portuguese-to-English translation dataset that is compiled by TED talk from all the TED talks given in Portuguese. The model has a very good prediction than our previous two language translations. We have a max token per example of 394 as in Fig 5. Portuguese is the input language and English is the target language. Visualization of the attention heatmap for the start and end of the input and output token is given after the reference section.

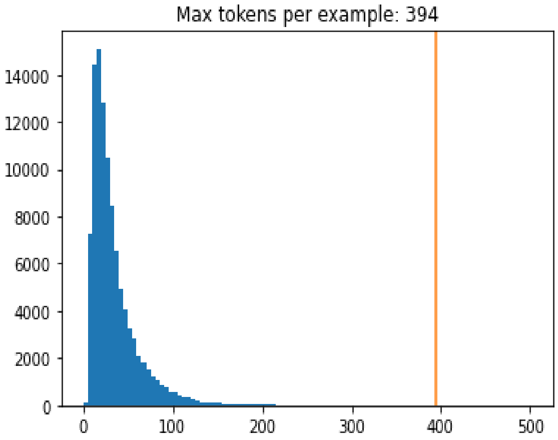


*Fig 6:*Max Token per Example .

We apply positional encoding to our Transformer model since there is no Convolutional or Recurrent layer involved. Positional Encoding gives the model information about the relative position of the tokens in the sentence. Adding a positional encoding makes the tokens stay closer to each other based on their meaning similarity and position in the sentence. The chart below represents position encoding on tokens.

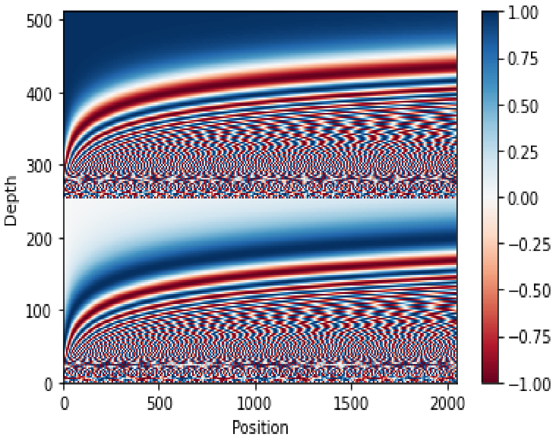


Fig 7: Position and Depth per token

The model has a loss of 700 batches per epoch with each epoch starting at batch 0 and incrementing at the rate of 50. The average loss over 20 epochs was 1.5196 and accuracy of 67%. An attention heatmap of a single head is given below.

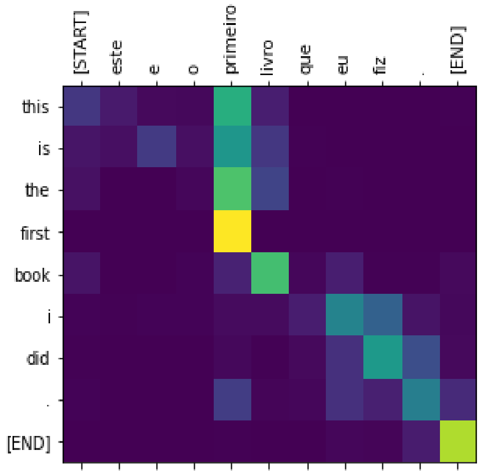


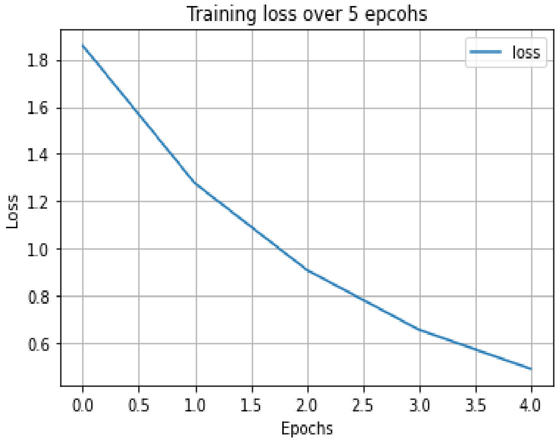
*Fig 8:*Single Head Attention Heatmap

5 Sentiment Analysis

In this part, we explain the work on Text Classification. Based on datasets from Question-Answering, Sentiment140, and IMDB review we will train and predict the polarity of text.

5.1 Question-Answer Tasks.

The dataset, Squad-Explorer, is in JSON file format with Questions and Answers about the history of the University of Notre Dame, and Superbowl history, used for training and validation respectively. We use the Exact Match Score (EMS) as performance metrics. We imported the BERT pre-trained model and ran our training and validation dataset.



5.2 IMDB Review Sentiment.

Running our custom Transformer model on the IMDB movie review dataset from Stanford we were able to get great training and decent validation accuracy, but the model shows significant overfitting. The result of our predictions for polarity isn't as good as the BERT model.

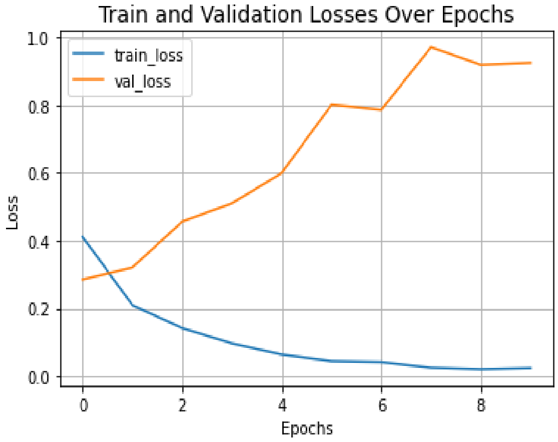


Fig 9: Training/Validation loss on IMDB review dataset.

Despite the training accuracy being great over 10 epochs with an average of 95%, the model's validation accuracy is also quite below the training accuracy. The model can't properly predict the polarity of movie reviews. However, if you import the BERT pre-trained model and use it to do the prediction of the movie review we can see great classification accuracy.

5.3 Named Entity Recognition

We used the CoNLL2003 dataset. We used the HuggingFace dataset library to train our model since the CoNLL2003 dataset isn't publicly available. The dataset has certain categories such as "Person", "Location", "And organization". Our model showed stagnant training accuracy of 40%, training loss was also almost similar for 50 epochs. We can run the model for 5 epochs and still get similar results.

Table 1 Exact Match Score on QA dataset

Epochs	Loss	EM Score
1	1.8570	0.78
2	1.2766	0.78
3	0.9111	0.78
4	0.6591	0.78
5	0.4934	0.77

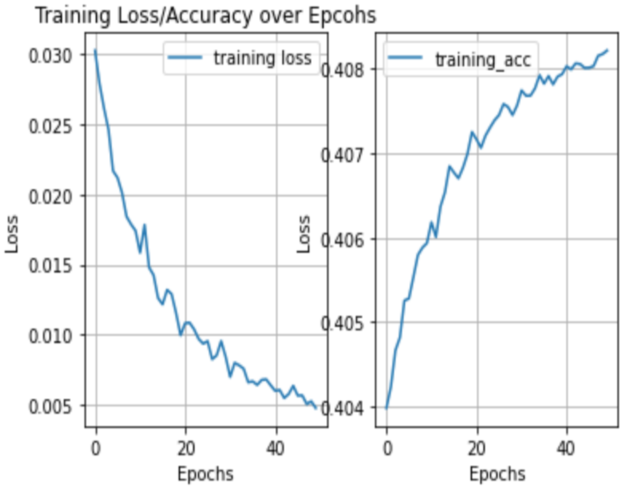


*Fig 10:*Training Loss and Accuracy for NER task.

We get a decent overall F1 score of 68%, and good F1, accuracy and recall for named entity recognition on categories like location, person, and organization as shown in table 2.

6 Object Detection Task

We ran the Vision Transformer model on the CIFAR-10 dataset and also tried to run MobileViT. MobileViT model needed huge computational power, GPU, and RAM memory but we had limitations with Google Colab and we left the experiment. As per research, MobileViT tends to perform much better than regular ViT.

6.1 Regular Vision Transformer

When used on the CIFAR-10 dataset the ViT model takes about an hour to run for 60 epochs and gives great results in comparison to CNN and other CNN derivatives. We achieve great train accuracy and test accuracy is also competitive.

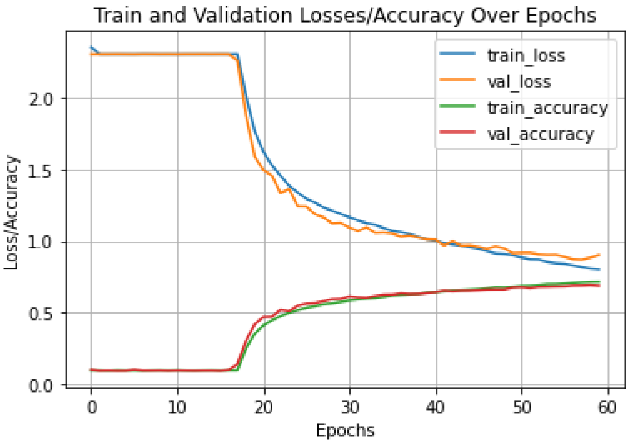


Table 2 Performance metrics based on Entity

LOC	Precision	70.28	Recall	80.19	F1	74.90
MISC	Precision	62.51	Recall	68.55	F1	65.39
ORG	Precision	49.23	Recall	64.35	F1	55.79
PER	Precision	58.42	Recall	57.06	F1	57.73

Fig 11: Training and Validation metrics over Epoch.

Using the MobileViT for training the flower dataset the model took over 3 hours to complete 30 epochs and was successful in getting top accuracy of near 85%. We will update the code in the Github repo.

7 Result and Analysis

Experiments were conducted to find out the training and validation accuracy for object detection tasks (ViT). The NLP task consisted of Sentiment Analysis like checking the polarity of text, movie reviews, and language translation. For NLP work we used the IMDB dataset from Stanford, three different language translation datasets like english-to-spanish, english-tofrench, portuguese-to-english. We also used a Q-A dataset publicly available on GitHub.The BLEU score for some of the translations was decent given the size of the dataset the model is trained on. If we compare the prediction result from the model pre-trained on the huge dataset, a corpus like BERT, and its types, we can get near-perfect language translation and sentiment detection.

The HuggingFace Transformer model available is trained and coded based on every research that is attention-based. HuggingFace can not only perform better NLP tasks but also perform much better object detection tasks. There are about 33 available spaces on HuggingFace web that can done online NLP or Computer Vision tasks. It seems like there is no topic related to NLP or Computer Vision that HuggingFace hasn't covered. HuggingFace Transformer are great tool for research experimentation in NLP and Vision.

Computational resources are a key to getting better performance on Switch Transformer and MobileViT. We ran most of our implementations on Google Colab with 16GB GPU and Sandbox with 16GB GPU assigned per node with 4 nodes being available to use. We ran our MobileViT model on the flower dataset from the TensorFlow dataset library and the model took 3 hours to run 30 epochs with 16GB GPU. This discouraged us from further experimenting with hyperparameter tuning. The lack of a faster GPU prohibited the frequent change and use of the network layer for training. Access to a good computational resource is most for testing out deep neural networks.

Pre-training the model with an established network that has been trained on millions of parameters can accelerate the learning rate of any neural network. We can also improve the model's efficiency by increasing the number of epochs, increasing the number of layers in the transformer layer, and changing the learning rate, and the weight decay. Fine-tuning a model using a large pre-trained high-resolution dataset can also improve the overall efficiency of the network. Having a high number of epochs won't necessarily increase the accuracy metrics, often after reaching to certain level of validation and testing accuracy, the model stays stagnant.

Unless we find an unexplored area of research we shouldn't shy away from using the pre-trained model like BERT, DistillBERT or even use or import the huggingface transformer API and spaces for NLP and image classification work. We failed to properly implement some of the clean codes that were available to explore.

The results from our NLP tasks like Q-A, sentiment analysis, and named entity recognition were quite decent and it gives us the motivation to conduct improved research and write better code implementation for certain NLP and object detection problems. We couldn't find anything that would work better for tiny object detection. We have mentioned that we failed to implement work on several datasets and failed to accomplish our target.

The BLEU score (Bilingual Score Evaluation Understudy), despite, being a widely used metric for language translation, is not good for sentence-level translation, it was never meant for sentence-level translation evaluation. To get sentence level evaluation BLEU score we use the smoothing function introduced by [5].

8 Conclusion

Transformers and other attention-based models are definitely the future base for research expansion in the field of Natural Language Processing. In this paper, we have carried out several intriguing NLP and Computer Vision tasks, but there were several implementation challenges. Going back to our previous research we definitely believe that Transformer and pre-trained models can greatly assist in learning from small data. We underestimated the time and challenge of the NLP task's code implementation. Training a model on a dataset for NLP requires huge computational power and it restricted us from frequent hyper-parameter tuning. We were very optimistic about getting many things done but failed to achieve most them. We have to acknowledge the incredible power of HuggingFace Transformers and their implementation of all attention-based models. HuggingFace has blown away the NLP task like OpenAI GPT. Despite not being able to run the MobileViT models and Switch Transformer for NLP tasks, we believe that these models can outperform their conventional counterparts.

9 Future Work

Nothing goes as expected, and that's true in our case too. We conducted some of the experiments with the belief that it would give out the optimal results we were looking for, but we were disappointed. Because of the time constraints, we couldn't subsequently make any more changes to our model to see whether it works or not. The MobileViT and Switch Transformer model stays as our top priority for future code implementation using more advanced computational resources. We certainly plan to keep working on changing a few things in our code for all tasks to make them more readable and understandable. We will certainly continue to explore more datasets to custom-built a model that can give us intuition into the development of a pre-trained model. One of the main things that we plan to do is run the transformer model on our custom dataset for object detection.

10 Acknowledgments

We mostly like to thank Dr. Maida for his tutorial on transformers and for covering most of the used topics in this paper. We would like to express our sincere gratitude to the free version of Google Colab that allowed us to use GPU power for some experiments. We would also like to thank the authors of the few articles we read for the implementation of neural networks. The articles and tutorials in some of the AI/Data Science websites like medium.com, and towardsdatascience.com helped a lot in understanding the NLP.

11 References

1 Fedus, William, Barret Zoph, and Noam Shazeer. "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity." arXiv preprint arXiv:2101.03961 (2021).

2 Stahlberg, Felix. "Neural machine translation: A review." Journal of Artificial Intelligence Research 69 (2020): 343-418.

3 Koehn, Philipp, and Rebecca Knowles. "Six challenges for neural machine translation." arXiv preprint arXiv:1706.03872 (2017).

4 Pouransari, Hadi, and Saman Ghili. "Deep learning for sentiment analysis of movie reviews." CS224N Proj (2014): 1-8.

5 Chen, Boxing, and Colin Cherry. "A systematic comparison of smoothing techniques for sentence-level bleu." Proceedings of the ninth workshop on statistical machine translation. 2014.

6 Ganesh, Prakhar, et al. "Compressing large-scale transformer-based models: A case study on bert." Transactions of the Association for Computational Linguistics

9 (2021): 1061-1080.

7 Feldman, Ronen. "Techniques and applications for sentiment analysis." Communications of the ACM 56.4 (2013): 82-89.

8 Zhang, Aston, Zachary C. Lipton, Mu Li, and Alexander J. Smola. "Dive into deep learning." arXiv preprint arXiv:2106.11342 (2021).

9 Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. Deep Learning. Cambridge (EE. UU.): MIT Press.

10 Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

11 Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

12 Mehta, Sachin, and Mohammad Rastegari. "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer." arXiv preprint arXiv:2110.02178 (2021).