Sentiment Analysis using Transformer and BLSTM

Anonymous Author(s)
Wangfan Li

1 Introduction

As deep learning advances, particularly in the area of NLP, the task of sentiment analysis as a classification task always an interesting challenge to tackle and for testing models.

The dataset ArtEmis created by Achlioptas, et al. (2021) is particularly challenging, in that it pairs the impression of a painting with the feeling that the painting is trying to convey. With up to 454k entries gathered through crowd sourcing and the vague nature of the emotional language used, for a model that can correctly classify the description "The boy walks down the path, the clouds at his back like old friends following him home" as contentment, it can presumably understand some aspect of the emotional language.

With 9 emotional categories of sadness, contentment, awe, anger, fear, amusement, disgust, excitement and something else, it is not an easy task when compared to more clear cut sentiment classification. In this work we will be trying to compare the performance of two models that are proven to be effective in NLP problems, and see if the difference in architecture can lead to better results in this sentiment analysis task.

The goal of this paper is to examine if the model can perform better by combining bi-directional long short term memory and transformer, than just transformer only, especially in areas of sentiment classification.

2 Related Work

Through many other literature, such as one by Zhang, et al.(2018), It is shown that Long Short Term Memory, LSTM for short, is capable of achieve excellent results for text classification, and is a solid foundation for sentiment analysis.

The famous paper by Vaswani, et al. (2107) introduced attention and how it could be used in a transformer architecture, and shows why it might function better than LSTM for NLP domain problems as the result showns.

The study done by Huang, F., et al. (2021) further shows how attention can improve the result from LSTM, by combining attention with LSTM it achieved state of the art result on sentiment analysis tasks.

Another papaer done by Vateekul, et al (2016) shows the effectiveness of different approaches on twitter data for sentiment analysis, and shows that both LSTM and transformers can be valid deep learning approaches for analyzing twitter text.

The study down by Devlin, et al.(2018) introduces the concept of BERT, a bi-directional transformer where its shown to be very effective for language understanding tasks, which includes sentiment analysis. The architecture in this paper will be used as our baseline.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

The approach that we take inspiration from is done by Huang, Z, et al. (2020), specifically the TRANS-BLSTM-SMALL model, where a combination of transformer and BLSTM is shown in the case of question-answering dataset to be superior to a pure transformer model. The reasoning behind the combination is that the authors theorize a joint model is better than a transformer baseline, as it has more accurate sequential modeling by complementing each other..

3 Methods

Taking the previous study into account, we will implement the architecture proposed by Huang, Z, et al. (2020) to see if we can replicate similar level of improvement over a base transformer classifier. The original logic for combining bidirectional LSTM with transformer is that it produces better joint models, and that the two architecture is complementary in capturing sequence information, because the two models combined is more effective for sequence modeling. We will see if it is also effective for high ambiguous emotional text included in ArtEmis.

Since we are less concerned about optimizing the model, but instead to compare the performances, we did a 90/10 training/test split of the 454k data, and feed the network with every words tokenized in the training set,

We used a base transformer classifier with architecture exactly shown in Figure 1

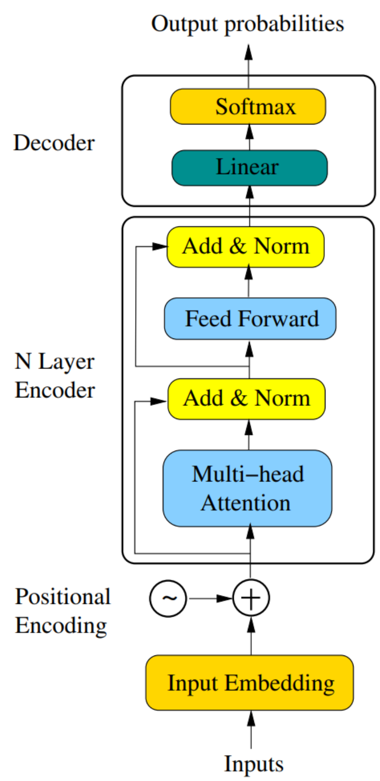

Figure 1: Transformer

The model described in Huang, Z, et al. (2020) is very similar to the above, except that a single birectional LSTM layer is added to the layer encoder block, of which the output from the two directions will be averaged as input for the following feed forward layer. The output of said linear layer will be added with the outputs from the residual connection and the feed froward before being normalized, and then either feed to the following encoder block, or the decoder if the current block is the last block. The architecture is described in Figure 2.

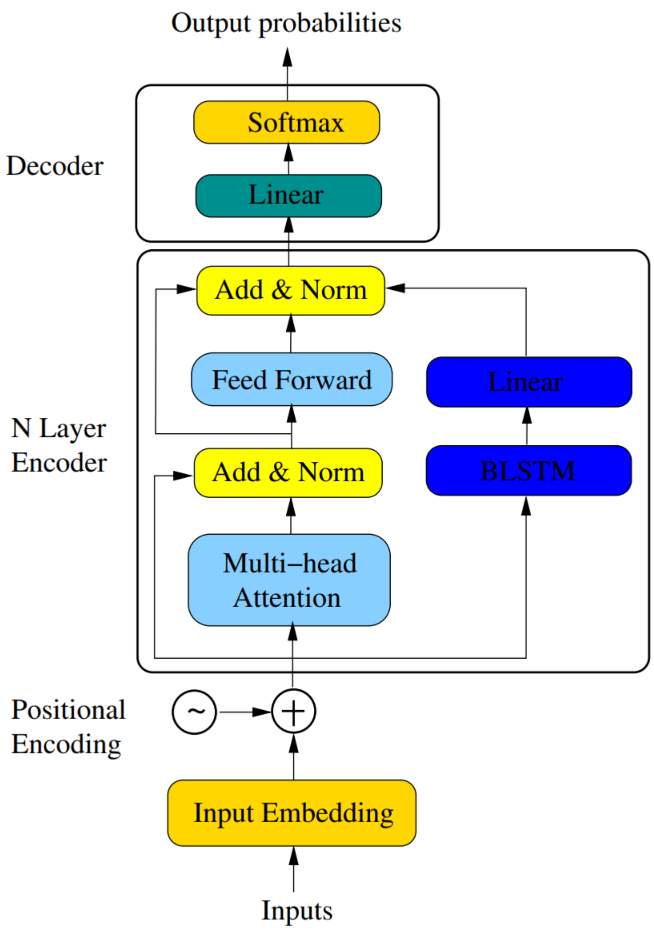

Figure 2: Transformer+LSTM

The two models share the same parameter regarding the overall structure, in that they both have 120 input embedding size fed into an positional encoding, 4 encoder block each with 8 heads for multi-head attention, and every feed forward layer has 28 dimension with a leaky ReLU output.

4 Results

The goal is to compared the two models and see if the added BLSTM component can increase the accuracy of the model. For fast training, we trained both models for 10 epochs with 64 batch each, with 0.0005 learning rate, adamW optimizer and cosine annealing scheduler. The loss is then calculated using multi-class cross entropy at the end. No pre-training is used, or is any pre-existing trained network used. Fine tuning is kept at minimum, where the parameter values are establish only after a few runs of the model, and are kept the same for both models for comparison.

Aside from accuracy and loss, we also used F1 score, which takes both precision and recall into account in light of the unbalanced dataset, to measure the performance of the model. The class label "anger" is very under represented while "contentment" is very over represented, as shown in figure 3.

The results are shown here after 10 epoches for both model.

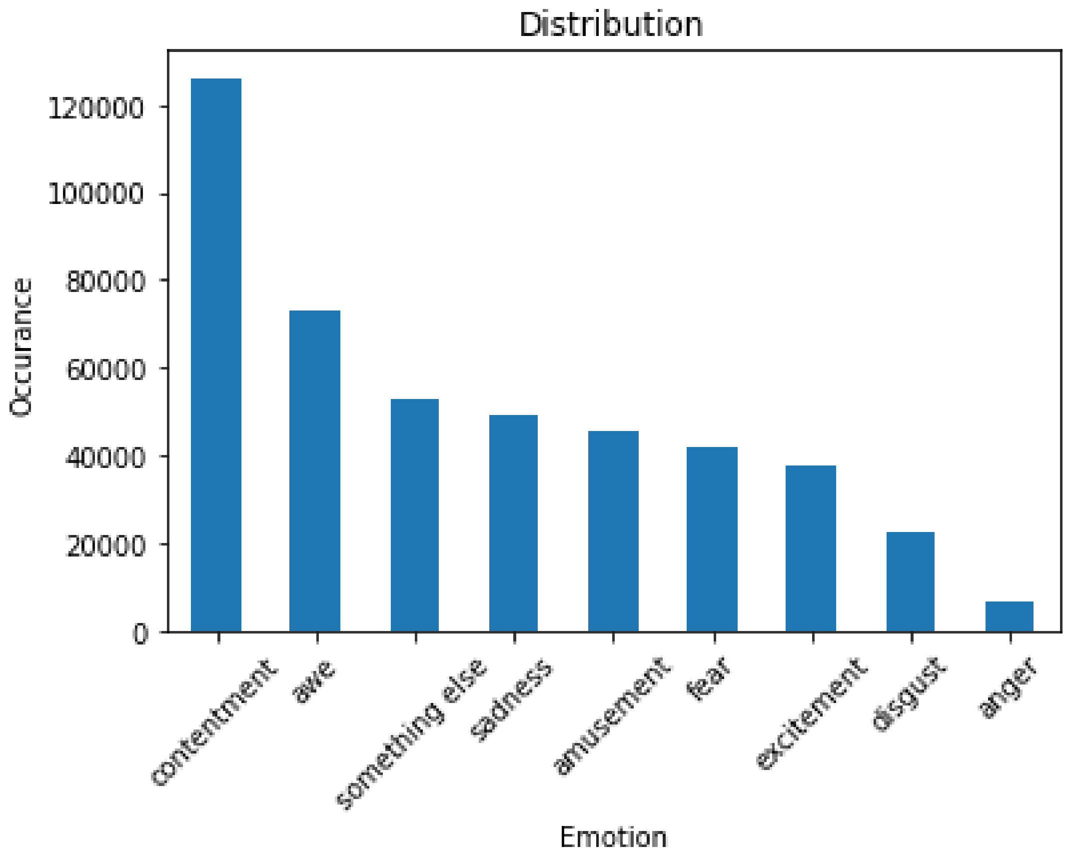

Figure 3: class distribution

Model	Training accuracy	Testing accuracy	F1
Baseline	0.688	0.564	0.643
Transformer + BLSTM	0.702	0.420	0.429

Figure 4: Result

As can be see from the table, the proposed model has better training accuracy but lower testing accuracy and F1 score, showing that it's worse when generalizing outside of the testing sets. This could also be an overfitting issue resulted from the increased complexity of the architecture, and that the network might be trying to remember the sequence of the training set when for sentiment analysis it is not as important for generalization.

It also takes more time to train, with the baseline taking 1150 secs and the Transformer + BLSTM taking 3080 secs and more resources, though the loss start off as less and decreases more rapidly when compared to the baseline transformer as seen in figure 5.

For actual inference on custom text, it's not obvious as to which model is better even in case where they differ to the label or to each other, as the emotional response to a description is very subjective. There is also the aspect that the model is trying to predict one label only from the data set limitation when in practical situation, a description might have multiple labels.

Finally, to make sure the result is consistent, both models are also run multiple times to see if the result hold, and it does through multiple training.

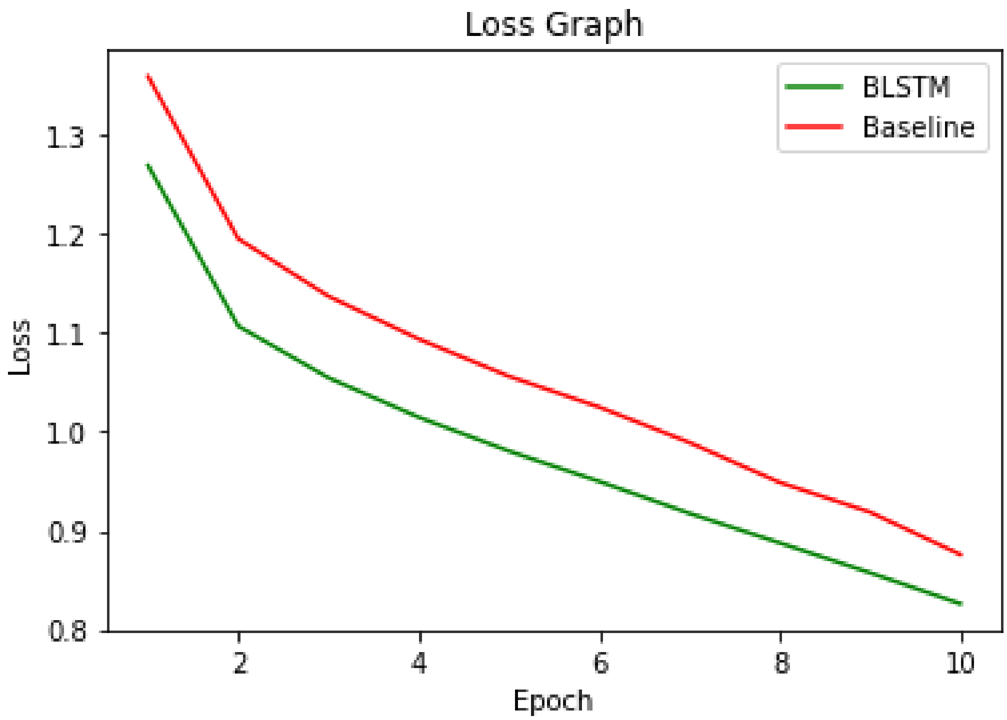

Figure 5: loss graph

5 Conclusion

We can see that the joint approach does not improve the effectiveness of the model significantly

One possible cause for failing to replicate the result could be that sequential information is less important for emotion analysis, so that adding bidirectional LSTM does not benefit the performance, and could instead harm it if LSTM is inferior in sentiment analysis. Where bidirectional LSTM is useful for question-answering since it provides different sequential information and that a joint model would make more sense.

Another possibility is that the result is from the incorrectly labeled data points instead of the joint approach being ineffective, in that multiple people might interpret different descriptions differently, in that one crowd worker might label something as "sadness" while another could label it as something totally different. There is no measurement on human performance in the original study, but it could be that a human can do no better when predicting what others labeled the description.

Finally, the amount of labels might be insufficient for the range of emotions, with one label being "something else" that includes any other emotions not labeled, which could interfere with the results.

For future work, it might be beneficial to try this approach on a more clear cut database, in that there is 100 percent a right label and a wrong label instead of the subjective label and judgement from ArtEmis, and see if the conclusion changes at all. It could also be interesting to see how the joint model might benefit from a learned positional embedding, or a relative positional embedding, instead of the absolute positional encoding used in this paper.

References

[1] Achlioptas, P., Ovsjanikov, M., Haydarov, K., Elhoseiny, M., Guibas, L. J. (2021). Artemis: Affective language for visual art. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11569-11579).

[2] Huang, F., Li, X., Yuan, C., Zhang, S., Zhang, J., Qiao, S. (2021). Attention-emotion-enhanced convolutional LSTM for sentiment analysis. IEEE Transactions on Neural Networks and Learning Systems.

[3] Huang, Z., Xu, P., Liang, D., Mishra, A., Xiang, B. (2020). TRANS-BLSTM: Transformer with bidirectional LSTM for language understanding. arXiv preprint arXiv:2003.07000.

[5] Zhang, L., Wang, S., Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.

[6] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[7] Vateekul, Peerapon, and Thanabhat Koomsubha. (2016). "A study of sentiment analysis using deep learning techniques on Thai Twitter data." 13th international joint conference on computer science and software engineering (JCSSE). IEEE, 2016.

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.