COMP551 - Assignment 4 Report

Abstract

In this project, we investigated the performance of the BERT base-uncased model on the AG News dataset, comparing two approaches for text classification: probing and fine-tuning. Probing involved freezing BERT's parameters and using extracted embeddings to train downstream classifiers, while fine-tuning updated BERT's weights end-to-end. We experimented with four probing strategies—mean, CLS, first token, and last token—and tested them using both KNN and multiclass logistic regression. The best performance was achieved using the mean pooling strategy with K=3 in KNN, which yielded a test accuracy of 93.4%, compared to the 92.4% of fine-tuning. Attention heatmaps also revealed meaningful distinctions in how the model attended to tokens in correctly versus incorrectly classified examples. These results suggest that for classification tasks on well-structured datasets, carefully designed probing methods can rival or exceed the accuracy of fine-tuning.

Introduction

The task at hand necessitated an investigation into how two mainstream ideas for deep learning text classification (probing vs fine-tuning) fared against each other, with the help of the AG News dataset, which consisted of over 1 million news articles covering a variety of topics, including passages in the realms of world news, sports, business, and science/technology.

We found that probing performed marginally better here, with the best probing strategy achieving a test accuracy of 93.4%, compared to BERT's test accuracy of 92.4%. The aforementioned best probing strategy was determined to be a combination of taking the mean over all token representation, along with a value of k=3 in KNN. However, we also observed that the BERT model took slightly less time to train, (~15 minutes for probing compared to ~8 minutes for BERT) which could account for the slight difference in accuracy.

The concept of probing, which involves "freezing" a model's (typically pre-trained) parameters by not allowing them to update, permits us to test different embedding/representation strategies and see which one best allows us to classify text.[1] On the other hand, the idea of fine-tuning takes an oppositional approach; the parameters of the pre-trained model are allowed to update themselves as the model is fine-tuned (i.e. a pre-trained model is trained on a small, task-specific dataset to adapt itself to the new task at hand).[2]

Datasets

The dataset used for this assignment was the AG news dataset accessed via the HuggingFace library. Comprising of over 1 million news articles from more than 2000 news sources, and featuring articles from four categories (World, Sports, Business, Science/Technology), the AG news dataset provided us with an

[1] Yonatan Belinkov; Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics* 2022; 48 (1): 207–219. doi: https://doi.org/10.1162/coli_a_00422

[2] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? Proceedings of the 18th China National Conference on Computational Linguistics (CCL 2019), Lecture Notes in Computer Science, 11856, 194–206. Springer, Cham. https://doi.org/10.1007/978-3-030-32381-3_16

ample, comprehensive collection of articles, that in turn allowed us to thoroughly compare the performance of a probing approach against fine-tuning a pre-trained BERT model. The articles were preprocessed using an uncased pre-trained tokenizer, pre-trained on textual input with two main objectives in mind:

- Masked Language Modelling → masking random words in a sentence and having the model learn which words should be in place of the placeholders
- Next Sentence Prediction → two masked sentences are concatenated, with the aim of determining whether or not they were adjacent/following each other in the original text

To better understand the dataset, we devised a barplot showing the class distribution among the three splits: training, testing, and validation. We found that the classes were distributed perfectly evenly among the three splits, which helped ensure that we could compare probing vs fine-tuning without worrying that one class would "pull" the model in a certain direction.

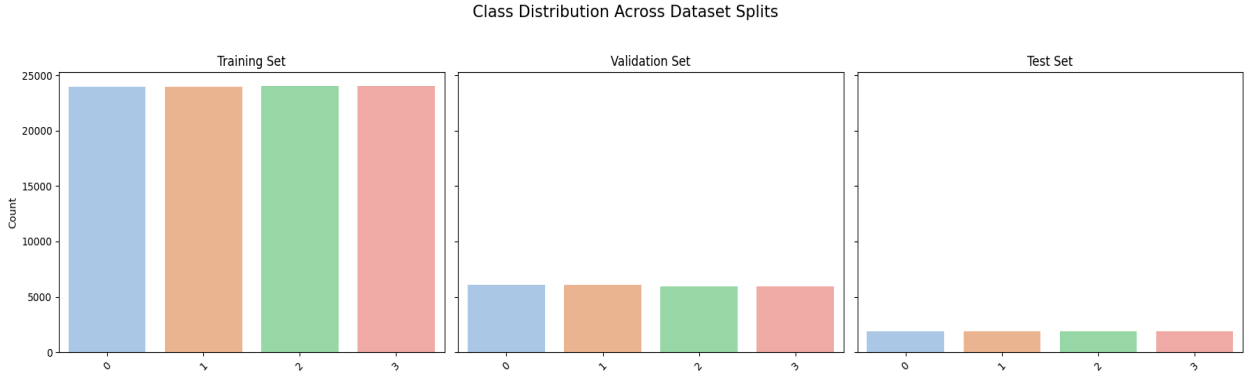

Figure 1: Class distribution across training, validation, and testing set

Benchmark

In this project, we used the pre-trained BERT base-uncased-model from HuggingFace. This version of BERT is composed of 12 transformer layers, each with 12 attention heads and a hidden size of 768, using GELU for the activation function. The model architecture only had encoders (no decoders) to output a contextualized representation for each token. For probing, we froze the model and extracted the embeddings using various strategies, including taking the [CLS] token from the final hidden layer, taking the first (and last) token embedding, and computing the mean over all token embeddings. The performance of each method was then evaluated by using them as input features for K Nearest Neighbors and multi-class logistic regression classifier. We tested a wide range of values for K (3-10) across the different strategies, allowing us to estimate the effectiveness of one extraction method over another.

Similarly, we also tested all the extraction strategies using the logistic regression classifier to measure which extraction led to the best accuracy. Now, to test the full capabilities of BERT, we unfroze the model and fine-tuned all the parameters using the AG News dataset, with 10,000 samples for training and 2,000 for validation. During fine-tuning, all model parameters were updated, enabling BERT to adapt its internal representations for the news classification task.

By evaluating the accuracy of BERT's fixed language representations by first probing and then again by fine-tuning our model, we were able to analyze its performance before and after the news classification adaptation.

Results

Validation accuracies for probing strategies for KNN

We explored four embedding extraction strategies for KNN. From the best to worst performing, the four strategies were mean, cls, last, and first. The best performer on KNN was with the 'mean' strategy using K = 3 (93.4% accuracy), while the worst was with the 'first' strategy using K = 10 (76.8%).

	strategy	k	accuracy
0	mean	3	0.934
1	mean	4	0.925
2	mean	5	0.915
3	mean	6	0.909
4	mean	7	0.908
5	mean	8	0.907
6	mean	9	0.905
7	mean	10	0.903

Table 1: Accuracy using mean

8	cls	3	0.904
9	cls	4	0.897
10	cls	5	0.890
11	cls	6	0.888
12	cls	7	0.885
13	cls	8	0.883
14	cls	9	0.878
15	cls	10	0.877

Table 2: Accuracy using CLS

16	first	3	0.838
17	first	4	0.819
18	first	5	0.793
19	first	6	0.797
20	first	7	0.787
21	first	8	0.776
22	first	9	0.771
23	first	10	0.768

Table 3: Accuracy using first

24	last	3	0.896
25	last	4	0.887
26	last	5	0.886
27	last	6	0.876
28	last	7	0.866
29	last	8	0.860
30	last	9	0.855
31	last	10	0.857

Table 4: Accuracy using last

Validation accuracies for probing strategies for Logistic Regression

We also tested the different strategies for logistic regression and got 'cls' as the best performing strategy (97.7%) and 'last' as the worst performer (94.0%). The high test accuracy goes to show the usefulness of pre-trained BERT's powerful language representations.

	strategy	accuracy
0	mean	0.964
1	cls	0.977
2	first	0.960
3	last	0.940

Figure 5: Logistic regression accuracies using different strategies

Test accuracies for the best probing strategy and fine-tuned BERT

After fine-tuning, the model achieved an accuracy of 92.3%, not as performant as probing (97.7%), but not a far cry either, it is still very solid in terms of accuracy.

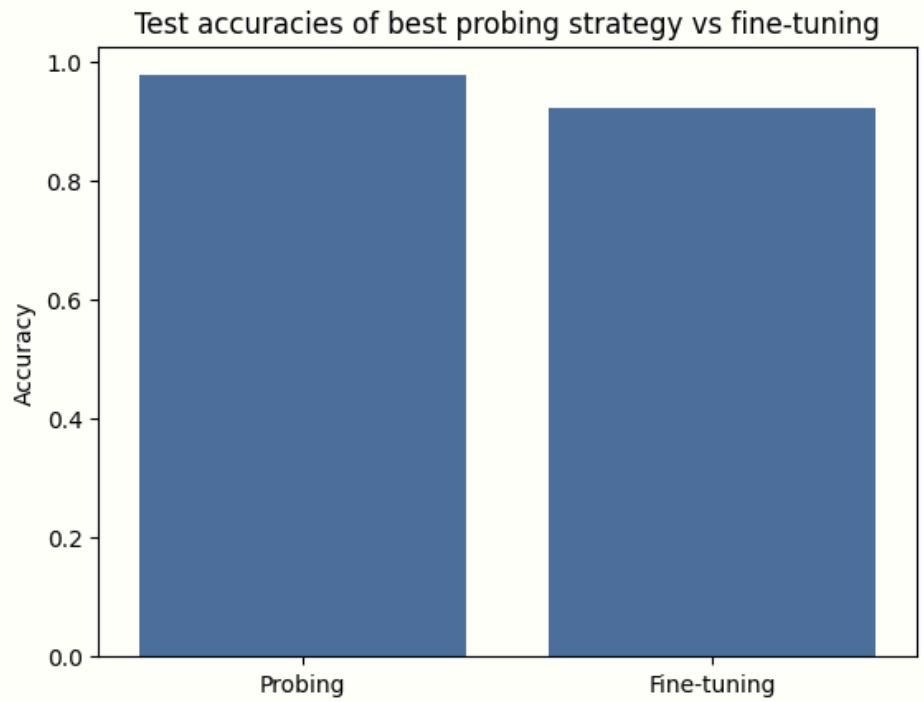

Figure 2: Barplot of best probing strategy (logistic regression using cls strategy) and fine-tuning

Attention heatmap for correct positive and negative predictions

The heatmaps for class 0 (world news) and class 1 (sports) shows how the CLS token distributes its attention across the different tokens for correct predictions. As expected, for world news, the CLS token pays more attention to words such as “Israel,” “palestinian,” and “strip,” and for sports, the focus is on the words “(real) madrid,” “football,” and “pitch.”

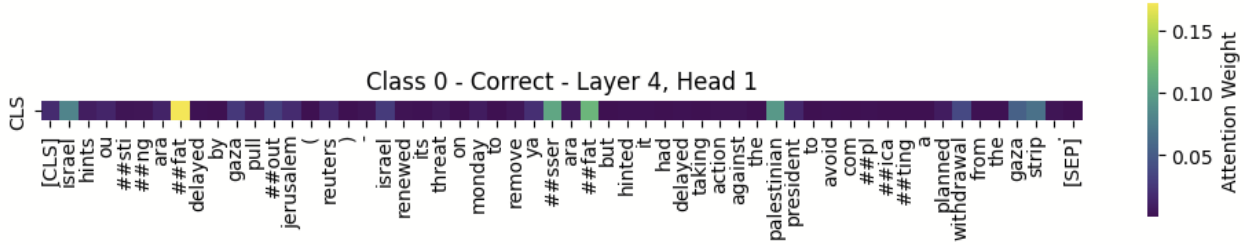

Figure 3: Correct positive predictions for world news

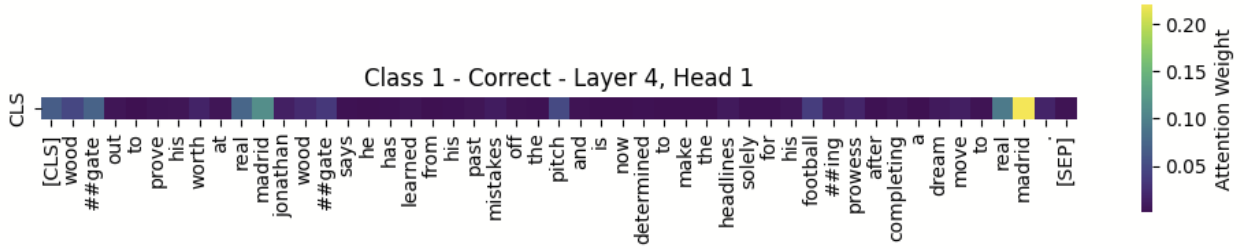

Figure 4: Correct negative predictions for sports

Attention heatmap for incorrect positive and negative predictions

The heatmaps for incorrect predictions are shown below. For example, in figure 5, we can see the heatmap for a sample that was mistakenly classified as world news. Its CLS token paid attention to the words “shares,” “equity,” and “institutional,” which may have been misidentified for what was actually business.

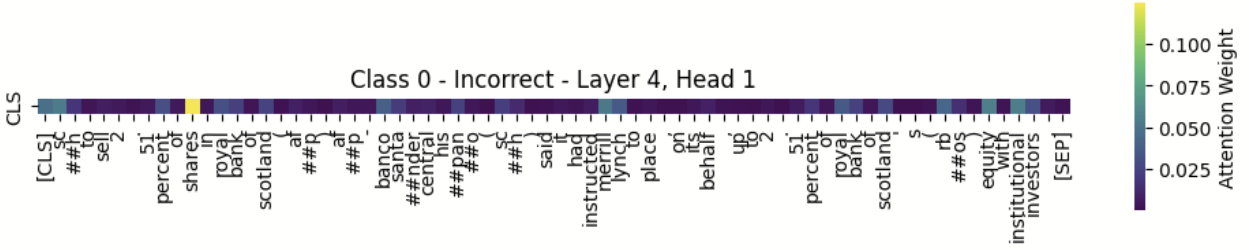

Figure 5: Incorrect positive predictions for world news

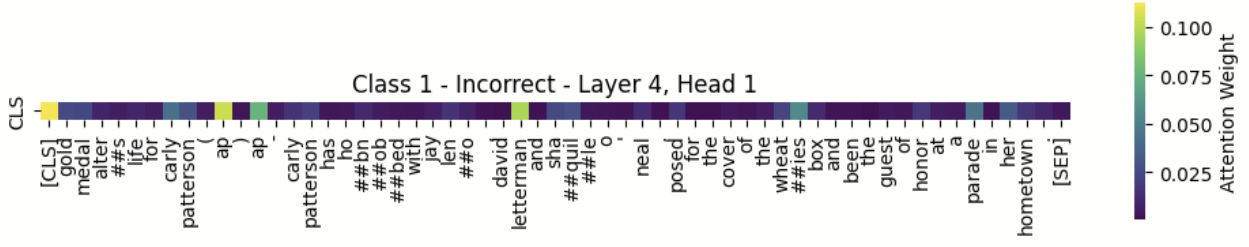

Figure 6: Incorrect negative predictions for sports

Discussion & Conclusion

Our experiments demonstrate that the BERT base-uncased model offers strong performance for text classification on the AG News dataset, regardless of whether it is fine-tuned or probed. Among the probing strategies tested, taking the mean over all token embeddings with K=3 in KNN achieved the highest test accuracy at 93.4%, slightly outperforming full fine-tuning, which reached 92.4%. While probing yielded better accuracy, it was also more computationally intensive, requiring nearly twice the training time (approximately 15 minutes) compared to fine-tuning (approximately 8 minutes). This highlights a practical trade-off between accuracy and efficiency, especially in scenarios where computational resources or inference time are constrained.

The performance gap between probing strategies was also notable. Using only the CLS token or the first/last token embedding consistently underperformed the mean strategy, suggesting that relying on a single position token fails to capture the full contextual richness encoded in BERT's hidden states. Fine-tuning the model allowed all parameters to adapt to the classification task, enabling a streamlined pipeline with good accuracy despite no additional classifier being trained.

Attention heatmaps revealed that correct predictions often corresponded with sharply focused attention on task-relevant words, whereas incorrect predictions showed weaker or misaligned attention patterns. These visualizations provided qualitative support for the quantitative results and offered insight into how BERT arrives at its decisions.

All together, probing with carefully chosen strategies can rival and even surpass fine-tuning in accuracy at a higher computational cost. Fine-tuning, while slightly less accurate, remains the faster and more integrated approach. Possible future directions include approaches that fall between these methods, such as partial unfreezing. Deeper layer-wise probing and evaluating these methods on less clean or low-resource datasets could also be utilized to explore their generalizability.