Exploring the Impact of Model Distillation on Explanations for Deep Neural Networks

1. Abstract

This project investigates the impact of model distillation on the explanations of deep neural networks, focusing on sentiment classification using BERT and its distilled counterpart, DistilBERT. While model distillation improves computational efficiency by transferring knowledge from a larger model to a smaller one, this process introduces qualitative changes in the models' explanations that are not well-understood. Leveraging SHAP (SHapley Additive exPlanations), we compare feature attributions between BERT and DistilBERT, examining how the distillation process alters decision-making logic, particularly in cases where small accuracy losses occur. Our results reveal that while both models prioritize similar key features, DistilBERT assigns higher magnitudes of importance to salient features, likely due to its streamlined architecture. This shift raises concerns about fairness and generalization in high-stakes applications. The study underscores the importance of balancing efficiency with comprehensive representations in deploying compressed models, advocating for further research into the trade-offs of model compression techniques.

2. Introduction and Background

The use of different architectures of deep neural networks has become an integral part of most academic and professional domains today, especially with the development of large language models that have taken the world by storm in their most recent form of chatbots. However, the accessibility of these models is restricted by their high storage and computation requirements for both training and inference. This restriction drove the development of model compression methodologies, which aim to reduce the required memory to store and run these models and improve their computational efficiency. However, this improvement comes at a cost, as it nearly always causes a drop in performance, which is most commonly measured in accuracy or through other benchmarks, depending on the use case addressed.

When compressing models, the concerns are usually purely quantitative, aiming to explore how many percentage points the accuracy or other applicable metrics are affected. However, there is not as much research exploring qualitatively which part of the potential input spectrum is influenced by this and why.

This qualitative exploration becomes more important as the use of deep neural networks becomes ubiquitous and influences high-stakes decision-making areas and sensitive use cases where human lives or other human rights can be on the line depending on the outcome of a neural network. To probe this space, we look into the task of sentiment classification using BERT as we attempt to answer the following questions: How do model compression techniques such as quantization and distillation affect explanations in deep neural networks (DNNs) for sentiment analysis? Specifically, how do small accuracy losses caused by compression affect different parts of the prediction spectrum? By comparing explanations from the original and compressed models, we aim to understand how these accuracy losses manifest and whether they impact the reliability of model decisions in specific cases.

3. Motivation

Model compression techniques are crucial for deploying deep learning models in resource-constrained environments, reducing computational demands without significantly sacrificing accuracy. However, little is known about how these techniques impact the explanations of the model's predictions, particularly when small accuracy losses occur. Since most research focuses on quantifying the accuracy loss, there is limited knowledge of the qualitative impact of the compression process, and whether it affects certain parts of the decision boundary disproportionately compared to others. This is what we aim to investigate by looking at explanations. We want to understand not only what the accuracy losses are after compression, but also how the explanations are impacted by the compression process. This is especially important for high-stakes tasks, where understanding the model's decision-making process is vital for trust and accountability. Gaining insights into how compressed models differ in their feature attribution can improve confidence in using these models in practical, real-world applications.

While previous work has addressed the computational efficiency and accuracy trade-offs of DNN compression techniques, little attention has been paid to how they affect model explanations. By using Shapley values to compare the explanations from the original and compressed models, this research aims to investigate whether and how compression alters decision-making logic in specific cases, especially where small accuracy losses occur.

4. Methodology

The methodology we use involves analyzing model explanations using BERT and DistilBERT for sentiment analysis. We utilize the Hugging Face Transformers library to load two pre-trained models: the original BERT base model and the compressed DistilBERT model, both fine-tuned on the Stanford Sentiment Treebank (SST-2) dataset.

For dataset selection, we employ the IMDB movie review dataset, which provides a rich and diverse set of sentiment-labeled text samples. The data loading process ensures a balanced selection of samples, randomly choosing an equal number of positive and negative reviews. Specifically, the implementation selects 10 samples (5 positive, 5 negative) with a reproducible random seed to maintain consistency across experiments.

The explanation generation relies on the SHAP (SHapley Additive exPlanations) framework, a model-agnostic interpretation method that provides detailed insights into feature attributions. Our implementation supports both a fast mode for quick analysis and a full-depth mode for comprehensive explanation generation. The SHAP explainer is carefully designed to handle tokenization, sequence length constraints, and computational efficiency, with built-in support for GPU acceleration when available.

The analysis process involves the following steps:

- Model prediction generation for each sample
- Assessing the classification performance for both models
- SHAP explanation computation
- Generation of various visualization outputs to compare model explanations

Computational considerations are addressed through efficient caching mechanisms, allowing repeated analyses to leverage previously computed results, reducing computational overhead.

5. Results

The results go through the performance of both models in the classification task, and then analyzes their explanations for further insights.

Table 1: Models Classification Performance

Model	**Class**	**Precision**	**Recall**	**F1-Score**
BERT	Negative	0.904	0.904	0.904
BERT	Positive	0.904192	0.904192	0.904192
DistilBERT	Negative	0.856343	0.918	0.8861
DistilBERT	Positive	0.911828	0.846307	0.877487

Table 1 shows the classification performance of BERT and DistilBERT for both positive and negative sentiment classes on 1001 randomly sampled prompts from the IMDB dataset. We notice that there is a small dip in performance from BERT to DistilBERT as the F1-Score shows, which is expected since the BERT model is the teacher model. However, the more interesting aspect is the fact that BERT is balanced and performs consistently across both classes, while DistilBERT does not only drop in accuracy but it also loses balance compared to BERT.

BERT shows consistent precision, recall, and F1-score values of 0.904 for both negative and positive classes, indicating balanced performance across both classes. DistilBERT, while slightly less precise for the negative class (precision of 0.856343) and slightly lower in recall for the positive class (recall of 0.846), still maintains robust performance with F1-scores of 0.8861 for negative and 0.877487 for positive classes. The positive class is more affected by the distillation process, as evidenced by the larger drop in recall. This comparison highlights the fact that distillation does not necessarily affect all classes symmetrically, but it can have a disproportionate effect on some classes, which we will attempt to explore more through explanations.

By using the same set of sentences (the 10 reviews described earlier), we generate explanations using SHAP for both BERT and DistilBERT. Below are the results for the feature importance for both models.

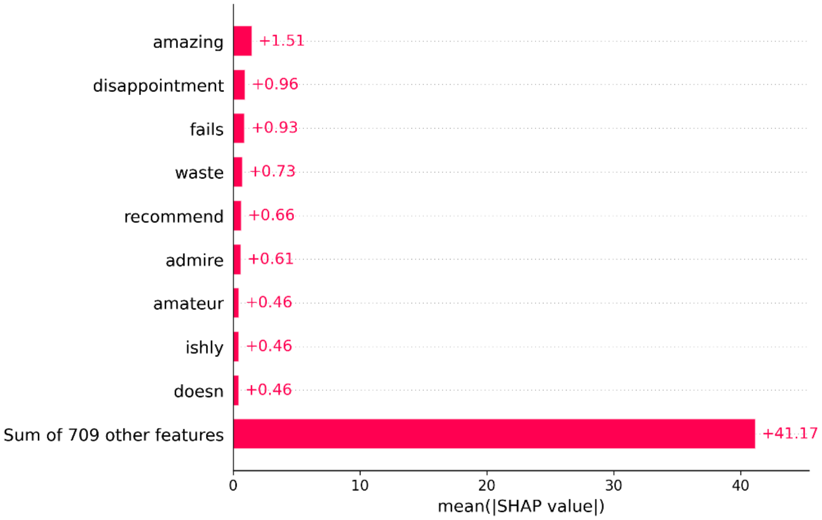

Figure 1: BERT Feature Importance

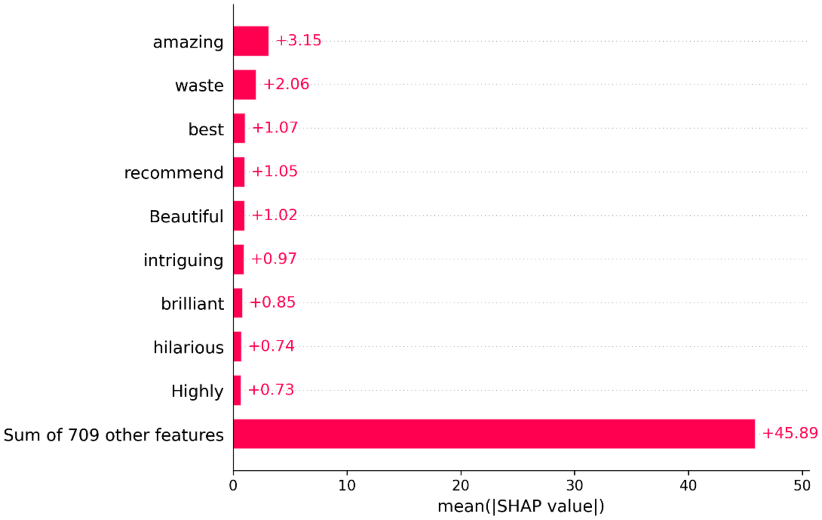

Figure 2: DistilBERT Feature Importance

As figures 1 and 2 show, we see some of the words with the largest importance appear in both models, and the two models also share the most important feature or word. However, we notice that across the board, the magnitude of the importance seems larger for DistilBERT. To explore this further, we look at the distributions of all the SHAP values generated for the two models for the same sets of inputs.

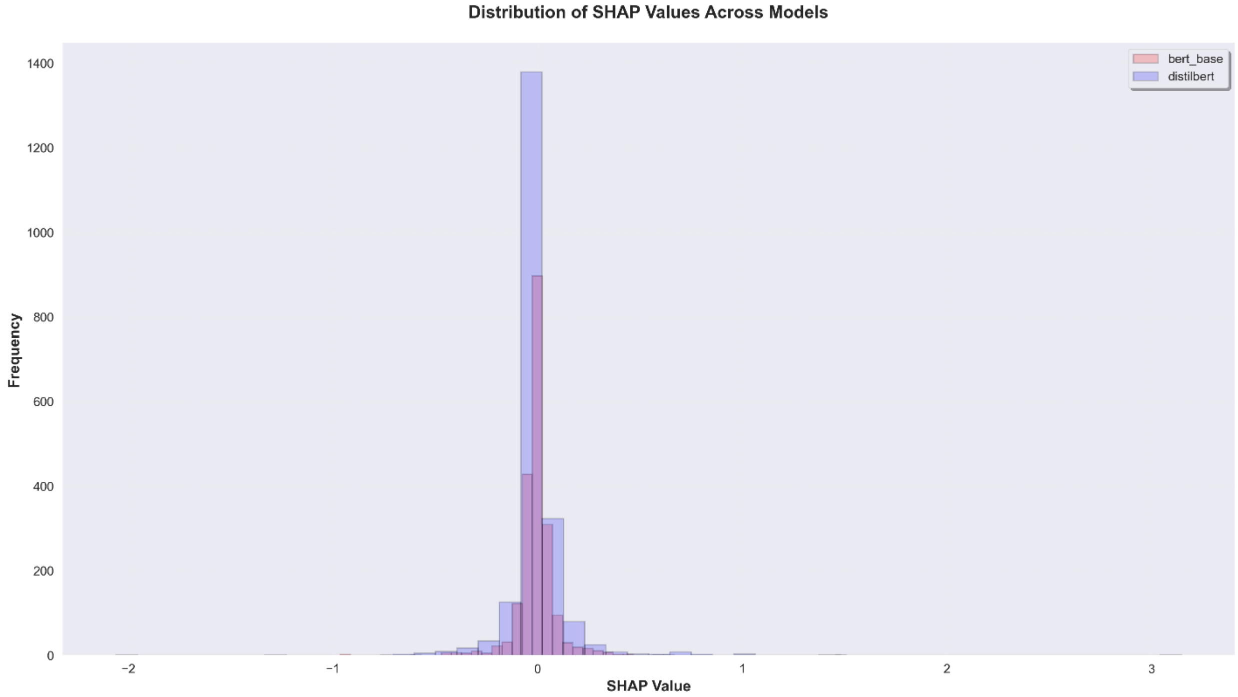

Figure 3: Distribution of SHAP Values generated for the same set of inputs for BERT and DistilBERT

Figure 3 confirms our doubts, showing that although the SHAP values of both models span approximately the same area, they do have different magnitudes with DistilBERT almost always having higher values than BERT. This can be potentially attributed to the underlying distillation process. Model distillation in deep neural networks aims to transfer the knowledge from a large, complex model (teacher) to a smaller, more efficient model (student). This process involves training the student model to replicate the teacher model's behavior using soft labels, which are probability distributions rather than hard labels. By mimicking the teacher, the student model aims to achieve comparable performance while being lightweight and faster.

For the sentiment classification task at hand, DistilBERT having higher SHAP values than full BERT can be attributed to the model distillation process. DistilBERT, being a distilled version of BERT, focuses on capturing the most salient features from the input data, leading to more concentrated and aggressive interpretations of the input features, since less salient features that can tone down the predictions are not taken along in the lighter model. This can result in higher SHAP values, indicating stronger contributions of individual features to the model's predictions.

To further examine this phenomenon, we get the difference between the SHAP values attributed to all features or tokens for both models, and then we get the difference between these and observe the top 20 features that have the largest differences between their BERT SHAP values and DistilBERT SHAP values.

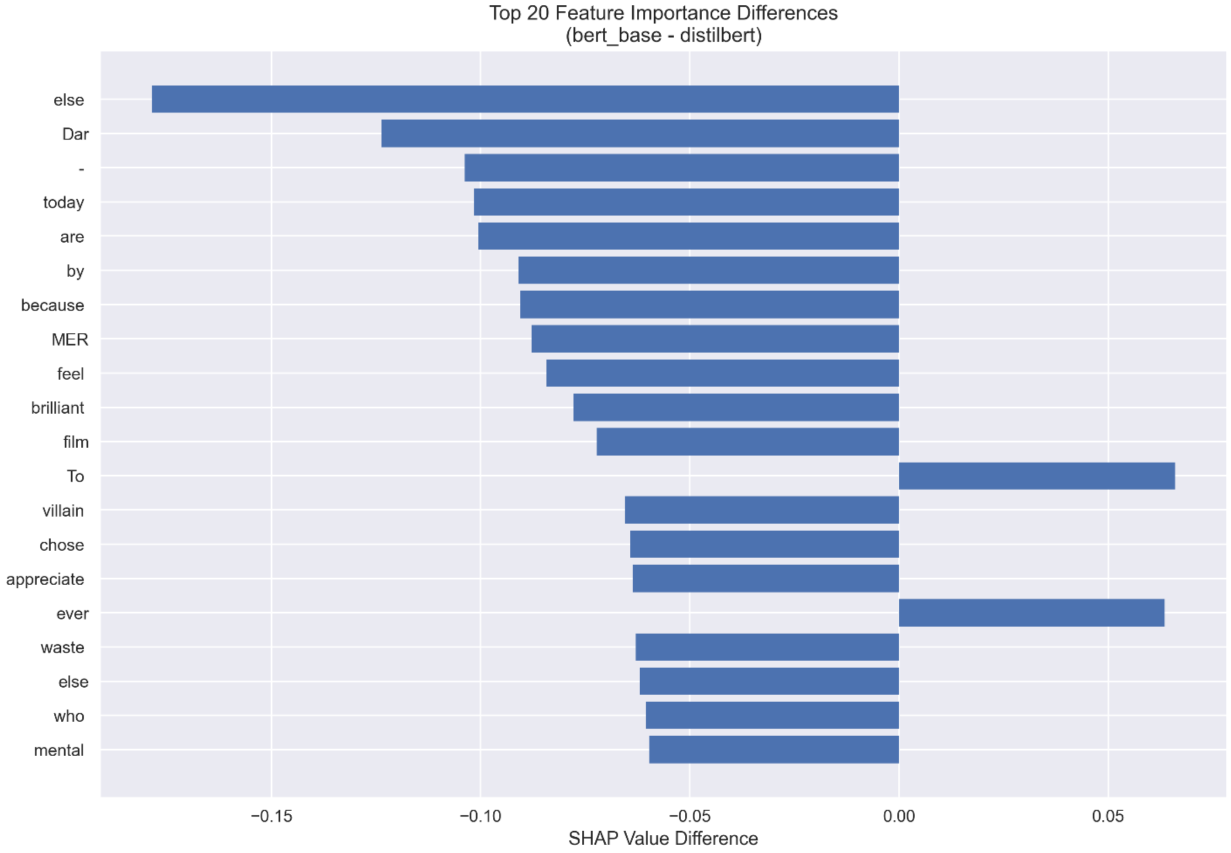

Figure 4: Feature Importance Differences between BERT and DistilBERT

As figure 3 shows, when features are ranked by the difference they exhibit between BERT and DistilBERT, we see that the values are almost always negative in the first 20 features, indicating that DistilBERT almost always assigns greater importance to more salient features than BERT to effectively Distill it as discussed earlier.

6. Discussion

The finding that DistilBERT assigns higher SHAP values to features compared to BERT provides insight into the mechanics of model distillation. While the student model appears to effectively replicate the teacher model's predictions with greater computational efficiency, this raises concerns about the trade-offs involved in this process. Model distillation, by design, prioritizes salient features while often discarding less prominent ones, effectively "distilling" the knowledge of the larger model. This can be likened to a crash course that enables a student to grasp the essentials of a subject without delving into its complexities or subtleties. While this approach improves efficiency, it raises questions about the reliability of these models in high-stakes decision-making contexts.

For instance, in a sentiment analysis task, DistilBERT demonstrates strong alignment with BERT on clear-cut cases of positive or negative sentiment. However, it begins to falter in more nuanced scenarios, such as inputs with multiple negations or subtle linguistic intricacies. While this discrepancy may be inconsequential in tasks with low stakes and inherent ambiguity—such as sentiment analysis, where even human annotators often disagree—it could have profound implications in domains where precision and fairness are paramount. Examples include credit scoring, job application screening, or medical imaging, where the cost of errors can be significant.

The observed tendency of the distilled model to emphasize the most salient features while de-emphasizing or pruning less significant ones may exacerbate bias issues. Deep neural networks are already known to inherit biases from the datasets on which they are trained. In the case of DistilBERT, the compression process could disproportionately affect underrepresented or less frequent data points. By focusing on optimizing accuracy and reducing complexity, the model risks disregarding subtle, less frequent representations in the data. This can lead to poorer generalization in sparse regions of the decision boundary—analogous to neglecting minority groups or edge cases in human contexts.

The implications extend beyond bias. Distillation and other compression techniques, such as quantization, are often presented as efficiency optimizations. However, they also involve a significant reduction in the contextual richness captured by the original model. This raises critical questions about the suitability of using distilled models for tasks requiring nuanced reasoning or intricate contextual understanding. How do we define the acceptable trade-off between computational efficiency and interpretive fidelity? What criteria should guide the decision to distill or quantize a model for specific use cases?

These concerns call for further research into how feature representations are altered during distillation and quantization, particularly in high-stakes applications. A deeper understanding of these processes is necessary to ensure that models maintain robust performance and fair treatment across diverse scenarios. Additionally, developing benchmarks to assess the appropriateness of model size and context for specific tasks could help guide the responsible deployment of distilled or compressed models.

7. Conclusion

This study investigated the impact of model distillation on the interpretability of deep neural networks, with a focus on sentiment classification tasks using BERT and its distilled counterpart, DistilBERT. By employing SHAP to analyze feature attributions, we uncovered meaningful differences in how the two models assign importance to input features.

Our findings reveal that while both models identify similar key features, DistilBERT consistently assigns higher magnitudes of importance. This behavior is likely a direct result of the distillation process, which prioritizes capturing the most salient features while discarding less critical information. The increased SHAP values in DistilBERT reflect a more concentrated attribution to specific input features, potentially due to its streamlined architecture.

These insights emphasize the need for careful consideration when deploying compressed models in high-stakes applications. Although the computational efficiency and accuracy trade-offs of distillation are well-documented, this study highlights the qualitative changes in interpretability that can arise. Understanding these changes is crucial for ensuring trust and accountability in real-world deployments, particularly when compressed models are used in sensitive or decision-critical domains.

Future work can probe deeper into the impact of compression on feature representations and it could extend this analysis to other compression techniques, such as quantization or pruning, and explore their effects on interpretability across diverse tasks and datasets. By building a deeper understanding of the interplay between compression and model explanations, we can better navigate the trade-offs inherent in deploying efficient yet reliable machine learning models.

8. References

- 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.

- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1631-1642).

- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

- Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems arXiv:1705.07874v2

- Kokalj, E., Škrlj, B., Lavrač, N., Pollak, S., & Robnik-Šikonja, M. (2021). BERT meets Shapley: Extending SHAP Explanations to Transformer-based Classifiers. In Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation (pp. 16-21).

- Guo, Y. (2018). A Survey on Methods and Theories of Quantized Neural Networks. arXiv preprint arXiv:1808.04752.

- Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.

- I used the help of Copilot in debugging and organizing my code, since I started it in a notebook then switched to scripts. I also used ChatGPT to help proofread and organize the report and used Perplexity to format the references.