Reading Between the LIMEs: Understanding the Transfer of Language Understanding Through Logit- and Attention-Based Distillation

Hanna Lee
Stanford University
hylee719@stanford.edu

Zhiyin Lin
Stanford University
zhiyinl@stanford.edu

Joy Yun
Stanford University
joyyun@stanford.edu

Abstract

Previous work has shown that knowledge distillation is an effective technique for improving small-model natural language benchmark performance by training it to emulate a large teacher model's predictions. There is little clarity around whether a teacher model's deep understanding of language is truly distilled and represented within student models or if the student models are simply learning heuristics to match the teacher's outputs. In this paper, we explore whether using attention-based knowledge distillation, in place of "vanilla" logit-based distillation, can guarantee a more thorough transfer of linguistic logic and understanding from teacher to student. We provide evaluations of alignments using quatitative metric Cohen's Kappa and qualitative LIME analysis. We find that 1) models can attain the same performance accuracy on the CoLA benchmark with very misaligned reasoning, 2) performing attention-based distillation using different attention layers can lead to significantly different logical alignments between the student and teacher model, and 3) attention-based distillation using first-layer attention most effectively transfers feature-level reasoning from the teacher to the student.

1 Introduction

As large language models continue to grow in size and capability with modern day technology, the importance of considering deployment feasibility and ethical energy usage grows at an equally rapid rate. Knowledge distillation is a technique that involves training a small, high-performing student model off of the outputs of a large, teacher model in place of the one-hot labels normally used for training (Hinton et al., 2015). The intuition here is that the rich knowledge and more specific language rules learned by the larger teacher model can be passed down to the student model, allowing for a more effective training of the student model.

Previous research has shown that knowledge distillation effectively trains higher-performing small-models on NLP benchmark tasks, such as in the case of DistilBERT (Sanh et al., 2020), a distilled version of the larger model BERT (Devlin et al., 2019) that is 60% of the size while retaining 97% of BERT's linguistic understanding. A deep natural language understanding comes from the sheer complexity of large language models (Tamkin et al., 2021), but there is still uncertainty regarding how the same complexity of understanding is contained within a smaller model of less parameters (Belinkov and Glass, 2019). Bender and Koller (2020) offers the idea that student models are simply mimicking the behavior of their teachers by learning heuristics in place of true linguistic rules. From an NLU standpoint, while knowledge distillation optimizes for small-model end-performance, previous work provides little clarity into how or what knowledge is being ‘distilled‘.

In our efforts to understand the internal workings of large language models (Belinkov and Glass, 2019), knowledge distillation further abstracts the representation of this same knowledge within student models. Driven by the desire for interpretable student models that reflect the true natural language understanding of their teachers rather than linguistic shortcuts, in this paper, we will explore knowledge distillation using a teacher model's attention map, rather than logits. In this way, we hope there will be more of a focus on transferring over a teacher model's logical framework in addition to its final answers. We will use accuracy, Cohen's Kappa coefficients, and LIME (feature attribution) scores to evaluate agreement and the similarity of logical frameworks shared between the teacher and student models, using logit ("vanilla") and attention-based distillation.

2 Prior Literature

Current literature in the field of model interpretability explores using feature attribution as a useful tool to give insight into the inner workings of a model. (Sundararajan et al., 2017) introduces Integrated Gradients as an attribution method used to assign importance scores to the input features of a machine learning model for a specific prediction, keeping the axioms of model sensitivity and implementation invariance in mind.

Ribeiro et al., 2016 emphasize that model transparency is crucial in developing trust in both a model's predictions and its behaviors. The authors introduce LIME, an algorithm that approximates classifiers locally with interpretable models in order to explain their predictions, and SP-LIME, which is intended to address the model trust problem by selecting a set of representative explanations via submodular optimization.

Shrikumar et al., 2017 also attempts to address the "black box" nature of neural networks with DeepLIFT, (Deep Learning Important FeaTures). It decomposes the output prediction of a neural network for a specific input by backpropagating the contributions of all neurons in the network to every feature of the input, essentially providing importance scores to inputs. DeepLIFT can separately account for the effects of positive and negative contributions at non-linearities, which can lead to finding dependencies that could be overlooked by other methods.

3 Data

We use the The Corpus of Linguistic Acceptability (CoLA) dataset (Warstadt et al., 2018), which is one of the nine GLUE benchmark tasks. The public version of CoLA contains 10657 sentences from 23 linguistics publications, which are annotated by experts for acceptability (whether it is grammatically correct or not) by their original authors. 9594 sentences in CoLA belong to the training and development sets, and the other 1063 sentences make up a held out test set. In the HuggingFace version (HuggingFace, 2023) of the CoLA dataset, which we used for our project, the data samples have the following format, with a label of 1 being an acceptable sentence, and 0 being unacceptable:

sentence(string)	label	idx
"They drank the pub dry."	1	17
"They drank the pub."	0	18

Table 1: An example of CoLA labeling.

4 Methods

4.1 Models

In this paper, we will work off of the pretrained BERT (Devlin et al., 2019) and DistilBERT (Sanh et al., 2020) models. While significant interpretability work has been done to explore the inner-workings of BERT (Rogers et al., 2020), less is known about DistilBERT's representations and decision processes, especially given its comparative performance to BERT on natural language benchmark tasks.

BERT-base has 12 Transformer layers and 12 attention heads per layer, while DistilBERT has 6 Transformer layers and 12 attention heads per layer. Since the two models share the same BERT-tokenizer and have the same number of attention heads, their attention maps can be directly compared for similarity on a given input. However, it is important to note that since the models have differing numbers of layers, we decided it would be most informative to compare attention maps for the first and last layers, as the representations for the the middle layers is abstracted away. More discussion regarding using attention from the first and last layer will follow in the results and analysis section.

4.2 Metrics

We use two metrics - accuracy and Cohen's Kappa to evaluate our models. Considering the motive for this exploration is to better understand what features BERT and DistilBERT are paying attention to in their CoLA predictions and to what extent they agree with each other, maximizing accuracy was not a main objective like in many works in this space. We report the model accuracies on the CoLA validation set to show that the models share a similar relative performance.

The Cohen's Kappa rater-similarity metric was first introduced in 1960 by Jacob Cohen in the journal of Educational and Psychological Measurement. Instead of simply reporting the percentage of predictions that two raters (models) share, Cohen's Kappa considers the fact that raters may happen to agree on some labels purely by chance. The score is calculated using the formula below:

$$k = \frac{p_o - p_e}{1 - p_e}$$

where,
$k =$ Cohen's Kappa coefficient
$p_o =$ probability of agreement between raters
$p_e =$ probability of chance agreement

The Cohen's Kappa coefficient will range between 0 and 1, where 0 indicates no agreement between raters and 1 indicates perfect agreement.

4.3 Baselines

Our baseline for model performance on the CoLA benchmark task is determined by the performance accuracies of BERT and DistilBERT after being finetuned on CoLA, which were 82.74% and 82.07%, respectively. Additionally, our baseline for agreement between BERT and DistilBERT's predictions on the CoLA validation set, using Cohen's Kappa coefficient, is calculated between the BERT and DistilBERT models, both independently finetuned on the COLA dataset, giving us -0.0003.

5 Knowledge Distillation

Knowledge distillation (KD) involves training a student model using the internal representations of a larger, teacher model, with the intention of the rich knowledge from the teacher improving the effectiveness of student training. Knowledge distillation does not always lead to perfect emulation of teacher predictions or reasoning by the student (Stanton et al., 2021), and in this paper we explore how different methods of knowledge distillation affect this. The teacher model is BERT-base, loaded in as `bert-base-cased` on HuggingFace, that we finetuned on CoLA. The student model we are finetuning is DistilBERT, loaded in as `distilbert-base-cased` on HuggingFace. It is important to note and should be kept in mind during analysis that the pretrained DistilBERT model has already been distilled from BERT-base, before being distilled from the CoLA-fintuned BERT in our experiments.

5.1 "Vanilla" Distillation Using BERT Logits

We first finetune a version of DistilBERT using "vanilla" distillation, which involves matching a student model's logits to a teacher model's predictive probabilities. We use the cross-entropy loss between these values (Sanh et al., 2020):

$$\mathcal{L}_{dist} = \sum_i t_i * \log s_i$$

where,
$t_i =$ teacher probability for class i,
$s_i =$ student's probability for class i.
This distilled version of DistilBERT serves as a baseline distillation performance to compare our later attention-distillation methods to.

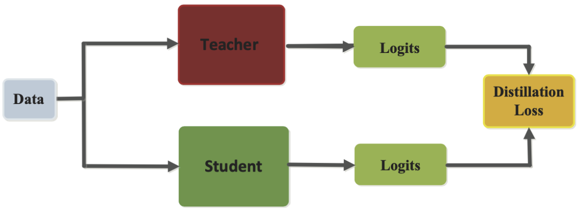

Figure 1: Flow of Knowledge Distillation | Source: (Gou et al., 2020)

5.2 Distillation Using First-Layer Attention Maps

We finetune a version of DistilBERT using attention-based distillation by taking the cross-entropy loss between the student and teacher model's **first-layer** attention maps for a given input, instead of the logits used above. The intuition for using the first-layer attention maps is that, at this stage, the internal representation of the input most directly connects with interpretable features in the original sentence. This is a valuable point of alignment between the teacher and student models because the linguistic reasoning established here propagates into the remaining layers and attention maps of the model.

5.3 Distillation Using Final-Layer Attention Maps

We also finetune a version of DistilBERT by taking the cross-entropy loss between the student and teacher model's **final-Transformer-layer** attention maps. The intuition for this method of distillation was that the final-layer attention map attends to the latest representation of the input across all attention layers, revealing most about how a model reaches its final output.

6 Results

We report the accuracy and the Cohen's Kappa of baselines and knowledge distillation models in the

Model	$Accuracy$	$Cohen'sKappa$
BERT	82.74	1.0000
DistilBERT	82.07	-0.0003
DistilBERT KD (logits)	82.84	0.7666
DistilBERT KD (attention last layer)	82.07	0.0201
DistilBERT KD (attention first layer)	82.65	0.7567

Table 2: Summary statistics of the models from top to bottom: BERT finetuned on CoLA, DistilBERT finetuned on CoLA, DistilBERT with logits-based knowledge distillation, DistilBERT with attention(last layer)-based knowledge distillation, and DistilBERT with attention(first layer)-based knowledge distillation. Cohen's Kappa is measured with respective to BERT.

summary statistics table below (Table 2). It is worth noting that the accuracy performance are similar across models, but the Cohen's Kappas with respect to BERT are drastically different. Since we use BERT as the ultimate teacher, its perfectly agrees with itself and thus has a Cohen's Kappa score of 1.0. The rest of the Cohen's Kappa discussion is as follows:

6.1 Ill-performed DistilBERT baseline

To start with, the baseline of our teacher-student paradigm is a teacher-less student. In this paper, we use DistilBERT pre-trained weights finetuned on CoLA as the baseline student. According to Table 2, baseline DistilBERT's Cohen's Kappa coefficient is slightly below zero, statistically indicating no agreement between itself and BERT. With the poor alignment between DistilBERT and BERT, we realize that they make mistakes on different data samples and probably have misaligned reasoning frameworks (more on this in section 5.2).

6.2 Vanilla KD: successful yet questionable

DistilBERT distilled with logits loss and DistilBERT distilled with attention first-layer loss both achieve a Cohen's Kappa score over 0.75, which statistically counts as substantial agreement with the teacher BERT. But, we suspect that their pathways to success are distinct. On one hand, the logits, by definition, characterize a distribution of predictions over all classes and occurs in the penultimate layer of a model. It is a more informational version of the output. Using aligned logits as an objective directly forces two models to match their predicted distribution over all classes in the penultimate layer. However, it remains unclear how to ensure the alignment of outputs propagates to the alignment of model reasoning process. With the nature of the logits approach implicitly being to retrieve feedback from some form of the model outputs, the logits approach could be blunt and essentially a shortcut. It remains unclear on how aligned these models' actual reasoning processes are.

6.3 Attention KD: versatile yet variable

As such, we have pondered this issue lengthily, and we hereby present experiments on attention-based knowledge distillation, hoping it opens up the blackbox of the teacher model's thinking process. When aligning the last layer attention weights, we arrive at a Cohen's Kappa score of 0.0201, indicating almost no agreement. However, if we align the first layer attention weights, we obtain a score of 0.7567 which is similar to that of the logits knowledge distillation approach. This drastic variation of alignment performance indicates the versatility of the attention approach.

Last layer attention alignment fails because the teacher and the student are already reasoning on a different playground when they reach the last attention layer, as the previous attention layers are not forced to be aligned. Suddenly forcing last attention layer alignment might not impact much performance and for the same reason, aligning the first layer of attention map might be more effective.

Since the first attention layer works with tokens that are fresh from inputs, they have not yet diverged from later transformations. As the Cohen's Kappa score indicates, using first layer attention alignment for distillation is much more effective than that for the last layer.

If guided early on, models are more likely to take the same correct path. Despite that this also depends on task difficulty and the number of potential pathways to the correct solution, early guidance is important for alignments.

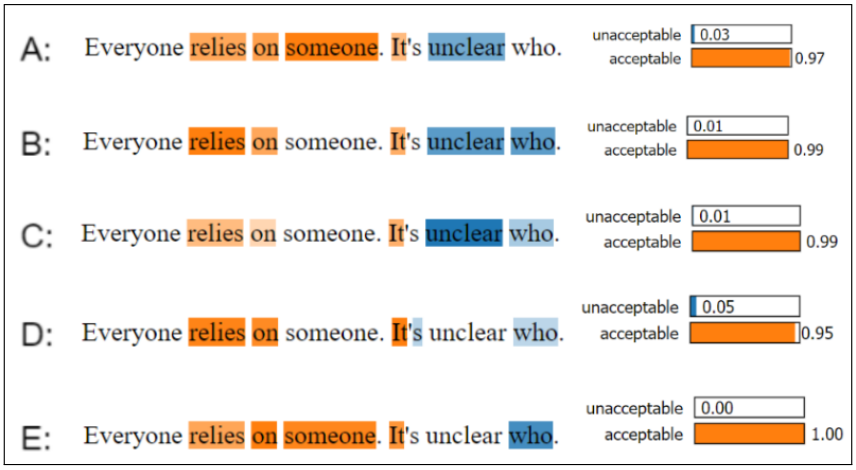

Figure 2: LIME visualizations from one out-of-domain sample used for interpretation. The models represented are as follows: **A**- BERT Finetune, **B**- DistilBERT, **C**- DistilBERT KD (logits), **D**- DistilBERT KD (attention last layer), **E**- DistilBERT KD (attention first layer). Words highlighted in **blue** contribute to the "unacceptable" class, while words highlighted in **orange** contribute to the "acceptable" class. The darker a word is highlighted, the more weight it has. Each sentence is followed by the model's overall prediction probabilities for that sentence. Visualizations are edited for clarity and ease of comparison.

7 LIME Visualizations

7.1 Motivation

In addition to using accuracy and Cohen's Kappa as quantitative metrics for comparing the results of different versions of DistilBERT and BERT, we wanted to qualitatively analyze intuitive feature attributions for several sample sentences. Including feature attribution in our analysis contributes to our multi-faceted approach for exploring how models learn from each other and how this can be done meaningfully. We chose LIME (Ribeiro et al., 2016) in order to provide interpretable visualizations of attribution scores that could be easily compared across different models. LIME samples instances, gets predictions using the model's decision function (in this case, from the logits), and weighs them by the proximity to the instance being explained.

7.2 Selected Samples

Four human-annotated, out-of-domain samples from CoLA (Warstadt et al., 2018) were chosen to be tested among all the models based on the following criteria: 1) 2 samples had true labels of "acceptable", and 2 samples had labels of "unacceptable", and 2) For each label, 1 sample was a consensus among all the human annotators and true label, while the other sample generated conflicting answers (at least 2 human annotators disagreed with the others).

We did this to have an equal analysis of both unacceptable and acceptable sentences, in addition to garnering insight into possible effects of "easier" vs "harder" samples. We sampled from out-of-domain samples to evaluate feature attributions that are somewhat distinct from the ones covered in the CoLA training and validation sets. A feature size of 5 was chosen based on the average word length of the samples.

7.3 Results

After closely examining the results of these 4 samples across all models, we found that the finetuned BERT model seemed to be in more agreement with the DistilBERT first layer attention model, in terms of the contributing words, as well as prediction probabilities. This makes intuitive sense, as we conjecture that the attention to input features maps to words more clearly in the initial layers, while the relation between attention scores and words becomes more abstract in the later layers.

On the other hand, we generally observed that the logit-distilled DistilBERT model yielded similar feature attribution scores to the finetuned distilBERT model. This may indicate that the logit distillation method did not result in a significant change in distilBERT's internal "reasoning".

Figure 2 illustrates the LIME attributions for each of the five models on one sample. In this example, the true label of the sentence is "acceptable". As we can see, all five models predicted "acceptable" with high confidence. BERT Finetune (Model A) and DistilBERT KD (attention first layer) (Model E) place high importance on the words "someone" and "on" (a pronoun and preposition). Conversely, both the DistilBERT (Model B) and DistilBERT KD (attention last layer) models place the highest importance on the words "relies" (a verb).

7.4 Analysis

Based on some previous work in grammar in the NLP field, specifically part-of-speech tagging, three distinct part-of-speech tags can be considered to be the most important for formulating grammar rules: nouns, verbs, and modals (a type of auxiliary/helping verb). (Brill, 1995) In the given example presented in Figure 2, all five models placed importance on the verb "relies", and the nouns "someone", "who", and "it" were all considered consequential by at least 2 different models.

This provides some promising evidence that the DistilBERT models we investigated have some intuition for focusing on the more crucial parts of speech for determining whether a sentence is grammatically correct or not. Additionally, we see that DistilBERT KD (attention first layer) may be the most promising out of all 4 distilled models in terms of transferring grammatical reasoning from BERT, based on its higher feature attribution similarity.

8 Conclusion

In summary, this project investigated the effects of numerous distillation techniques with the CoLA benchmark, including knowledge distillation (KD) with logits, KD with last layer attention, and KD with first layer attention. Given the somewhat nebulous nature of whether the knowledge transfer during distillation is actually conveying a deep understanding of language, our iterative exploration of different attention-based distillation methods sheds some light on how language understanding can be transferred between models effectively.

Our findings elucidate the ability of DistilBERT KD (attention first layer) to transfer word-level feature understanding knowledge from BERT to DistilBERT by using the first layer attention scores, according to our Cohen's Kappa annotator agreement metric and LIME analysis. Despite the similar accuracies between distilled models, we discovered great variations in how models approach the CoLA dataset, reflected by the varying Cohen's Kappa scores. The annotator agreement between BERT and DistilBERT is found to be very low, but the Cohen's Kappa score increases greatly using attention first layer distillation compared to attention last layer distillation, which indicates a higher word-level grammar understanding transfer when using earlier layers.

Our LIME analysis further suggests higher word-level understanding transfer using first layer attention, compared to using last layer attention. The feature attribution scores and the words chosen by each model provide a greater sense of interpretability for these various distillation approaches in evaluating the grammatical correctness of a sentence.

Future work may include testing on other datasets that show significant gaps in performance between BERT and DistilBERT, as well as examining our distillation techniques for other models such as RoBERTa. We would also aim to evaluate other methods of distillation, in addition to testing different quantitative and qualitative metrics for agreement and feature attribution for a more comprehensive analysis.

9 Failures

Before arriving at this paper, we took multiple routes to study knowledge distillation, and failed. Below, we listed some attempted ideas and the mistakes that we made.

9.1 Feature attribution loss

The initial idea of this project was to use integrated gradients feature attribution joint loss during the knowledge distillation process. This was inspired by the aforementioned literature on feature attribution techniques and their significance. The idea was that if we could generate feature attribution scores that somewhat capture a model's understanding of grammar, these could be used during training of the distilled model to impart some additional insight. However, after some experimentation, we saw that since the feature attribution scores calculated through the integrated gradients method are with respect to the inputs, and backpropagation is concerned with gradients with respect to the weights, this approach is perhaps not the most intuitive or compatible with formulating a training loss.

9.2 RoBERTa can't teach DistilBERT

One should be reminded that differing BERT-based models have different architectures, tokenizers, and more. As a result , it is challenging to for some distillation methods to use teacher and student of different a properties. One such example is RoBERTa and DistilBERT, since they have different tokenizers. Additional designs in how the teaching works are needed to run such experiments.

Known Project Limitations

We want to point out some limitations of the current project for future NLP practitioners who seek to make use of the data, models, or findings of this paper.

In retrospect, we could have take more holistic considerations on what makes good model alignments. In this project, it is primarily evaluated based on the Cohen's Kappa metric and analyzed via LIME. The Cohen's Kappa metric is measured on the output predictions (i.e. 0 or 1 since CoLA is a binary classification task). In addition, LIME analysis gives a qualitative lexical evaluation of alignment. Nevertheless, this objective could be expanded further to examine different alignments aspects. Some future ideas include layer-level alignment for model-centered alignment, output-level alignment for different input types (gender-biased, quantitative, style, etc.) for content-centered alignment.

Moreover, it is worth keeping in mind that the empirical results are specific to BERT-DistilBERT as the teacher-student pair. Though one could speculate reasonable generalizations to neighboring language models such as RoBERTa-DistilRoBERTa and more, it is hard to say for sure without empirical experiments. Particularly, differing model architecture could affect the results of the proposed alignment methods.

Authorship Statement

All three team members made even efforts on arriving at this final product. The three authors set up, ran both baselines and knowledge distillation experiments, and iterated together. Some specific efforts include: Hanna carried out LIME analysis for model results. Zhiyin built in the attention-based knowledge distillation. Joy worked out Cohen's Kappa evaluation.

We owe a special thank you to Christos Polzak for his generous help on starter code, debugging, and advice. All three group members contributed equally toward the writing of the paper. Finally, we used AWS EC2 GPUs to train our models, and we appreciate how CS224U provide us with sufficient AWS credit, as we did not encounter any GPU out of credit, memory, or speed issues throughout the experimental process.

References

Yonatan Belinkov and James Glass. 2019. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. 2020. Knowledge distillation: A survey. *CoRR*, abs/2006.05525.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

HuggingFace. 2023. The glue cola dataset. `https://huggingface.co/datasets/glue/viewer/cola`.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how BERT works. *CoRR*, abs/2002.12327.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A. Alemi, and Andrew Gordon Wilson. 2021. Does knowledge distillation really work? *CoRR*, abs/2106.05945.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks.

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2018. Neural network acceptability judgments.