Are Distilled Models Just Deep-Sea Octopi? Probing Linguistic Representations of Distillation-Finetuned Models

Stanford CS224N Custom Project (Mentor: John Hewitt; no external collaborators, mentors, or shared classes)
All code available at `https://github.com/zy-f/nlp-distill-probing`.

Christos Polzak
Department of Electrical Engineering
Stanford University
clcp@stanford.edu

Joy Yun
Department of Computer Science
Stanford University
joyyun@stanford.edu

Abstract

While previous work has shown that knowledge distillation improves small-model performance on NLP benchmark tasks (Hinton et al., 2015), it is still unclear how smaller models learn from the knowledge of their teachers (Belinkov and Glass, 2019). Given that achieving a deep understanding of linguistic properties typically relies on the complexity of large language models (Tamkin et al., 2021), we explore whether DistilBERT models finetuned with distillation on natural language inference (NLI) are really learning deep language rules, or if they are simply picking up on heuristics they can use to mimic their teacher's (BERT's) outputs. We first verify that the in-distribution gains from finetuning with distillation generalize to other NLI datasets. Then, through function-word NLI probing and word-level edge probing, we demonstrate that during NLI distillation finetuning, student DistilBERT models do absorb understanding of linguistic properties from their teacher, both in positive and negative ways. In particular, we find that the gains in generalized NLI performance provided by distillation finetuning are at least partially because distillation improves DistilBERT's understanding of function words.

1 Introduction

As the capabilities of neural language models grow with modern access to data and computational resources, the increased size of such models also makes deployment to large numbers of users difficult. Knowledge distillation (Hinton et al., 2015) is a popular method of creating smaller high-performance models by using the predictive outputs of a frozen, pre-trained large model to try and "teach" a much smaller "student" model to match the outputs of the larger "teacher" model. The idea behind distillation in NLP is that incorporating a good teacher's predicted outputs, rather than only the strict one-hot labels provided by a dataset, captures nuances and uncertainty inherent in trying to label human language and brings it into the training process.

A popular distilled model is DistilBERT (Sanh et al., 2020) – a smaller, cheaper, and more general-purpose version of BERT (Devlin et al., 2019) trained by distilling BERT base. The original paper shows that even with a 40% size decrease from BERT, DistilBert still retains 97% of BERT's natural language understanding capabilities. While knowledge distillation has been proven to improve small-model performance on NLP benchmark tasks (Hinton et al., 2015), it is still unclear how smaller models learn from the knowledge of their teachers (Belinkov and Glass, 2019). Given that achieving a deep understanding of linguistic properties has greatly relied on the complexity of large language models (Tamkin et al., 2021), and inspired by the octopus thought experiment that compares statistical mimicry to true understanding of meaning (Bender and Koller, 2020), in this paper we explore whether distilled models are really learning deep language rules or if they are simply picking up on heuristics they use to mimic their teachers' outputs.

Using pretrained DistilBert as our student model, we finetune on the distilled knowledge from two versions of Bert finetuned on natural language inference – one that performs well on overall language understanding and the other on NLI-specific benchmarks. An effective distillation of knowledge would lead to the student model directly reflecting the strengths of its teacher, while a deceptive distillation could take the form of a student performing better on the benchmark tasks but showing no improvement in the areas the teacher was intended to teach.

To measure the type of knowledge the distilled model learns and to better understand its inner-workings, we probe (Belinkov, 2021; Hewitt et al., 2021) both versions of finetuned BERT and DistilBert, before and after finetuning, on a variety of syntactic and semantic tasks to evaluate the true extent to which distilled models retain the language understanding of their teachers.

2 Related Work

Existing work has explored probing as an effective method of evaluating a language model's true abilities to understand language. Specifically, conditional probing introduced by Hewitt et al. (2021) and edge probing introduced by Tenney et al. (2019) are amongst the most prevalent. Conditional probing captures new information in layers that did not carry over from a chosen baseline layer (e.g. embedding layer) regarding a specific linguistic property being examined. It has revealed that knowledge of simple syntactic properties, like part-of-speech, carry deeper into the layers of large language models than previous results have shown. Edge probing covers a selection of NLP-specific tasks that target a model's word-level contextual representations for syntactic and semantic knowledge.

Kim et al. (2019) introduces another type of NLI word-probing in the form of nine challenge tasks that each target a different dimension of natural language inference understanding (i.e. negation, comparatives, spacial expressions, etc.). Kim et al.'s research shows that model performance, across a breadth of linguistic tasks, depends closely on the objective for which it was pretrained.

Previous papers have also shown that finetuning BERT on downstream tasks leads to "catastrophic forgetting" from pretraining (Durrani et al., 2021). Merchant et al. (2020) finds that while finetuning BERT leads to a significant change in its knowledge representations, the changes are not negative but rather restructure the knowledge to be more tailored to the task. Hościłowicz et al. (2023) reports that, in general, probing is an insufficient metric for exploring model intepretability due to difficulty and inconsistency when decoding results of current probing methods.

3 Approach

Distillation. In this paper, we perform "vanilla" distillation, which only matches the logits of the teacher and student model. To do this, we compute a cross-entropy loss between the student's logits and the teacher's probabilities: $\mathcal{L}_{dist} = \sum_i t_i * \log s_i$, where t_i is the teacher's probability for class i, and s_i is the student's probability for class i (Sanh et al., 2020). We define the output probabilities as $p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$, where p_i is the softmax-temperature probability of class i, and z_i is the logit for class i. The temperature T smooths the class distribution; at evaluation, we set $T = 1$.
Sanh et al. (2020) also use a joint loss where we also include a weighted cross-entropy loss associated with the original ground truth $\mathcal{L}_{hard}$, resulting in a joint loss $\mathcal{L}_{joint} = \alpha_{hard}\mathcal{L}_{hard} + \alpha_{dist}\mathcal{L}_{dist}$. Our teacher model is the standard BERT pretrained on masked language modeling (`bert-base-cased` on HuggingFace), which we then finetune on mNLI. Our student is DistilBERT (`distilbert-base-cased` on HuggingFace). For both models, we make use of the official pretrained weights from masked language modeling (MLM) pretraining only, also from HuggingFace. Critically, we note that DistilBERT is also distilled during MLM pretraining.

Word-Level Edge Probing. We probe models by extracting their representations at different layers and then training an extremely simple model to predict linguistic properties based on those representations. In particular, we perform individual layer probing and conditional probing (Hewitt et al., 2021) as follows:
Let X be the vector representing our input tokens. Let $f_i(X)$ be the representation of X after the i-th layer of our model. In an NLP context, define baseline $B = f_0(x)$ as the representation of X after only token embedding. Given a probe model $\mathcal{P}_\theta$ with parameters θ, we can optimize θ on a **probing**

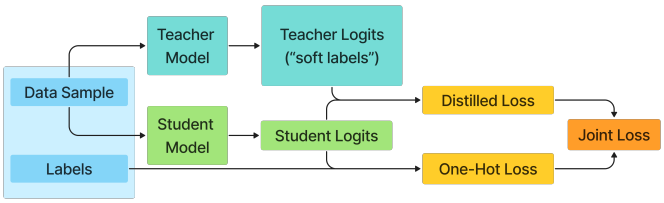


Figure 1: Flow of Knowledge Distillation

task with training dataset $\mathcal{D}_{tr} = \{(x_i, y_i)\}_i$, then evaluate the probe and the representations on a hold-out dataset $\mathcal{D}_{val} = \{(x_j, y_j)\}_j$. Let Perf$(R)$ be the performance of the probe on the hold-out dataset given some representation R. In individual-layer probing, we simply measure Perf$(f_i(X))$ for various i. However, in conditional probing, we attempt to capture the *new* information captured by a deep layer in the model relative to its baseline as Perf$([B; f_i(X)])$−Perf$([B; 0])$. In particular, our word-level probing tasks are: (i) coarse-grained universal part-of-speech (Nivre et al., 2020) (`pos`), (ii) Universal Dependencies edge prediction (`dep rel`), and (iii) named entity recognition (`ner`).

NLI Generalization. To try to determine why probe-measured changes to our models' understanding of linguistic properties matters, we evaluate our models out-of-distribution on other benchmark NLI datasets. In theory, we believe poorer understanding of part-of-speech, dependencies, named entities or NLI function words should be reflected by poorer generalization.

NLI Function Word Probing. To evaluate whether distillation-driven improvements in NLI benchmark performance are due to better understanding of language or ill-gotten gains from imprecise heuristics, we also evaluate our NLI-finetuned models on a suite of small challenge datasets that explicitly test for their understanding of linguistic properties. In particular, we evaluate on five tasks provided by Kim et al. (2019), which introduce small, function-word level mutations that invert entailing and contradicting hypothesis-premise pairs, as follows: (i) **prepositions** (`prep`), which swaps prepositions that are syntactically replaceable but semantically distinct (e.g. `with/without`); (ii) **quantification** (`quant`), which replaces quantifiers (e.g. `all/some`, `two/twenty`); (iii) **spatial** (`space`), which swaps relative object positions (e.g. `left/right`, `near/far`); (iv) **comparatives** (`comp`), which swaps expressions of differences in quantity or quality (e.g. `more/less`); and (v) **negation** (`neg`), which permutes of negations of the premise or hypothesis (e.g. inserting `not`).

Baselines Our word-level edge probing baseline is the probe accuracy of the popular pretrained BERT model (Devlin et al., 2019) (weights acquired from HuggingFace's `bert-base-cased`), as reported in Figure 2. We determine baseline accuracy on our finetuning and distillation task by finetuning BERT on the benchmark mNLI dataset (Williams et al., 2018) (Teacher `mnli_m` and `mnli_mm` in Table 1) and comparing it to the results from the original BERT paper (Devlin et al., 2019). We roughly match the original paper's scores of 84.6% and 83.4%. Finally, our baseline for NLI function word probing and generalization is the accuracy of BERT after mNLI finetuning (Teacher in Tables 1 and 2).

4 Experiments

4.1 Data

For natural language inference, we finetune using the benchmark multi-Natural Language Inference (mNLI) dataset (Williams et al., 2018) (available from HuggingFace). We evaluate on both the matched (`mnli_m`) and mismatched (`mnli_mm`) validation sets.
For `pos` and `dep rel` probing, we use the Universal Dependencies English Web Treebank (Silveira et al., 2014), available via `UniversalDependencies.org` (Nivre et al., 2020). Importantly, we perform minimal preprocessing, treating any subwords generated by BERT's WordPiece algorithm as independent words with the same label as the original word.

For `ner` probing, we use the OntoNotes V5 dataset (Pradhan et al., 2013), available on HuggingFace. In particular, we use the `english_v4` subset of the data, and then, due to memory constraints, randomly sampled 30% of the available train and validation data. As in the previous word-level tasks, we treat any subwords as independent words with their own BERT representations.
For the NLI function word probing, we simply evaluate NLI-finetuned models on the datasets provided by Kim et al. (2019). All available splits are concatenated and used to for evaluation.
Lastly, for generalization, we test on three NLI datasets: SNLI (`snli`) (Bowman et al., 2015), ANLI (`anli`)(Nie et al., 2020), and Jamaican Patois NLI (`jam_patois`) (Armstrong et al., 2022), all available through HuggingFace. In all three cases, we concatenate together all available splits and evaluate our NLI-finetuned models.
For clarity, we note that the function word datasets and Jamaican Patois NLI datasets are unusually small (<500 samples).

4.2 Evaluation method

For probing (both individual-layer and conditional), our Perf$(\cdot)$ function is simply validation set classification accuracy. Similarly, our mNLI performances are reported as the accuracies on the matched and mismatched validation sets.
For NLI function word probing and generalization, we report accuracy on the entire dataset.

4.3 Experimental details

For all models finetuned or distilled on mNLI, we use a batch size of 64 and no gradient accumulation, AdamW optimization, and train for 10 epochs, saving the model with the best average of mismatched and matched validation accuracy. During distillation, we use $T = 2, \alpha_{hard} = 2, \alpha_{dist} = 5$, matching DistilBERT's pretraining hyperparameters. Our probe is a two-layer feed-forward network with a hidden dimension of 45.

4.4 Results

Model	mnli_m		mnli_mm		snli		anli		jam_patois	
	UT	OT	UT	OT	UT	OT	UT	OT	UT	OT
Teacher (BERT)	82.5	84.0	83.4	83.9	76.6	78.0	64.8	66.3	51.2	53.7
Control (DistilBERT)	81.2	81.2	81.3	81.6	73.6	73.9	61.9	62.4	**49.9**	46.9
Student (DistilBERT)	**82.7**	**83.3**	**83.4**	**83.0**	**74.5**	**75.6**	**64.0**	**64.8**	46.8	**50.6**

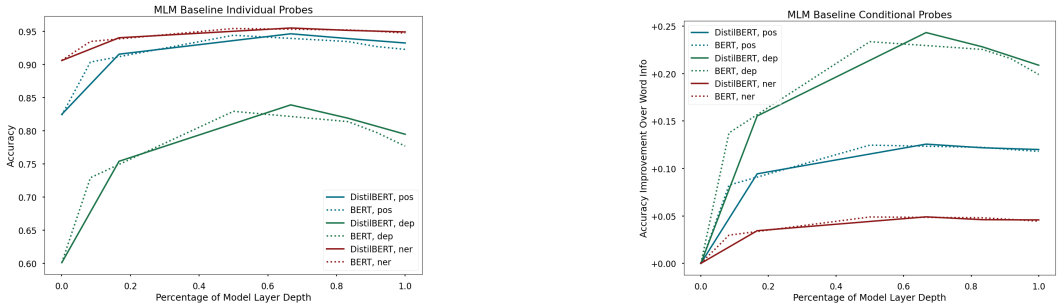
Table 1: NLI Benchmarks (UT: Undertuned, OT: Overtuned)

Distilled MLM Pretraining Probes. To evaluate whether DistilBERT pretrained with distilled MLM fundamentally understands language, we perform our conditional and individual layer-wise edge probes on the MLM-pretrained DistilBERT and BERT models. For DistilBERT, we probe the embedding layer (our DistilBERT conditional probe baseline) and the LayerNorm outputs of layers 1, 4, 5, and 6; for BERT, we probe the embedding layer (our BERT conditional probe baseline) and the LayerNorm outputs of layers 1, 6, 10, 11, and 12. Since DistilBERT is reported to perform almost as well as its teacher (BERT) on a range of downstream tasks, we expect it to also understand language with similar capability. Our results, as shown in Figure 2, find that DistilBERT's linguistic understanding across its layerwise depth closely matches and even surpasses BERT's across our suite of tests.

Finetuning on mNLI. In our mNLI finetuning experiments, we used three models:

- **Teacher:** pretrained BERT, finetuned without distillation
- **Control:** pretrained DistilBERT, finetuned without distillation
- **Student:** pretrained DistilBERT, finetuned with distillation on the finetuned BERT

We ran these experiments twice, once with a learning rate of 5e-5 and again with a learning rate of 1e-5 and a weight decay of 1e-5 (we re-ran because we realized the first run was not converging to fit mNLI properly; we include it regardless because it provides interesting context results). Our validation



(a) MLM Baseline Individual Probes

(b) MLM Baseline Conditional Probes

Figure 2: MLM Baseline Probes

performances are reported in Table 1. Note that we refer to the initial run as UT ("under-tuned") and the second run as OT ("over-tuned"). We find that in both runs, the student consistently outperforms the control by 1-2%. However, interestingly, we find that the superior teacher performance in the OT case does not correspond to a consistent performance improvement in the student, and that the student also sometimes outperforms the teacher, suggesting that we may have needed to finetune the teacher further.

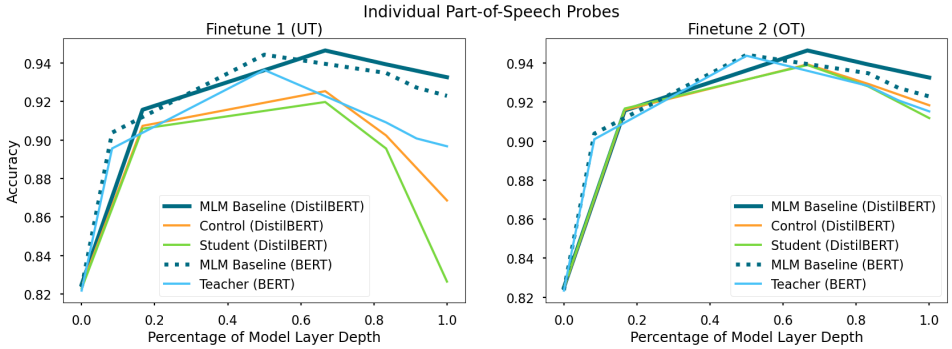
Finetuned Model Edge Probes. To determine the effects of finetuning on probe performance, we perform our suite of conditional and individual layer-wise edge probes across our six finetuned models (three from each finetuning run). Figures 3 show the probe performances relative to the BERT and DistilBERT pretraining-only baselines. We find that in our initial run (UT), across our probe tasks, late-layer performance drops significantly for all our models, with the teacher decreasing least relative to pretrained BERT, the control decreasing more relative to pretrained DistilBERT, and the student surprisingly decreasing even more than the control. In our re-run (OT), we still find that late-layer probe performance generally decreases after finetuning (with the exception of BERT on `ner`), but the differences are much smaller, and the student and control consistently have very similar probe performances. The primary correlation we observe is that decreases in BERT performance after finetuning are similar to the gap between our control and student probe performances. Critically, we **do not** find that distilled finetuning improves probe performance, and in the UT case find that it actually worsened late-layer performance.

Finetuned Model Generalization. For each of our six finetuned models, we run a single epoch of evaluation (with no further training) on SNLI, ANLI, and the Jamaican Patois NLI dataset. As shown in Table 1, we find that with the exception of the Jamaican Patois NLI dataset in the UT case, the student consistently matches or outperforms the control on generalization performance, and the teacher always outperforms both. Furthermore, unlike in the in-distribution case, we find that the universal improvement of the OT teacher consistently produces a superior OT student.

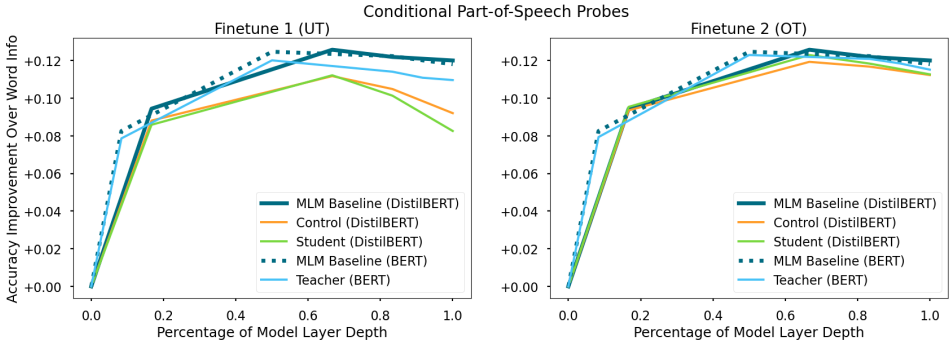
Finetuned Model Function Word Probing. Similar to generalization testing, for each finetuned model, we run a single epoch of evaluation on the five function word probing datasets. We find that the teacher and student always outperform the control, but as in the in-distribution case, the student sometimes outperforms the teacher. However, unlike in previous experiments, we find no clear correlations in the performance of the UT models vs the OT models across our probing datasets. In particular, unlike in the generalization case, we do not find that better teachers consistently produce better students, or that OT teachers or students are consistently better than their UT counterparts.

Model	prep		quant		space		comp		neg	
	UT	OT	UT	OT	UT	OT	UT	OT	UT	OT
Teacher (BERT)	51.7	51.7	78.0	77.7	76.4	73.3	64.0	65.2	63.8	65.0
Control (DistilBERT)	50.8	49.7	74.3	72.1	75.2	72.0	62.9	**59.6**	62.6	**63.1**
Student (DistilBERT)	**51.4**	**50.8**	**76.8**	**73.7**	**77.7**	**73.9**	**67.4**	**59.6**	**64.1**	**63.1**

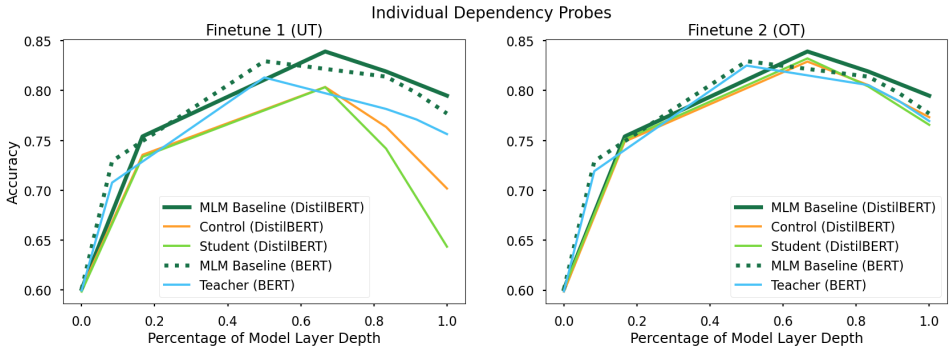
Table 2: Function-Word NLI Tasks (UT: Undertuned, OT: Overtuned)



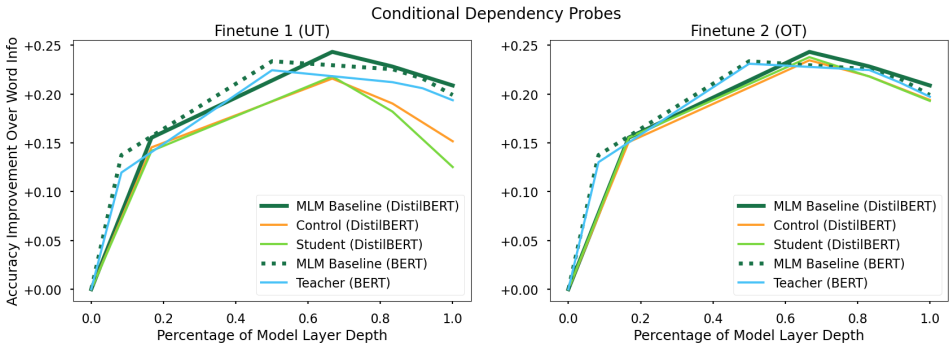
(a) Individual Part-of-Speech Probes



(b) Conditional Part-of-Speech Probes

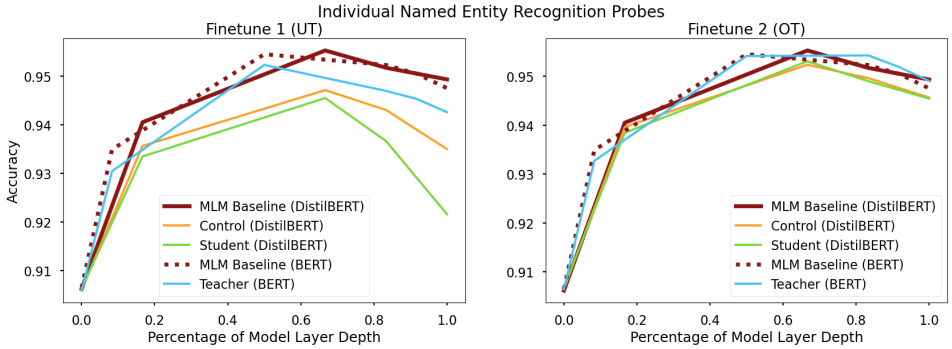


(c) Individual Dependency Relation Probes

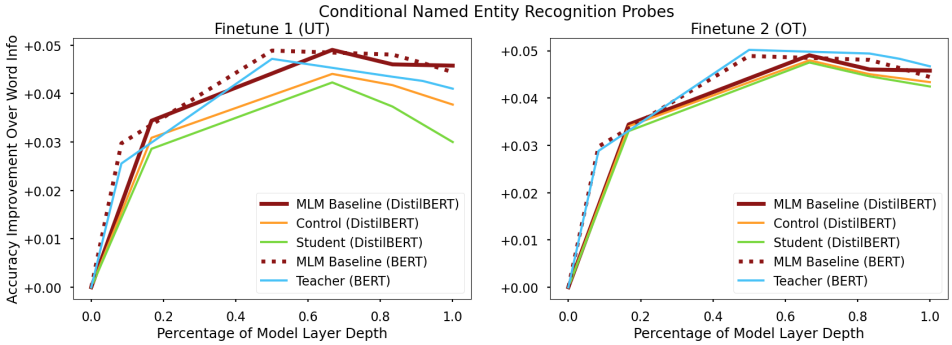


(d) Conditional Dependency Relation Probes

Figure 3: Individual and Conditional Layerwise Probes (part-of-speech and dependency relations)



(a) Individual Named-Entity Recognition Probes



(b) Conditional Named-Entity Recognition Probes

Figure 4: Individual and Conditional Layerwise Probes (named-entity recognition)

5 Analysis

First and foremost, we verify that **performing distilled NLI finetuning** on pretrained DistilBERT (student) by leveraging the output of a BERT model finetuned on mNLI (teacher) **consistently improves performance on in-distribution validation accuracy, and find that these improvements transfer to out-of-distribution generalization accuracy**, relative to a DistilBERT model with the same pretraining but finetuned without distillation (control). To determine where these gains are coming from, we proceed with an analysis of our probe results.

First, since the teacher and student both consistently outperform the control on function-word challenge datasets, we conclude that **the teacher possesses a deeper linguistic understanding of function-word information that it then successfully transfers to the student.** We then posit that this deeper knowledge of function words results in superior generalization to other NLI datasets. However, comparing the UT and OT teacher and student models within individual challenge datasets, we observe that the higher-performing teacher does not necessarily produce a higher-performing student (as shown in the `comp` and `neg` results of Table 2, where the worse-performing teacher produces the better student).

We further note that this improvement in generalization occurs despite the fact that in the UT case, the student is *worse* than the control at understanding word-level sytactic (`pos`, `dep`) and semantic (`ner`) information, as demonstrated through our layerwise edge probes. We hypothesize that this means that **the negative effects of inferior understanding of these word-level properties are outweighed by the improvements distillation finetuning brings to other kinds of linguistic understanding**, such as the aforementioned function-word understanding, and possibly other properties we did not test.

Nevertheless, we still find word-level understanding to be important for generalization performance. When we compare models across the OT and UT finetuning runs, we find that the models with better word-level probe performance also generalize better (i.e. the OT student, which has superior edge probe performance, generalizes better than its UT counterpart, as do the OT control and teacher). This effect does not transfer to the function-word challenge datasets, but we believe this is because the function word substitutions evaluated via the challenge datasets are intentionally designed to be grammatically equivalent, and thus largely do not change any dependencies, parts-of-speech, or named entities in the dataset sentences.

Our most difficult-to-explain result is that, despite the teacher's superior understanding of word-level linguistic properties over the control, distillation does not improve the student's performance (student and control are even in the OT case), and in the UT case, the student actually performs worse in later layers than the control does. We hypothesize that this relates to the fact that when finetunied, the teacher forgets word-level knowledge relative to its MLM-pretrained counterpart, as reflected in lower edge probe BERT performance after finetuning. Thus, **we suspect that the more the teacher forgets word-level knowledge reflected in the edge probes in favor of NLI-specific heuristics, the more the joint loss encourages the student to sacrifice its understanding of those same linguistic properties to instead match the heuristics of its teacher, ultimately resulting in a student that performs even worse than the control.** However, we reiterate that even in the UT case, this forgetting was outweighed by the useful NLI-related linguistic properties picked up by the teacher and transferred to the student during distillation, such as function-word understanding.

6 Conclusion and Future Work

In this paper, we have demonstrated that during NLI distillation finetuning, student DistilBERT models do not just learn linguistically-unfounded heuristics to mimic a BERT teacher, and instead find that student models actually absorb an understanding of linguistic properties from their teacher. We have observed that this can occur in both positive and negative ways: a teacher that forgets more word-level syntax and semantics (part-of-speech, word-level dependencies, and named entity recognition) produces a student that similarly forgets more word-level knowledge than a control an independently-finetuned model with the same pretraining as the student), while a teacher that outperforms the control on function-word understanding can transfer that knowledge to a student that then also outperforms the control. We have seen that these improvements in function-word understanding lead to a student that generalizes better to other NLI datasets, even in the presence of worse word-level understanding, suggesting that function-word understanding may be more important than word-level knowledge for good NLI performance.

However, due to our limited number of probes, we have not eliminated the possibility that distillation transfers other un-probed linguistic properties that substantially contribute to producing our better student models, and believe that further research should probe additional properties between teachers, students, and controls. We also believe that a separate investigation on how distillation during the MLM pretraining process affects the student's understanding of linguistic properties would provide a fuller picture of the relationship between distillation and the transfer of meaningful linguistic knowledge between teacher and student models. Finally, our model is limited in terms of finetuning scope (we only explore NLI) in teacher and student model architectures, and we believe further investigation could be done for other architecture pairs or on other downstream tasks.

References

Ruth-Ann Armstrong, John Hewitt, and Christopher Manning. 2022. Jampatoisnli: A jamaican patois natural language inference dataset.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and advances.

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.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

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.

Nadir Durrani, Hassan Sajjad, and Fahim Dalvi. 2021. How transfer learning impacts linguistic knowledge in deep nlp models?

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher Manning. 2021. Conditional probing: measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1639, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Jakub Hościłowicz, Marcin Sowański, Piotr Czubowski, and Artur Janicki. 2023. Can we use probing to better understand fine-tuning and knowledge distillation of the bert nlu?

Najoung Kim, Roma Patel, Adam Poliak, Alex Wang, Patrick Xia, R. Thomas McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different nlp tasks teach machines about function word comprehension.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to bert embeddings during fine-tuning?

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

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

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

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

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.