Are Distilled Models Just Deep-Sea Octopi?

Probing Linguistic Representations of Distilled Models

Christos Polzak, Joy Yun
{clcp, joyyun} @stanford.edu
Department of Computer Science

Overview

- While large language models are growing more powerful, small models are still necessary for general low-resource deployment.
- Knowledge distillation creates smaller high-performance models by using the predictive outputs of a frozen, pre-trained large model to "teach" a smaller "student" model to match its "teacher's" outputs.
- **To what extent are distilled models really learning deep language rules versus simply picking up on heuristics used to mimic their teachers' outputs?**
- Conclusion: Student DistilBERT models generalize better, and the student models absorb understanding of linguistic properties from their teacher, both positive and negative. Distilled understanding of function words contributes to improved generalization.

Datasets

Fine-tuning Dataset:

- Benchmark multi-Natural Language Inference (mNLI) dataset
  Train = 392,002 samples, matched (`mnli_m`) validation = 9800 samples, mismatched (`mnli_mm`) validation = 9800 samples

Layerwise Edge Conditional Probing [1]:

- Part-of-Speech (`pos`), Dependency Relations (`dep`) - Universal Dependencies English Web Treebank
- Named Entity Recognition (`ner`) - OntoNotes V5 dataset (english_v4)

NLI Function-Word (FW) Probing [2]:

- Evaluation on five small curated challenge datasets
  FW types: negation (`neg`), spatial (`spatial`), comparators (`comp`), quantitatives (`quant`), prepositions (`prep`)

Generalization Datasets:

- Stanford Natural Language Inference (`snli`)
- Adversarial Natural Language Inference (`anli`)
- Jamaican Patois Natural Language Inference (`jam_patois`)

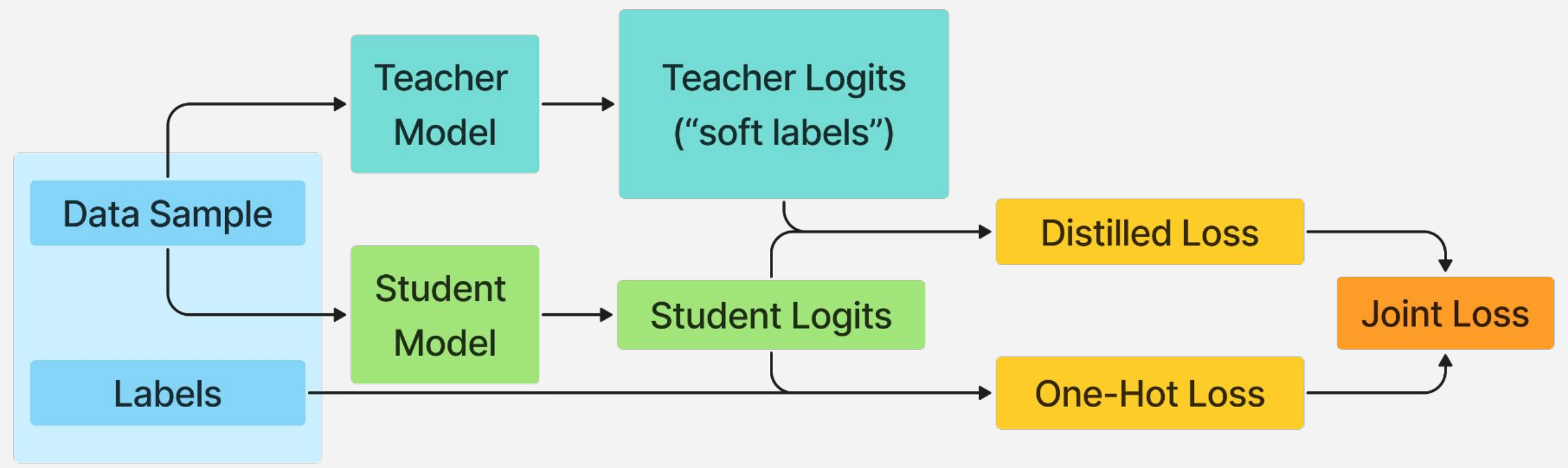
References

[1] 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.
[2] 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.
[3] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Methods

1. (Distillation) Finetuning with DistilBERT [3]

First, we finetune MLM-pretrained BERT on mNLI: the **teacher.**
Then, we finetune MLM-pretrained DistilBERT with distillation: the **student.**
Finally, we finetune MLM-pretrained DistilBERT independently: the **control.**



$\mathcal{L}_{\text{hard}} =$ standard one-hot cross-entropy loss | $\mathcal{L}_{\text{dist}} = \sum_i t_i * \log s_i$ | $\mathcal{L}_{\text{joint}} = \alpha_{\text{hard}}\mathcal{L}_{\text{hard}} + \alpha_{\text{dist}}\mathcal{L}_{\text{dist}}$

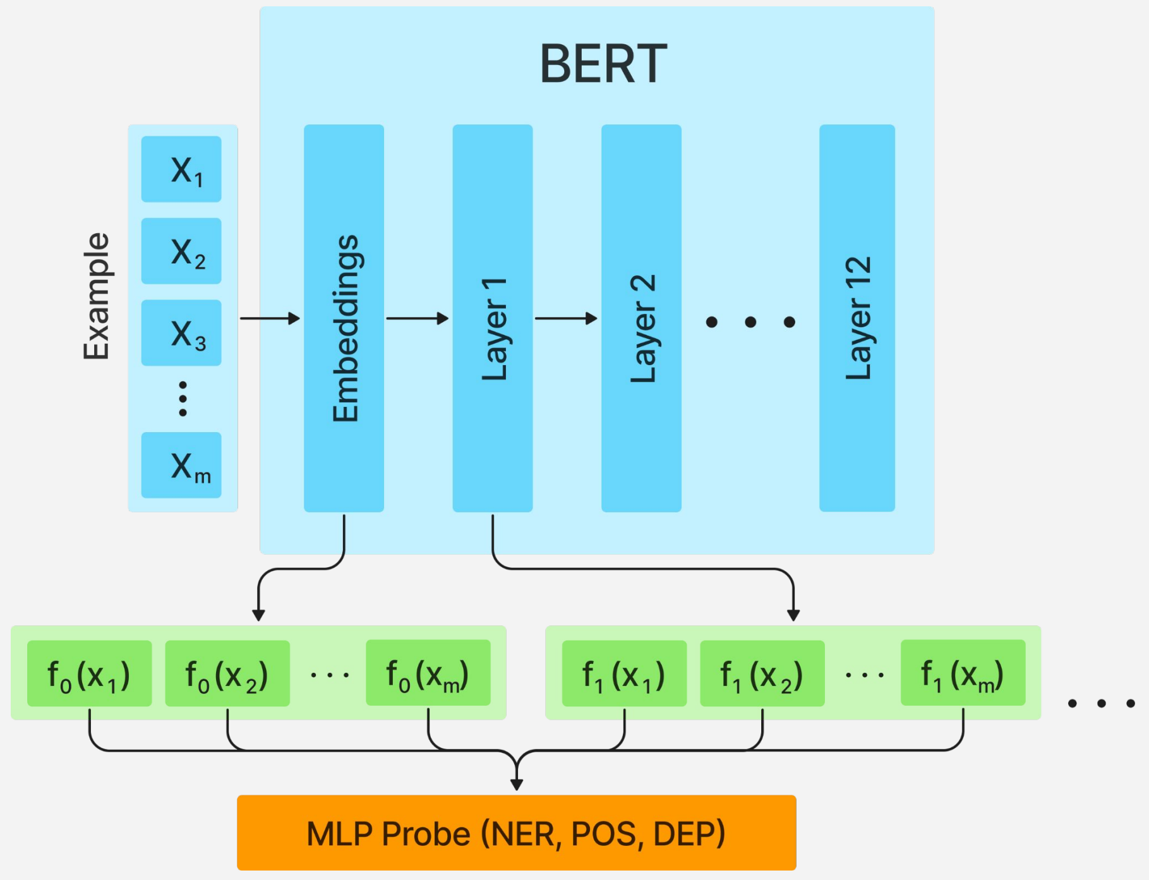
t, s = softmax probabilities

2. Layerwise Probing

Constructing probe datasets: Pass sentences with word-level `pos`/`dep`/`ner` labels through finetuned models. Store representations across layers.

Probing: Train a two-layer MLP to predict the label from the layer's representations + the embedding representation (baselining out word-identity-level information).

Interpretation: Probe accuracy ↑ = ↑ Understanding of label at that layer



3. Generalization and Function Word Probing

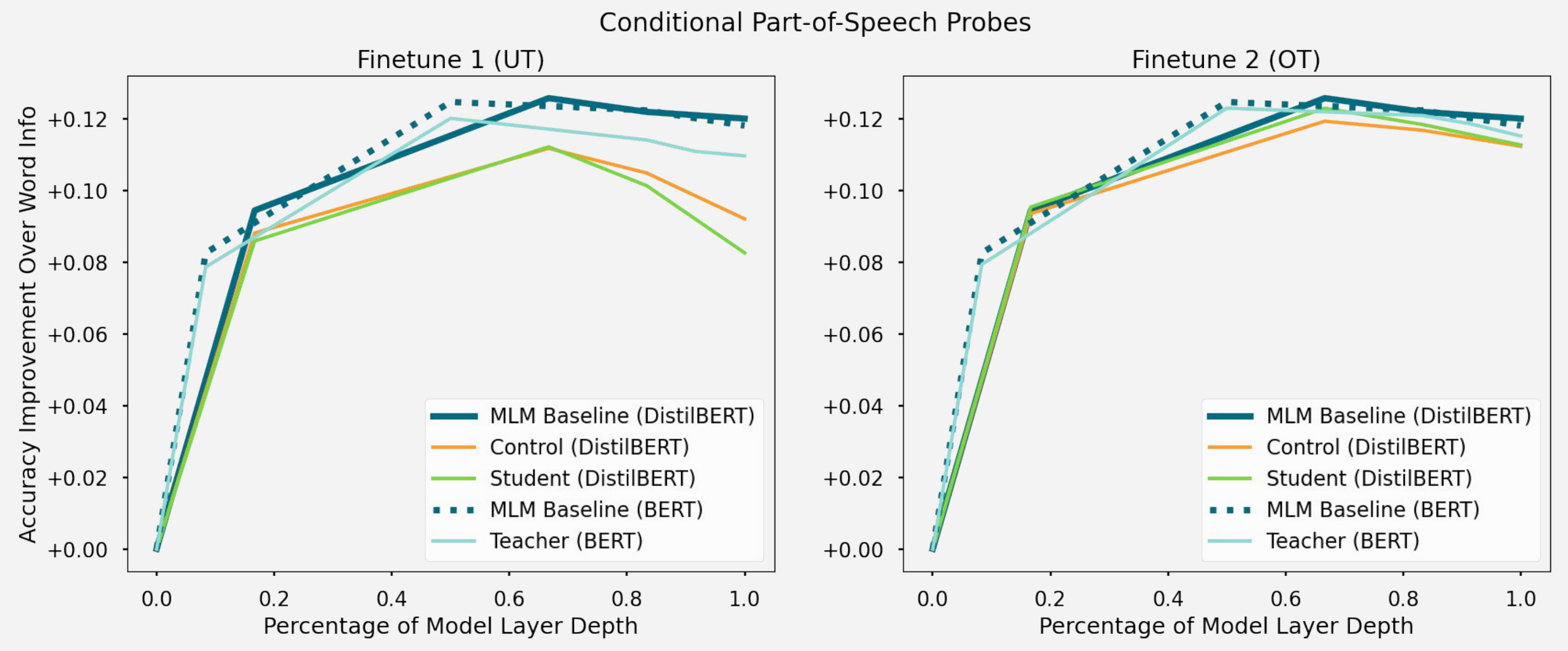
To obtain results, evaluate NLI models on entire datasets (all splits used)

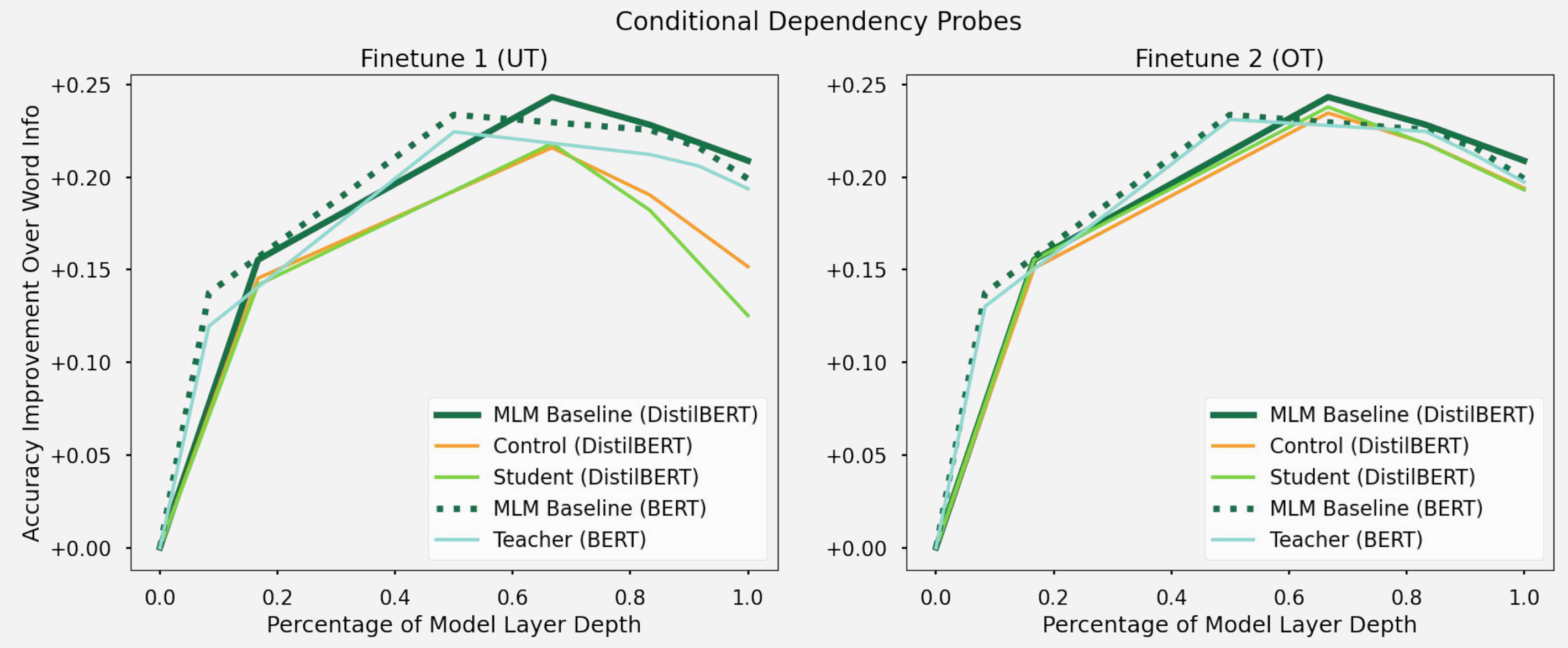
Results

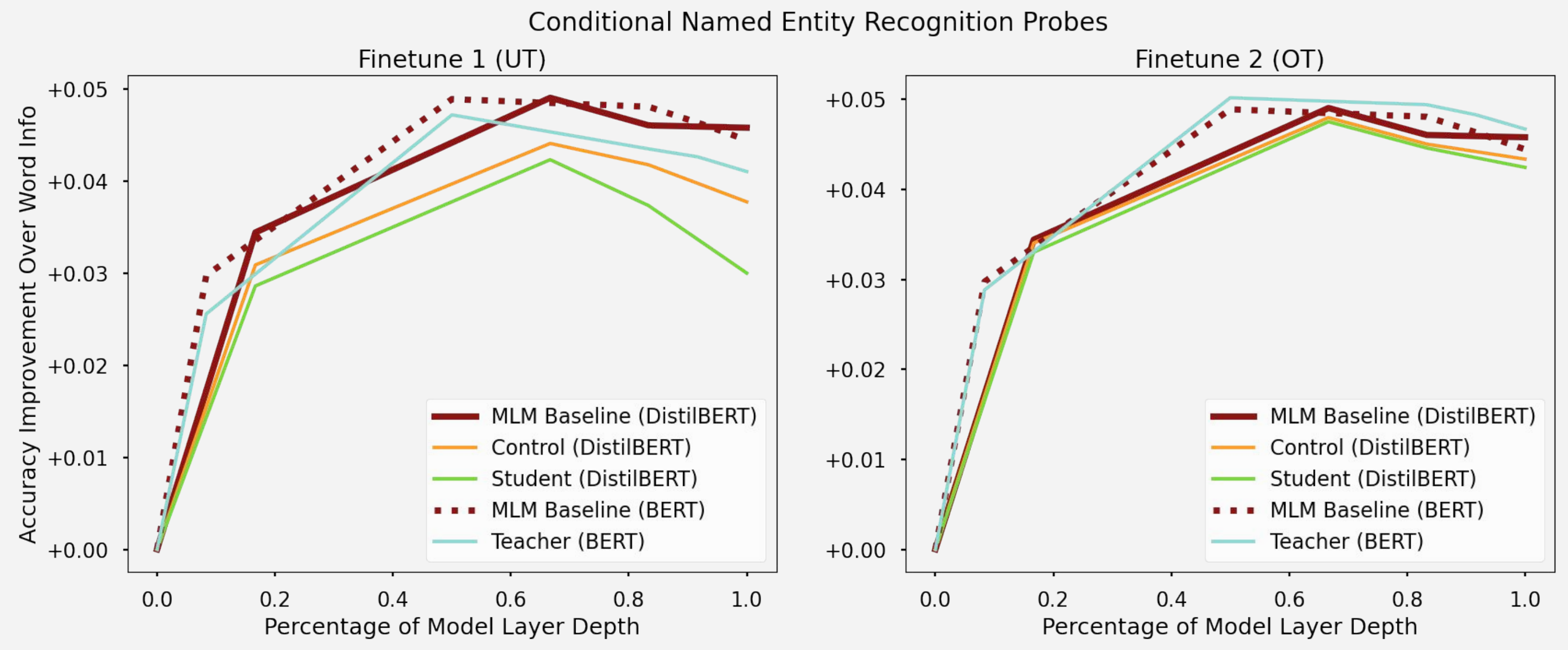
NLI	mnli_m		mnli_mm		snli		anli		jam_patois	
	UT	OT	UT	OT	UT	OT	UT	OT	UT	OT
Teacher	82.5	84.0	83.4	83.9	76.6	78.0	64.8	66.3	51.2	53.7
Control	81.2	81.2	81.3	81.6	73.6	73.9	61.9	62.4	**49.9**	46.9
Student	**82.7**	**83.3**	**83.4**	**83.0**	**74.5**	**75.6**	**64.0**	**64.8**	46.8	**50.6**

FW	prep		quant		spatial		comp		neg	
	UT	OT	UT	OT	UT	OT	UT	OT	UT	OT
Teacher	51.7	51.7	78.0	77.7	76.4	73.2	64.0	65.2	63.8	65.0
Control	50.8	49.7	74.3	72.1	75.2	72.0	62.9	**59.6**	62.6	**63.1**
Student	**51.4**	**50.8**	**76.8**	**73.7**	**77.7**	**73.9**	**67.4**	**59.6**	**64.1**	**63.1**

Results (cont'd)







Discussion

- Performing **distilled NLI finetuning consistently improves performance** on in-distribution validation and out-of-distribution generalization accuracy.
- The teacher model possesses a deeper linguistic understanding of **function-word information** that **is successfully transfered** to the student during distillation.
- Comparing OT and UT finetuning runs, **better word-level understanding correlates with better generalization performance**.
- In the UT case: Negative effects of inferior understanding of word-level properties are outweighed by the improvement distillation brings to other kinds of linguistic understanding (i.e. function words).