On the Interplay Between Fine-tuning and Composition in Transformers

Lang Yu[1], Allyson Ettinger[1]
[1]University of Chicago

Contact: langyu@uchicago.edu, twitter: @langyu94

Paper: https://arxiv.org/pdf/2105.14668.pdf

Code: https://github.com/yulang/fine-tuning-and-composition-in-transformers

THE UNIVERSITY OF CHICAGO

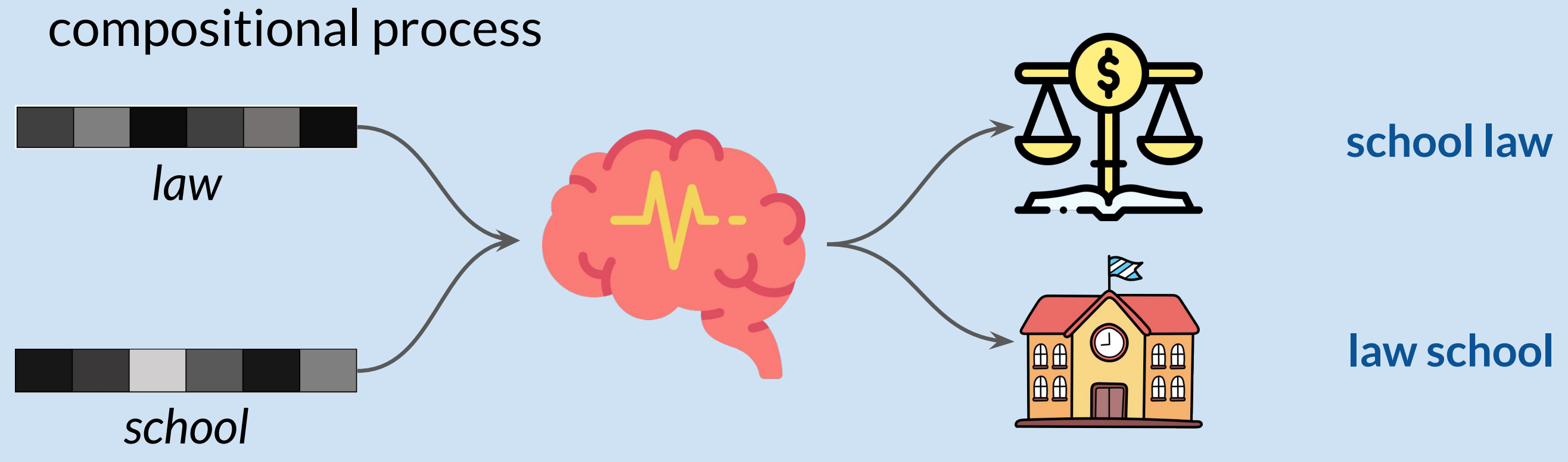
INTRODUCTION

- Phrase-level representations in transformers reflect heavy influences of **lexical content**, and lack evidence of sophisticated **compositional information** (Yu and Ettinger, 2020)
- Will models show better compositionality after fine-tuning on tasks that are good candidates for requiring composition?
- We experiment with 2 fine-tuning tasks, and present layer-wise analysis of 5 different transformers. We present further analysis of the impact of fine-tuning.

COMPOSITION & EVALUATION

Composition

- A fundamental component of language understanding
- Capacity to combine meaning units into larger units
- Composed representation should resemble output of human compositional process



Composition Evaluation (Yu and Ettinger, 2020)

- Capture correspondence of phrase representation with **human judgment** on phrase pair similarity
- Evaluation consists of two types of tasks
  - **Similarity correlation**: correlate representation cosines with human-annotated similarity ratings from BiRD (Asaadi et al 2019)
  - **Paraphrase classification**: train a MLP classifier to identify paraphrases versus non-paraphrases from PPDB (Pavlick et al., 2015)
- Each test has uncontrolled and controlled variations : latter constitutes model-agnostic schemes to remove cues of word overlap

Composition in Pre-trained Transformers (Yu and Ettinger, 2020)

- Models show non-trivial alignment with human judgment, but it seems to rely on **lexical information**
- With lexical overlap controlled, models show **severe performance drop**
- Suggests lack of sophisticated composition beyond word content encoding

FINE-TUNING

Can we improve compositionality via fine-tuning?

Fine-tune on promising tasks for requiring composition

- Paraphrase Adversaries from Word Scrambling (PAWS) (Zhang et al., 2019)
  - Quora Question Pairs subset (PAWS-QQP)
  - Binary classification on sentence pairs with **high lexical overlap**
- Stanford Sentiment Treebank (Socher et al., 2013)
  - 5-class classification on syntactic phrases of fine-grained sentiment labels
  - **Hierarchical structures** promote composition

IMPACT OF FINE-TUNING

● Pre-trained ● PAWS-tuned ● SST-tuned

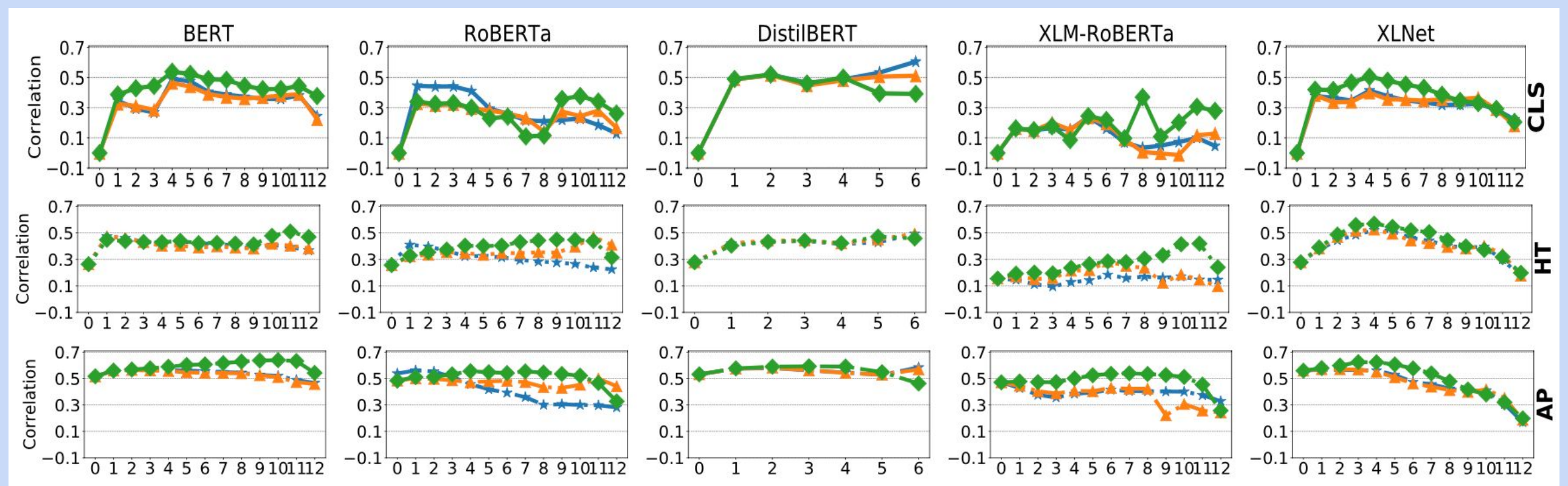


Figure 1: Correlation on uncontrolled BiRD dataset (phrase-only input)

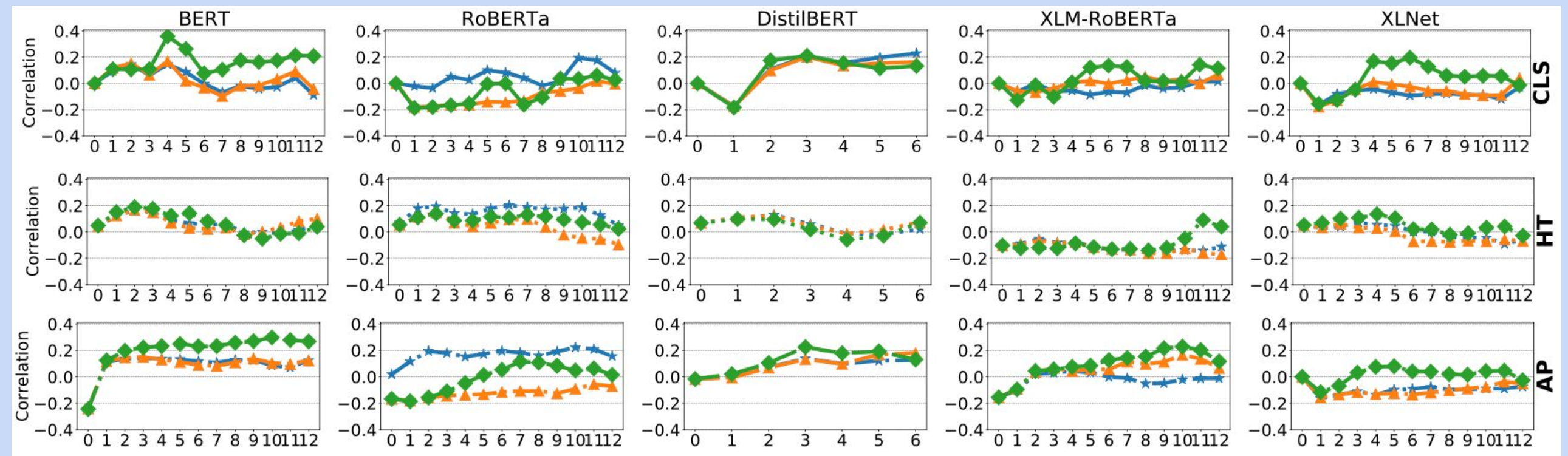


Figure 2: Correlation on controlled BiRD dataset (phrase-only input)

Note: "HT" = Head-token, "AP" = Avg-Phrase

Key findings:

- Fine-tuning consistently improves peak correlations among models in uncontrolled tests. Improvements are generally stronger tuning on SST than on PAWS
- In controlled tests, PAWS-QQP mostly harms performance, while SST shows localized benefits in **BERT's CLS token**

FAILURE OF PAWS-QQP

There are also specific discussion, public profile debates and project discussions.

There are also public discussion, profile specific discussions, and project discussions.

Swapping distance = 4

Table 1: Accuracy of fine-tuned models on PAWS-QQP test set.

Model	Accuracy (%)
BERT	80.13
RoBERTa	90.81
DistilBERT	81.98
XLM-RoBERTa	91.18
XLNet	88.24
Linear CLF	71.34

*A simple linear classifier with relative swapping distance as the **only** input feature*

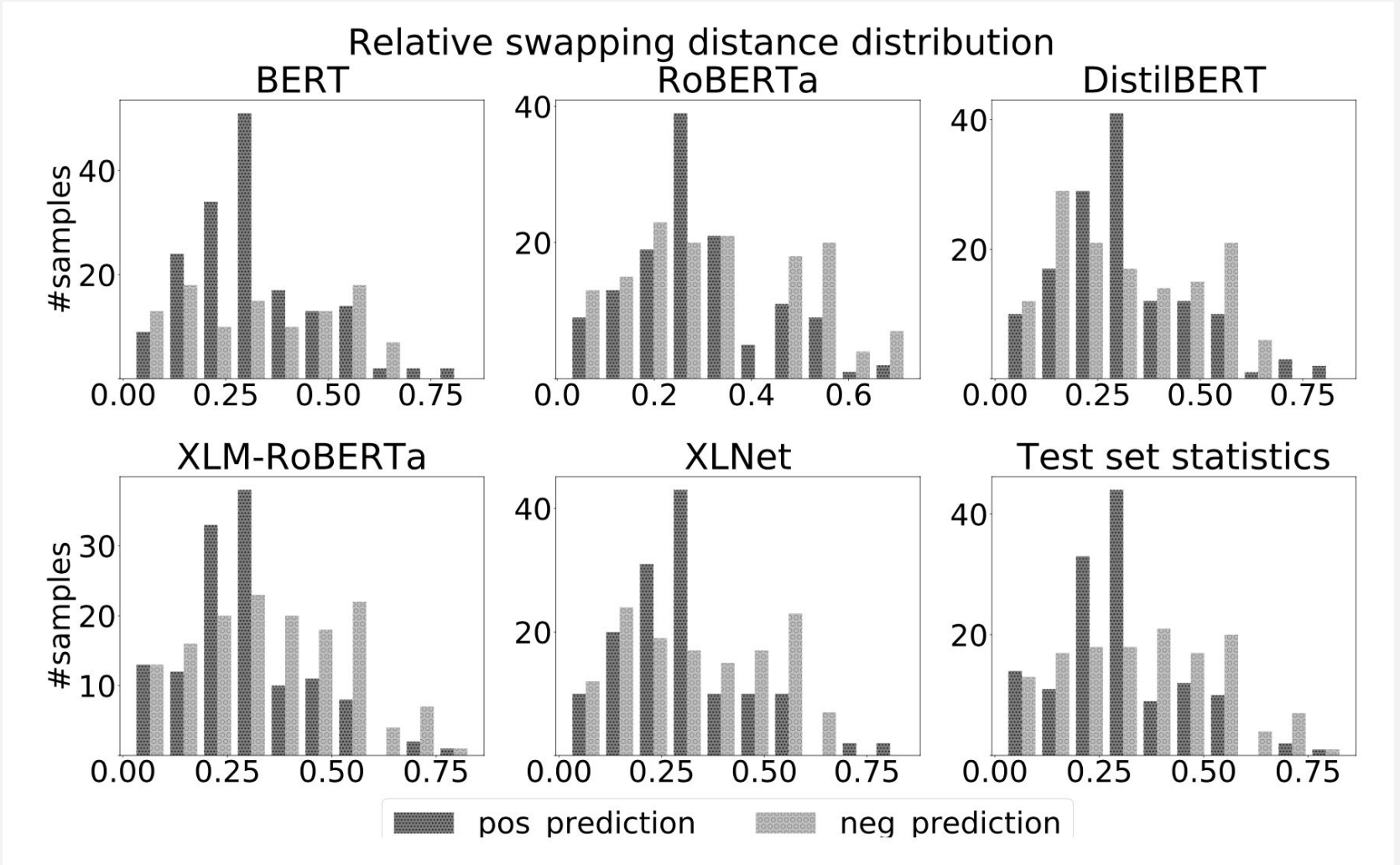


Figure 3: Distribution of positive and negative predictions/labels

LOCALIZED IMPACTS OF SST

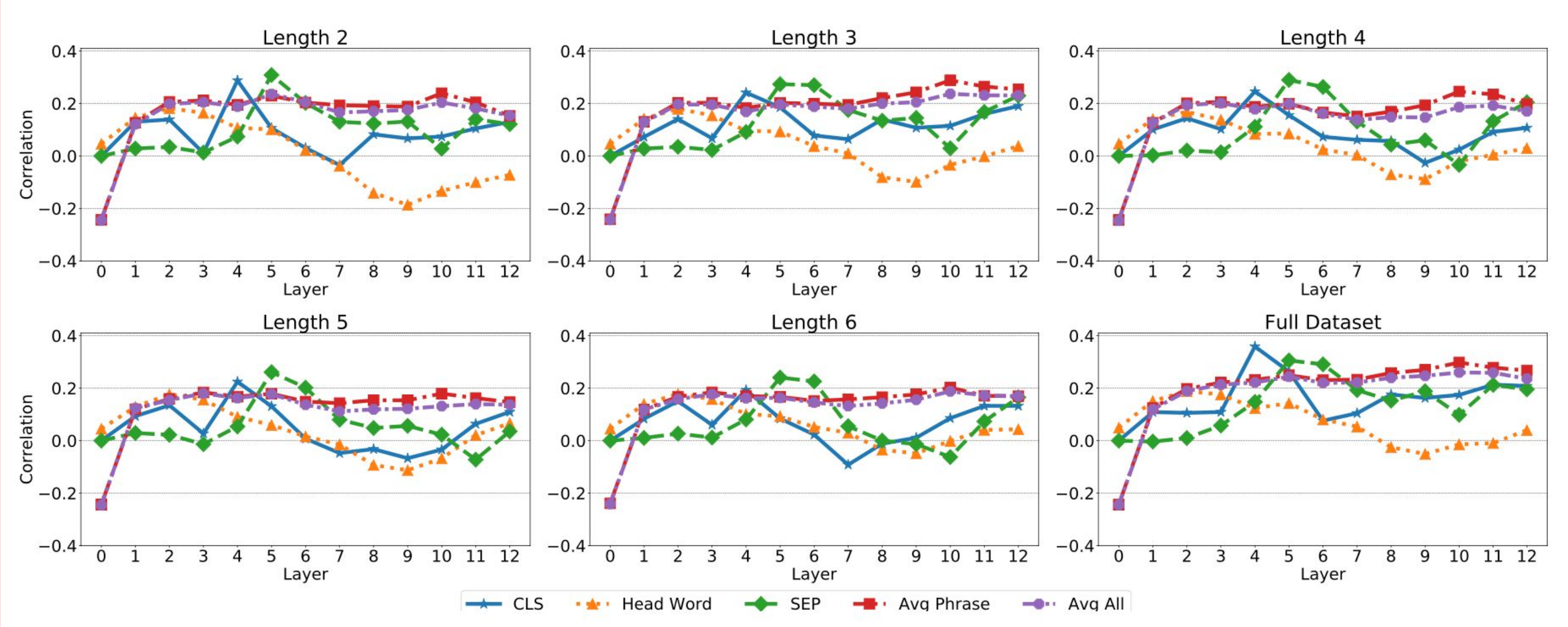


Figure 4: Layer-wise correlation of BERT fine-tuned on phrases of different lengths in SST

- Tuning on full dataset (mixed phrase lengths) gives the strongest boost
- Among filtered sets, length 2 training yields the highest peak, while length 6 the lowest.
- Training on diverse phrase sizes encourages fine-grained attention to compositionality, while training on phrases of similar size to test tasks may have slightly more direct benefit.

TAKEAWAYS

- Select tasks with promise to address composition weakness and reliance on word overlap
- Fine-tuned models show limited improvement
  - **PAWS-QQP** has spurious cues that undermined learning of meaning
  - **SST** shows small localized benefit, but the improvements do not extend to all model
- We predict that phrase-level training with meaning-rich annotations is a promising direction for learning composition

SELECTED REFERENCES

Lang Yu and Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4896–4907.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (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)* (pp. 4171-4186).

Yuan Zhang, Jason Baldridge, and Luheng He. 2019a. Paws: Paraphrase adversaries from word scrambling. 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 1298– 1308.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep mod- els for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1631–1642.