Analyzing the Domain Robustness of Pretrained Language Models - Layer by Layer

Abhinav Ramesh Kashyap, Laiba Mehnaz, Bhavitvya Malik, Abdul Waheed, Devamanyu Hazarika, Min-Yen Kan, Rajiv Ratn Shah

National University of Singapore, IIIT-Delhi, Maharaja Agrasen Institute of Technology

Introduction

- Pretrained Language Models are robust on OOD ***end-tasks***
- But we do not understand, the robustness of representations at different layers
- ***Invariance:*** Inherently, how domain invariant are representations at different layers of pretrained language models?
- ***Probing:*** Do they contain similar linguistic information for data from different domains?

Methodology

- Obtain representations from pretrained transformers for a pair of domains.
- Use Divergence Measures like ***Maximum Mean Discrepancy (MMD)*** , ***Correlational Alignment (CORAL)*** and ***Central Moment Discrepancy (CMD)*** to measure divergence between domains.
- ***Lower the divergence, greater the invariance***
- ***Probing:*** Train classifier probes on source domain and test on target domain (Zero shot probes)

Datasets

- ***Invariance:*** Toronto Book Corpus **(Standard),** PubMed articles **(Biomedical),** 2011 tweets **(Twitter)**
- We use 1000 samples from each and report results as the average of 5 runs for all exp.
- ***Probing:*** Onto-notes POS, NER (Standard), Workshop on Noisy User Generated Text-***WNUT for NER***(Twitter), Twitter POS tagging dataset (***Derczynski et al***)

1. Lower layers of Transformers are less domain variant than higher layers.

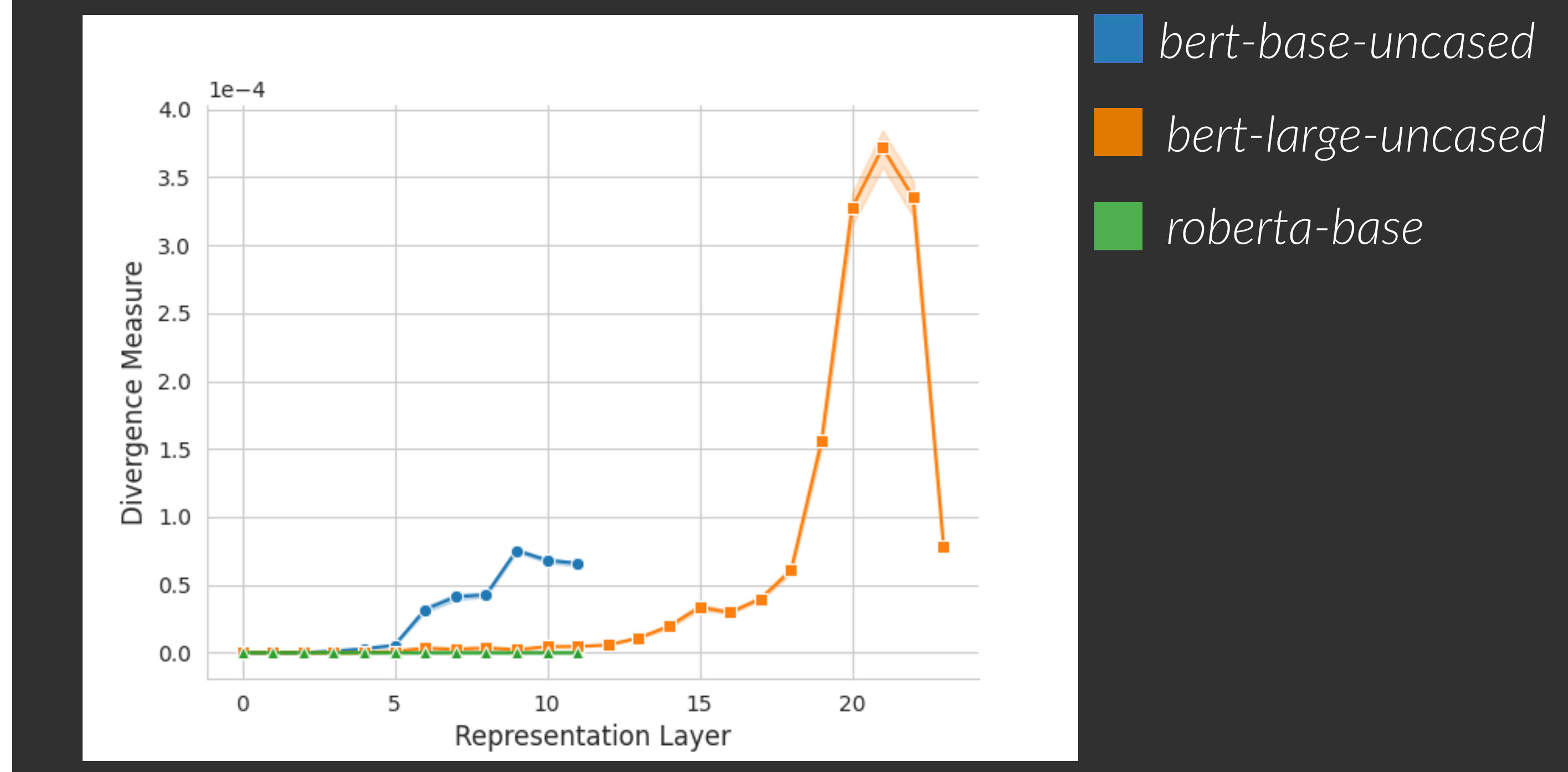
2. Domain Adaptive Transformers (DAPT) are more domain variant at certain higher layers compared to non-DAPT models

3. Distilbert is more domain variant than non-distilled models.

4. Zero shot classifier probes show that Transformers have similar amount of linguistic information at different layers for different domains.

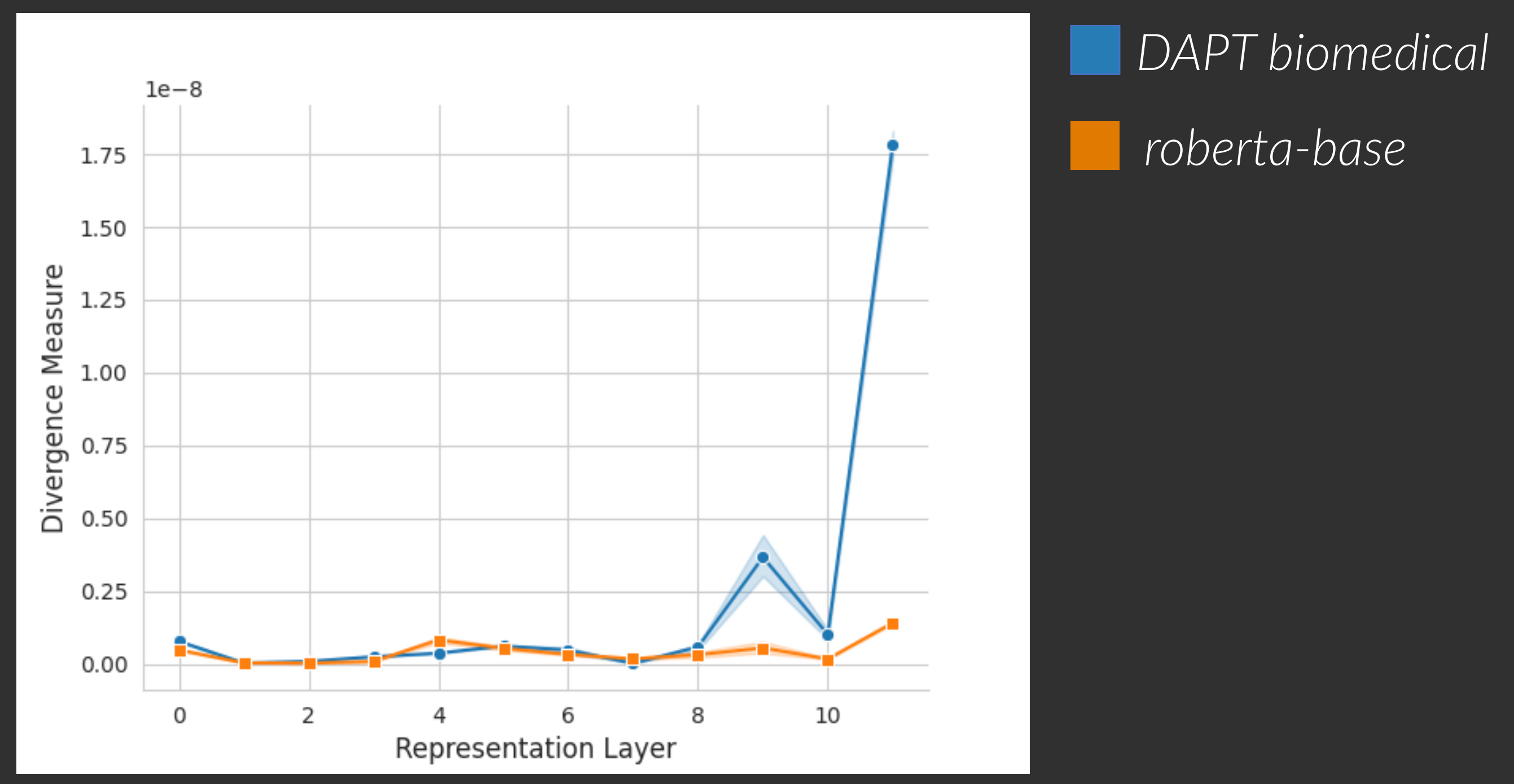
Main Results

Lower Layers are more domain invariant



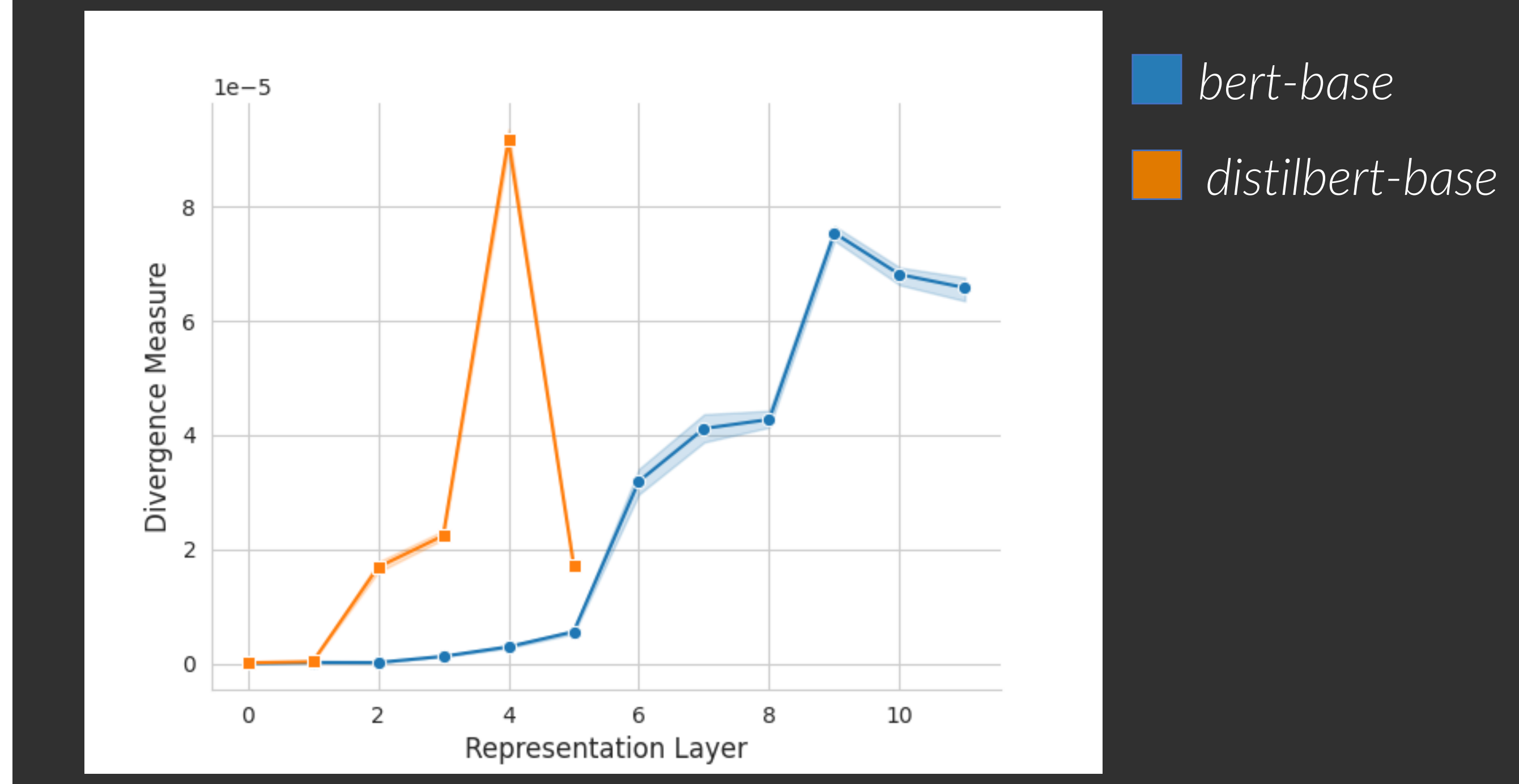
CORAL - standard v biomedical

DAPT models domain variant (higher layers)



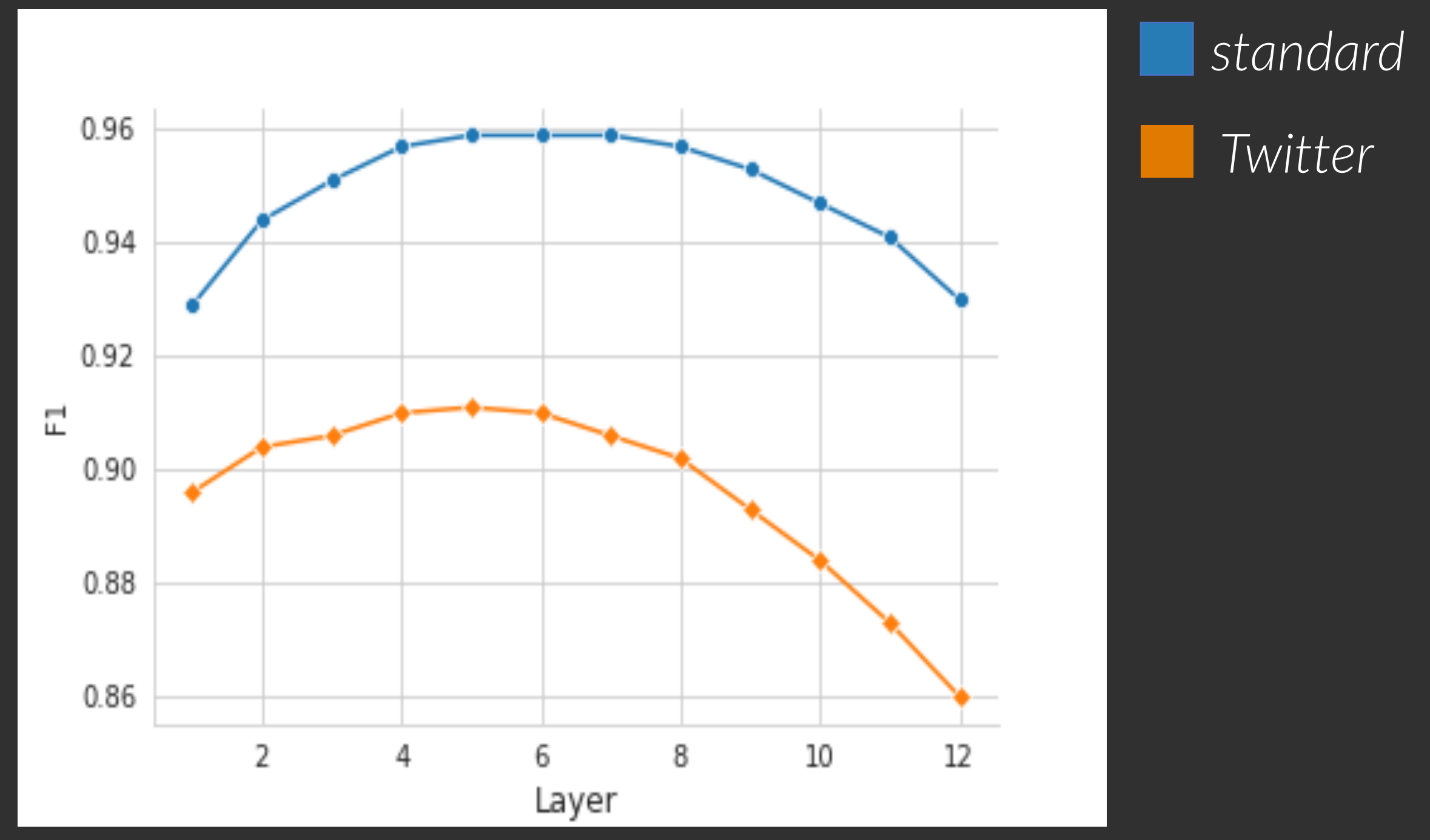
CORAL-roberta-base vs DAPT biomedical

Distilled Models are more domain variant



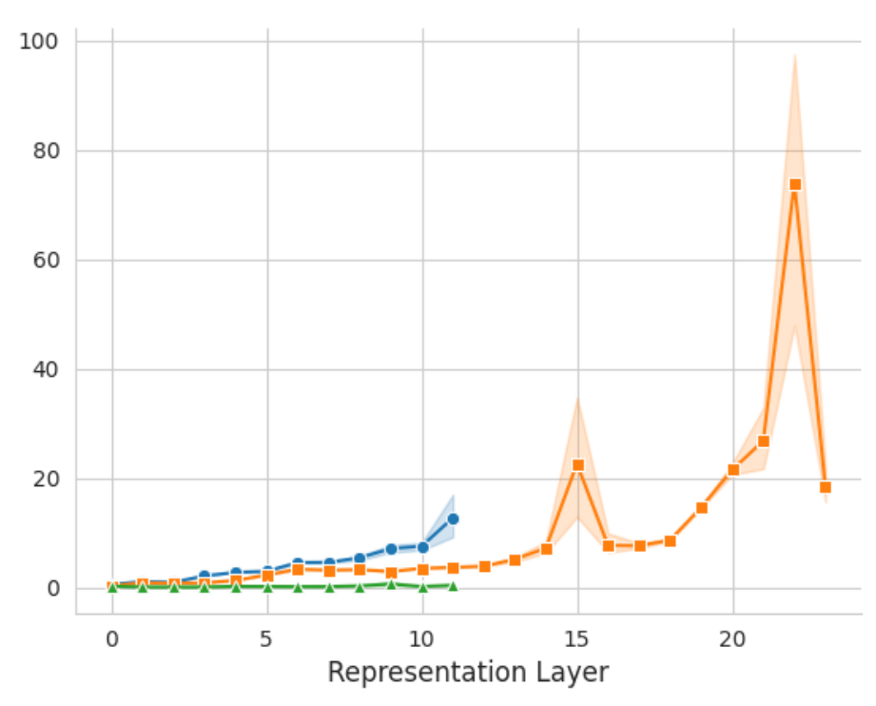
CORAL - distilbert v bert-base

F1 scores for POS peak at layer 5 for both domains

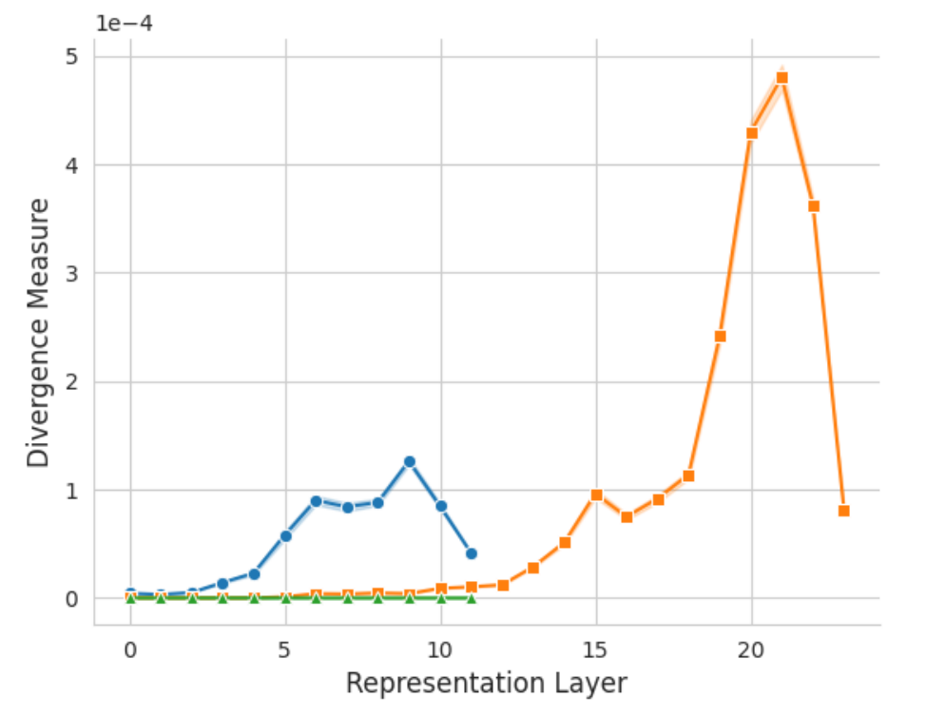


POS - F1 for bert-base standard vs twitter

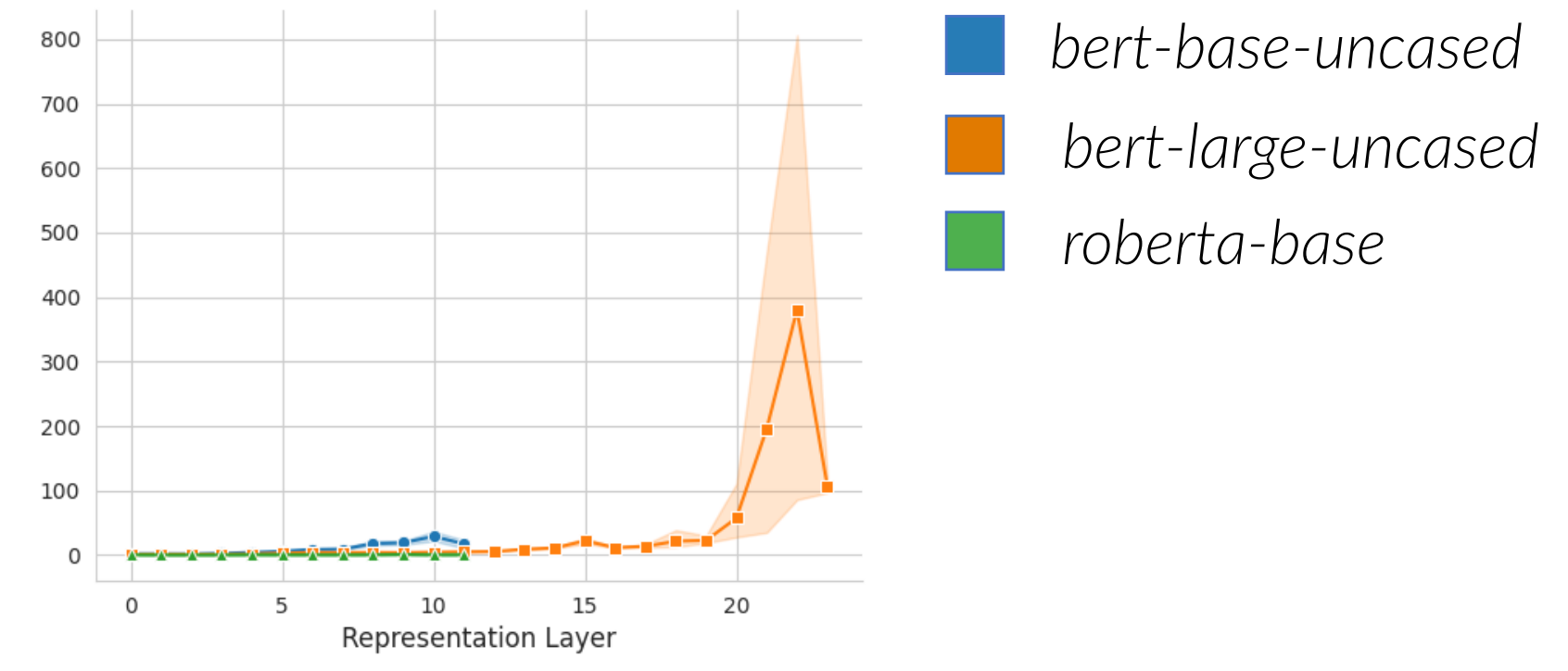
Other Domain Invariance Plots



CMD - standard v biomedical

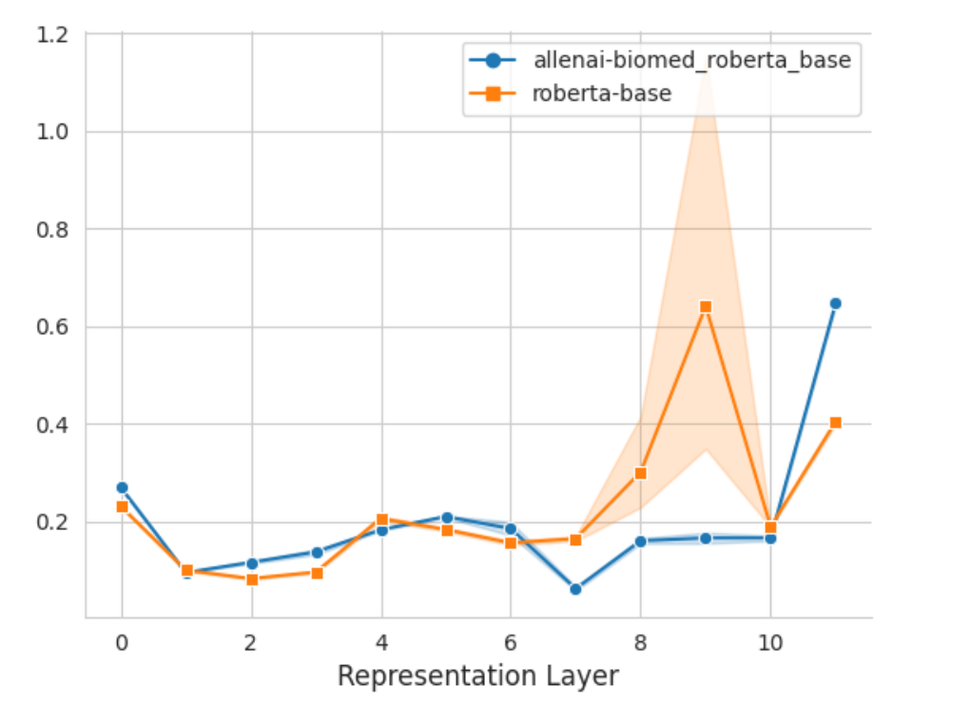


CORAL - standard vs twitter

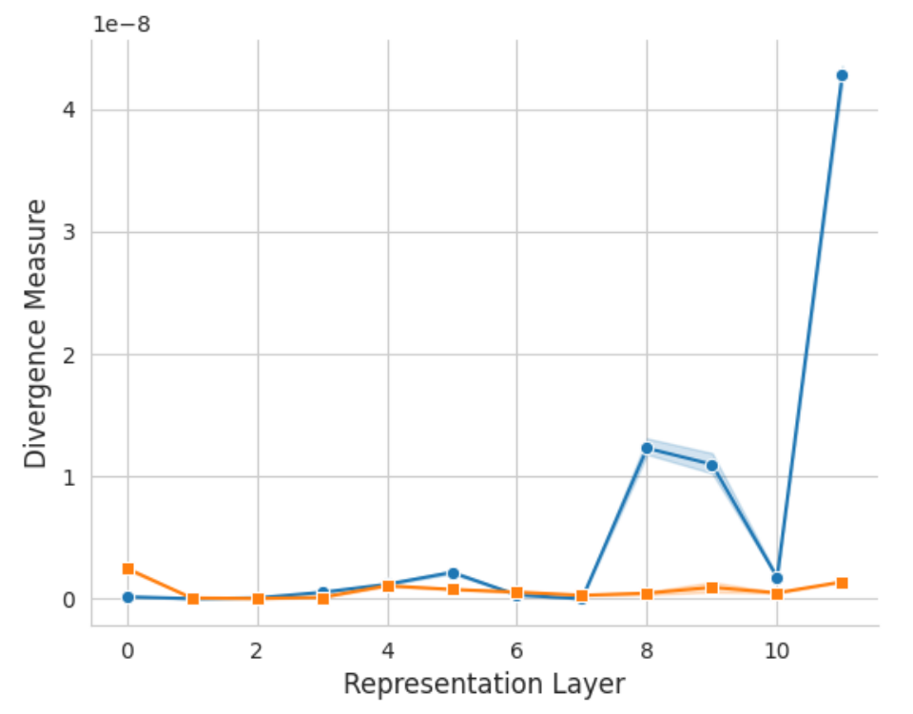


CMD - standard vs twitter

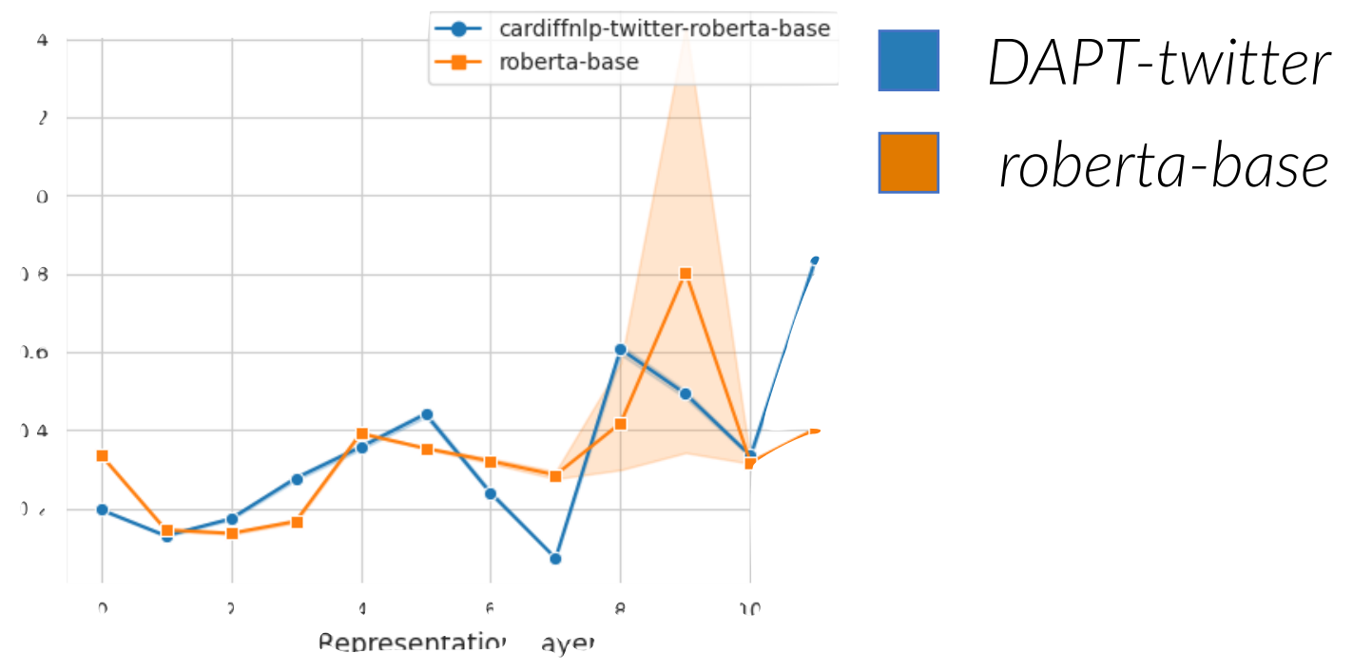
DAPT Plots



CMD - roberta-base vs DAPT biomed

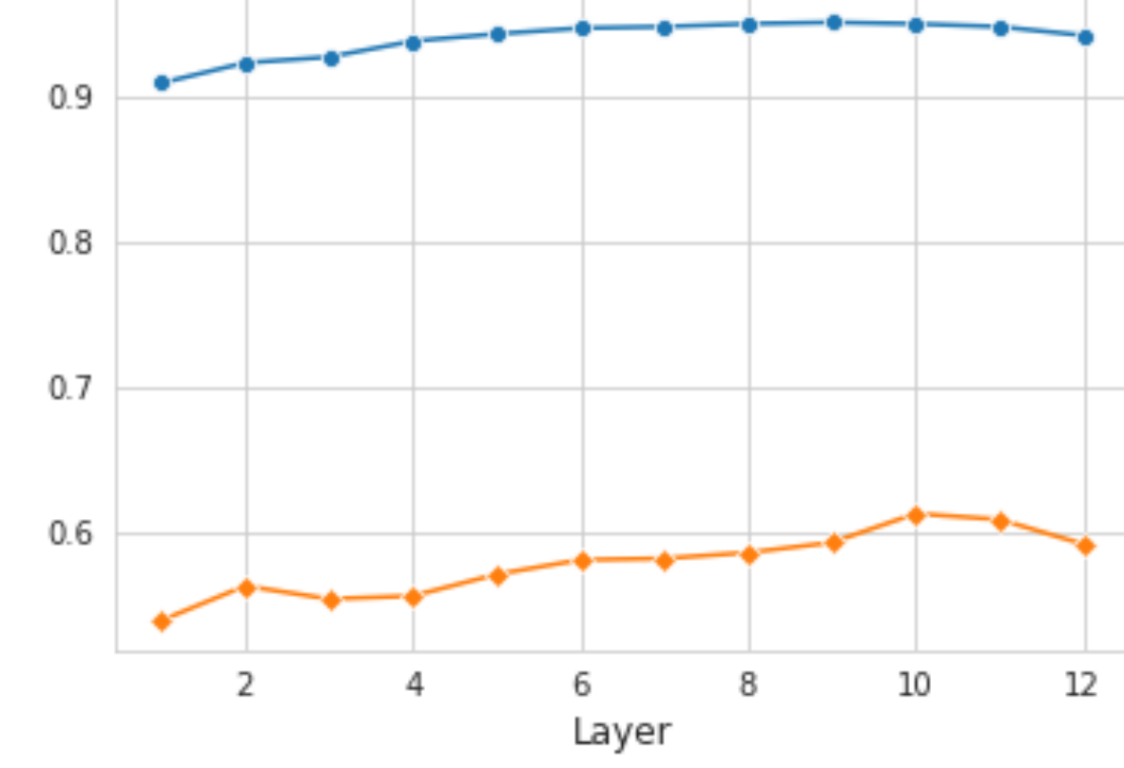


CORAL-roberta-base vs DAPT twitter

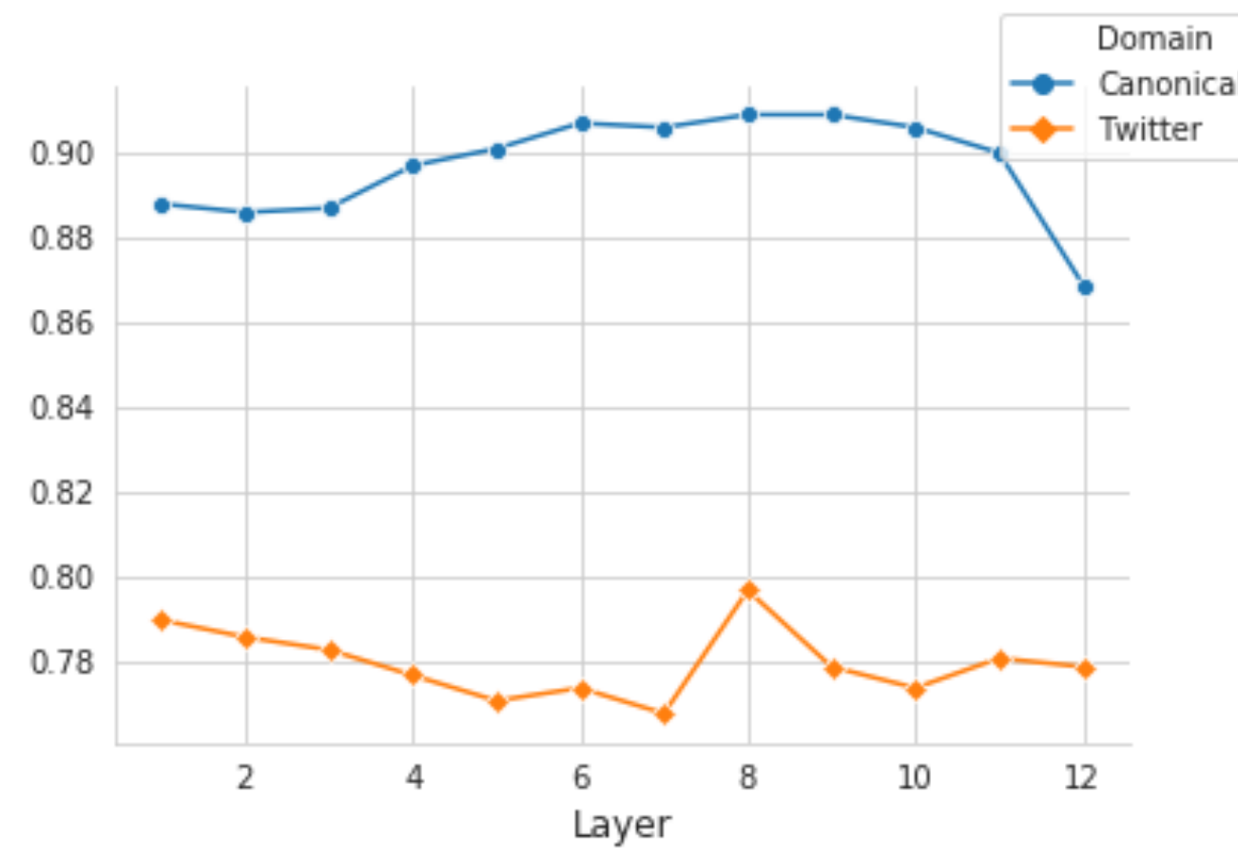


CMD roberta-base vs DAPT twitter

Zero-Shot Classifier Probes Plots



NER - F1 for bert-base standard vs twitter



Coref - F1 for bert-base standard vs twitter

NUS National University of Singapore

WING Web Information Retrieval Natural Language Processing Group

MIDAS@IIITD Multimodal Digital Media Analysis Lab