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

Abhinav Ramesh Kashyap[α], **Laiba Mehnaz**[β], **Bhavitvya Malik**[γ], **Abdul Waheed** [δ], **Devamanyu Hazarika**[α], **Min-Yen Kan**[α], **Rajiv Ratn Shah**[β]

[α] National University of Singapore, Singapore

[β] MIDAS Lab, IIIT-Delhi

[γ] Independent Researcher

[δ] Maharaja Agrasen Institute of Technology, New Delhi, India.

{abhinav, hazarika, kanmy}@comp.nus.edu.sg

{laibamehnaz, bhavitvya.malik, abdulwaheed1513}@gmail.com

{rajivratn}@iiitd.ac.in

Abstract

The robustness of pretrained language models (PLMs) is generally measured using performance drops on two or more domains. However, we do not yet understand the inherent robustness achieved by contributions from different layers of a PLM. We systematically analyze the robustness of these representations layer by layer from two perspectives. First, we measure the robustness of representations by using domain divergence between two domains. We find that *i)* Domain variance increases from the lower to the upper layers for vanilla PLMs; *ii)* Models continuously pretrained on domain-specific data (DAPT) (Gururangan et al., 2020) exhibit more variance than their pretrained PLM counterparts; and that *iii)* Distilled models (e.g.,DistilBERT) also show greater domain variance. Second, we investigate the robustness of representations by analyzing the encoded syntactic and semantic information using diagnostic probes. We find that similar layers have similar amounts of linguistic information for data from an unseen domain.

1 Introduction

Pretrained Language Models (PLMs) have improved the downstream performance of many natural language understanding tasks on *standard* data (Devlin et al., 2019).[1] Recent works attest to the surprising out-of-the-box robustness of PLMs on out-of-distribution tasks (Hendrycks et al., 2020; Brown et al., 2020; Miller et al., 2020). These works measure robustness in terms of the performance invariance of PLMs on end tasks like Natural Language Inference (Bowman et al., 2015; Williams et al., 2018), Sentiment Analysis (Maas et al., 2011), Question Answering (Zhang et al., 2018), among others. However, they do not investigate the domain invariance of PLM representations from different layers when presented with data from distinct domains. Studying the invariance of PLM representations has been useful in advancing methods for unsupervised domain adaptation. For example, in building domain adaptation models that explicitly reduce the divergence between layers of a neural network (Long et al., 2015; Shen et al., 2018a), for data selection (Aharoni and Goldberg, 2020; Ma et al., 2019) *et cetera*.

Given the importance of PLMs, a glass-box study of the internal robustness of PLM representations is overdue. We thus study these representations, dissecting them layer by layer, to uncover their internal contributions in domain adaptation. Firstly, we use the tools of domain divergence and domain invariance, without subscribing to the performance of a model on any end task. The theory of domain adaptation (Ben-David et al., 2010), shows that reducing $\mathcal{H}$-divergence between two domains results in higher performance in the target domain. Many works have since adopted this concept for domain adaptation in NLP (Ganin et al., 2016; Bousmalis et al., 2016). The aim is to learn representations that are invariant to the domain, while also being discriminative of a particular task. Other divergence measures such as Maximum Mean Discrepancy (MMD) (Gretton et al., 2012a), Correlational Alignment (CORAL), Central Moment Discrepancy (CMD) (Zellinger et al., 2017), have been subsequently defined and used (Ramponi and Plank, 2020; Kashyap et al., 2020). However, our community does not yet understand the inherent domain-invariance of PLM representations, particularly across different layers.

We ask key questions concerning domain invariance of PLM representations and find surpris-

[1] We borrow the term *standard data* from (Plank, 2016) to refer to news and web-like text and *non-standard data* to refer to other text like biomedical and Twitter.

Proceedings of the Second Workshop on Domain Adaptation for NLP, pages 222–244
April 20, 2021. ©2021 Association for Computational Linguistics

ing results. First, we consider vanilla PLM representations (e.g., BERT) which are trained on standard data like Wikipedia and Books (Plank, 2016). We ask: *do they exhibit domain invariance when presented with non-standard data like Twitter and biomedical text?* (§3), and *are lower layers of PLMs general and invariant compared to higher layers?* To answer these, we measure the domain divergence of PLM representations considering standard and non-standard data. We find that the lower layers of PLMs are more domain invariant compared to the upper layers (§3.2). We find that it is similar in spirit to computer vision models where lower layers of the neural network learn Gabor filters and extract edges irrespective of the image and are more transferable across tasks compared to the upper layers (Yosinski et al., 2014).

While PLMs improve the performance of tasks on standard data, Gururangan et al. (2020) improve on domain-specific tasks by continuing to pretrain RoBERTa on domain-specific data (DAPT). We thus also ask: *what happens to the domain invariance of DAPT models?* We find that compared to pretrained RoBERTa, the divergence of DAPT at a given layer either remains the same or increases, providing evidence of their specialization to a domain (§3.3). Lastly, given that standard PLMs have high training cost, we also consider the distilled model DistilBERT (Sanh et al., 2019). *What happens to the domain invariance in distilled model representations?* We find that such representations produce more domain-specific representations across layers (§3.4).

We further analyze the robustness of representations from the perspective of the encoded syntactic and semantic information across domains (§4). *Do contextualized word-level representations encode similar syntactic and semantic information even for unseen domains?* We experiment with zero-shot probes where the probes are trained on standard data only. We consider syntactic tasks like POS and NER and a semantic task – coreference resolution and find that the probes indicate similar layers encode similar amount of information, even on non-standard data.

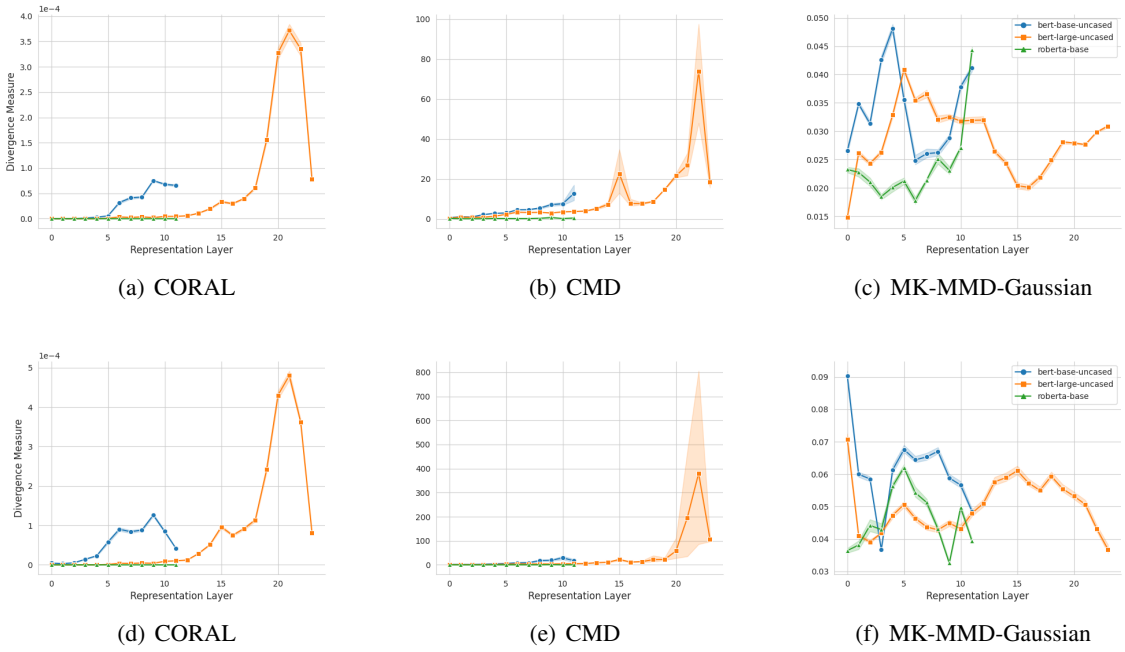
In summary, our contributions are as follows:

- We investigate the domain invariance of PLMs layer by layer and find that lower layers are more domain invariant than upper layers, which is useful for transfer learning and domain adaptation.
- Further, we analyze the robustness in terms of the syntactic and semantic information encoded in the representations across unseen domains and find that similar layers have similar amounts of linguistic information, which is a preliminary exposition of their overall performance robustness.

2 Experimental Setup

The majority of current PLMs like BERT (Devlin et al., 2019) are transformer (Vaswani et al., 2017) based models. As such we focus on the representations from different transformers. They are unsupervisedly pretrained using masked language-modeling and next sentence prediction objectives, over large amounts of English standard data such as the Books corpus (Zhu et al., 2015) and Wikipedia articles. We consider varitaion in size: two differently sized versions of the BERT model, `bert_base_uncased` — a 12 layer model and `bert_large_uncased` — a 24 layer model, both trained on lower-cased text, for comparing matters of size in representations. Next, to analyze whether training with larger data scale aids in robustness, we consider RoBERTa (Liu et al., 2019b), which is similar to BERT, but trained on a magnitude larger standard data. Further, we check the effect of distillation on domain-invariance and hence, consider DistilBERT (Sanh et al., 2019). Finally, training of models on domain-specific data is known to increase their performance on domain-specific tasks. To analyze the effect of continued fine-tuning on invariance, we consider RoBERTa pretrained on non-standard Biomedical (Gururangan et al., 2020), and Twitter (Barbieri et al., 2020) domain data. We refer to this as **DAPT-biomed** and **DAPT-tweet**, respectively. For our experiments, we use the models hosted on the `huggingface-transformer` library (Wolf et al., 2020).

Divergence Measures. We consider three different divergence measures that are widely used in the unsupervised domain adaptation literature. Correlation Alignment (CORAL) measures the difference between covariance of features – a second-order moment. Sun and Saenko (2016) reduce the distributional distance between features for unsupervised domain adaptation (UDA) in computer vision models. In contrast to CORAL, Central moment Discrepancy (CMD) considers higher-order moments of random variables to



(a) CORAL (b) CMD (c) MK-MMD-Gaussian

(d) CORAL (e) CMD (f) MK-MMD-Gaussian

Figure 1: **Top**: Comparing divergence for the standard vs. biomedical samples **Bottom**: Comparing divergence for the standard vs. twitter samples. The plots consider three divergence measures: CORAL, CMD, and MK-MMD Gaussian, for three encoders `bert-base-uncased`, `bert-large-uncased` and `roberta-base`. The values are the mean and standard deviation of divergence measures calculated over 5 splits of 1000 samples.

measure the distributional difference between features, and has been used in both NLP (Peng et al., 2018) and multimodal UDA (Hazarika et al., 2020). Finally we consider another popular measure of measuring divergence — Maximum Mean Discrepancy (Gretton et al., 2012a). Specifically, we consider the Multi-Kernel Gaussian variate (MK-MMD-Gaussian), which ensures that the statistical two sample test for the difference in distributions have high power and low test error (Gretton et al., 2012b; Long et al., 2015). We chose these measures because of their popularity, relevance and inexpensive calculations, and provide their technical details in Appendix A.

3 How Domain-Invariant are PLM Representations?

Most of the current techniques in unsupervised domain adaptation explicitly reduce the divergence between different layer representations during training (Yu et al., 2020). A common post-hoc analysis from such works shows the reduction of domain invariance at different layers. However, they do not pay much heed to the domain-invariance of representations that already exist in such models prior to domain-adapted training. Thus, we use domain divergence measures to investigate whether domain-invariance is an inherent property of pretrained transformer models, by the virtue of large-scale self-supervised learning.

3.1 Datasets and Method

We randomly sample 5000 standard data sentences from the Toronto Books corpus (Zhu et al., 2015), which is similar to the data used to train pretrained language models. We further split them into five groups of 1000 sentences for calculating our divergence measures and report the mean and variance of our results. We consider two non-standard domains. The biomedical domain similarly consists of 5000 sentences from publicly available PubMed abstracts[2] and for the Twitter domain, we sample 5000 tweets from the year 2011 made available by the archive team.[3] We follow the same procedure as Nguyen et al. (2020) to preprocess tweets: we use fastText (Joulin et al., 2017) to consider only English tweets and use the `emoji` package [4] to translate emojis into text strings, normalize all the user mentions to `@USER` and URLs to `HTTPURL`.

[2]https://www.nlm.nih.gov/databases/download/pubmed_medline.html
[3]https://archive.org/details/twitterstream
[4]https://pypi.org/project/emoji

Layer	1	2	3	4	5	6	7	8	9	10	11	12
NMI (standard-biomed)	0.004	0.365	0.046	0.63	0.722	0.596	0.245	0.092	0.164	0.312	0.63	0.588
NMI (standard-twitter)	0.256	0.252	0.215	0.233	0.698	0.699	0.727	0.695	0.753	0.772	0.762	0.68

Table 1: NMI values measuring the clustering performance at different layers of `bert-base-uncased`.

We make forward passes of 1000 samples from one pair of domains (standard–biomedical / standard–twitter) separately through the transformers, obtaining two sets of representations. We then use these to calculate divergence measures. We consider the representations of `[CLS]` token as the representation of a sentence, as done in other works. Note that we do not fine-tune any of our models on the non-standard data.

3.2 Results

Across Layers: Overall, the divergence measures increase from the lower layers to the upper layers (Figure 1). CORAL and CMD for `bert-base-uncased` and `bert-large-uncased` indicate that the divergence strictly increases. Surprisingly, the models trained on standard data extract invariant representations at the lower layers, becoming more domain-specific at the upper layers, irrespective of the domain. Both CORAL and CMD indicate a sharp decrease in divergence for the last layer for all the models. Since they are language models trained to predict the next word, they might encode representations related to the pretraining objective itself (Liu et al., 2019a). Compared to BERT-base, the divergence measures of BERT-large, at layers where they can be compared, is lower (c.f. Fig. 1 and Fig. 5 in Appendix B). Even though both the models are trained on a similar amount of data and similar training procedures, it is surprising that BERT-base has lower divergence than BERT-large.

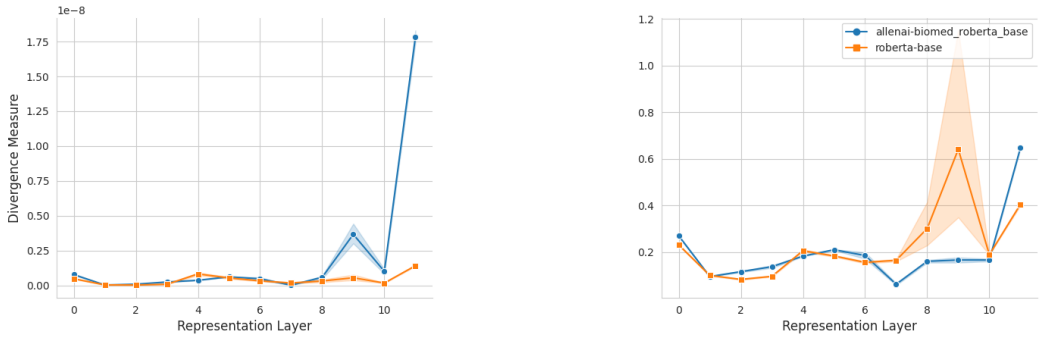
But, MK-MMD-Gaussian does not indicate a clear increase in divergence. We attribute this to the divergence measure, since MK-MMD-Gaussian is sensitive to the kernel and choosing an optimal value for its parameters is non-trivial (Gretton et al., 2012b). We confirm this by plotting the PCA representations of these data points (Figs. 8 to 13 in Appendix E.), which show that the representations from the two domains are interspersed in the lower layers and separated in the upper layers, as done in many previous works (Ganin et al., 2016; Long et al., 2015). We further quantify this by performing k-means clustering where $k = 2$ (the number of domains). We evaluate the clusters using Normalized Mutual Information (c.f. Table 1). Clustering quality is higher for upper layers compared to lower layers where representations are interspersed.

The increasing divergence across layers has plausible implications in making decisions in many scenarios. For example, in deciding the number of layers in the gradual unfreezing of layers in transfer learning (Howard and Ruder, 2018), in unsupervised domain adaptation where divergence between representations from different layers are reduced (Long et al., 2015). Recently, Aharoni and Goldberg (2020) show the final transformer layer representations cluster while Ma et al. (2019) consider penultimate layer representations. The high domain divergence of the upper layers is a plausible explanation for the clustering (Figs. 8 to 13 in Appendix E.). Clustering of representations plays a key role in downstream applications, such as data selection for machine translation and curriculum learning, data points in the source domain closest to the target domain are chosen (Axelrod et al., 2011; Moore and Lewis, 2010).

BERT vs. RoBERTa: Compared to BERT, RoBERTa has uniform divergence across layers (c.f. Fig. 1). RoBERTa is similar to BERT, but a major difference is the amount of pre-training data used (one magnitude; 160GB vs. 16GB). We speculate that the domain-invariance is because the pretraining data is an unintended mixture of different domains. Recent works have shown the impact of training models with large and diverse datasets on the robustness of image classification models (Taori et al., 2020) and text classification models (Tu et al., 2020) with similar trends observed where RoBERTa is more robust.

3.3 What happens to the domain-invariance of DAPT models?

To create domain-specific PLM, the simplest methods train models from scratch on domain-specific data like scientific publications (Beltagy et al., 2019), BioBERT (Lee et al., 2019), ClinicalBERT (Alsentzer et al., 2019) among others. In



(a) CORAL (b) CMD

Figure 2: Comparing CORAL and CMD divergences for `roberta-base` and DAPT-biomed (Gururangan et al., 2020). Considers standard and biomedical samples.

contrast, instead of pretraining from scratch, recent work shows impressive benefits of continuing to pretrain on domain-specific data — termed domain adaptive pretraining (DAPT) (Gururangan et al., 2020). Although there are improvements on domain-specific end tasks, the domain-invariance of these representations is not analyzed. We consider only the CORAL and CMD divergence measures from now, due to our observations from the previous section.

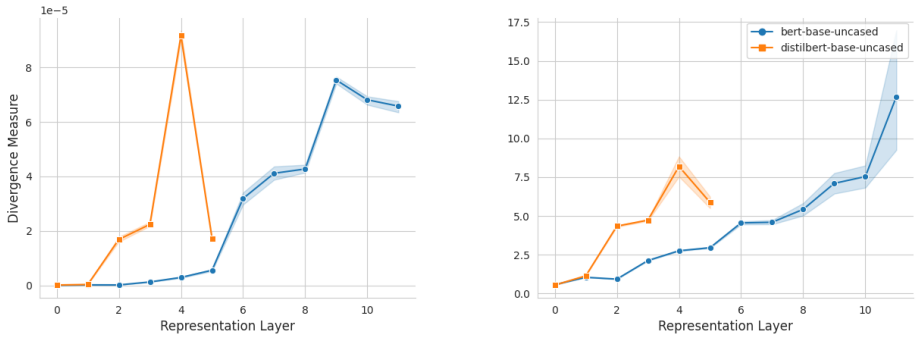
Figure 2 shows that the divergence across the layers for **DAPT-biomed** is the same as RoBERTa or is higher (c.f. Fig. 6 in Appendix C for **DAPT-twitter**). The main aim of continuing to pretrain is to make the models more domain-specific. We expect the representations to diverge from the standard representations after model training. DAPT representations possibly serve as good initial representations for fine-tuning on domain-specific end tasks like natural language inference, text classification *et cetera* (Hao et al., 2019). Teasing out the benefits of domain-specific pretraining from the task-specific fine-tuning is still unclear and warrants careful attention.

3.4 What happens to the domain-invariance after distillation?

Knowledge Distillation (Hinton et al., 2015) has been successfully used to reduce the size and inference time of PLMs. Here, a smaller student network mimics the output of a larger teacher network. We consider the DistilBERTmodel. Fig. 3 shows the comparison of divergence measures between DistilBERT with BERT for the standard and the biomedical domain pair (c.f. Fig. 7 in Appendix D for comparison with Twitter domain). DistilBERT contains half the number of layers compared to BERT. At a comparable layers, DistilBERT always has higher divergence values for both CMD and CORAL. Sanh et al. (2019) show that distillation loss that mimics the teacher's output and cosine embedding loss which aligns the student and teacher hidden states vectors, are the major contributors to the student's performance. Yet, we find that DistilBERT still has greater variance which may affect downstream tasks like text classification. Although a few models (Jiao et al., 2020; Sanh et al., 2019) reduce some notion of geometric distance between the intermediate representations of the student and the teacher, it does not guarantee that the entire linguistic knowledge and the domain-invariance of the teacher are transferred to the student model. Recent work in NLP have tried to incorporate rich information from teacher networks using contrastive learning (Tian et al., 2020; Sun et al., 2020) and by reducing the Earth Mover's distance between the hidden representations in the transformer architecture (Li et al., 2020). Related computer vision work also to impart adversarial robustness, even in the student network (Goldblum et al., 2020). The benefits of such enhanced distillation techniques on the robustness of the model is an under-explored area.

4 Robustness of Linguistic Information

How much linguistic information do representations from pretrained language models still encode for data from a different domain? Here, we evaluate the robustness of representations in `bert-base-uncased`. Do word-level repre-



(a) CORAL (b) CMD

Figure 3: CORAL and CMD divergence measures for standard vs biomedical samples. Two encoders are considered here: `bert-base-uncased` and `distilbert-base`.

sentations from PLMs encode similar levels of linguistic knowledge irrespective of the domain?

4.1 Datasets and Method

Edge Probes: Edge probes (Tenney et al., 2019b) measure the magnitude of linguistic information present in contextualized word representations. Representations of spans from a specific layer are passed through a shallow, multi-layer perceptron which predict their linguistic label. The performance of the probes indicates the magnitude of linguistic information.

To evaluate the linguistic information in BERT representations regardless of the domain, we train probes on source domain data and test on a held-out test dataset from the target domain. Since the *non-standard* data is *unseen* during training, this is a form of *zero-shot probing*, as also experimented in (Ravichander et al., 2020b). Training separate probes on every domain would yield inaccurate information about the linguistic information in the representations. The probes themselves may learn the linguistic task and overfit on the target domain data which can serve as a confounding factor.

A performance drop in probing performance between domains should not be interpreted as an absence of linguistic information. Other confounding factors like distribution difference (Recht et al., 2019; Miller et al., 2020) may be responsible. We are interested in the underlying pattern, and one has to exercise caution in interpreting the absolute performance numbers.

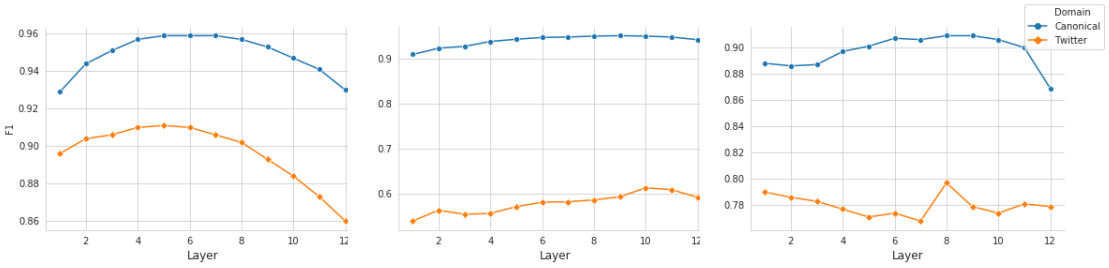
We chose three tasks from the suite of tasks defined by (Tenney et al., 2019b), where POS tagging (part-of-speech tagging), and NER (Named entity recognition) are considered syntactic, and where Coreference resolution (Coref) is considered a semantic task. We chose these tasks guided by the availability of similar datasets in both domains. For all our experiments involving probing, we use the `jiant` framework (Wang et al., 2019) .

Data: Following (Tenney et al., 2019b) we use the OntoNotes 5.0 corpus (Weischedel, Ralph et al., 2013) for probing. Since they are from newswire and web text, which is similar to the pretraining corpus of BERT (Devlin et al., 2019), we consider this dataset as *standard* data (source domain). We choose Twitter to represent *non-standard* data (target domain) for the probing task since our previous experiments showed a greater divergence, and thus are significantly different from the pretraining corpus used in BERT.

For POS tagging, we use the dataset described by (Derczynski et al., 2013). We remove the following POS tags from the dataset: USR, URL, HT, RT, "(", and ")", to normalize the labels across the domains. For NER, we use the dataset released for the shared task of the Workshop on Noisy User-generated Text (W-NUT) (Baldwin et al., 2015). For coreference resolution, we use the dataset presented in (Aktaş et al., 2018), whose annotations were later modified by (Aktaş et al., 2020) so that they were conceptually parallel to OntoNotes 5.0 corpus (Weischedel, Ralph et al., 2013). The size of the datasets across train, development and test splits were kept similar for both the domains (c.f. Appendix F).

4.2 Results

Even though the probes had not seen examples from the target domain, we observe from Fig. 4



(a) POS (b) NER (c) Coref

Figure 4: Micro F1 scores for probing POS, NER, and Coref information from `bert-base-uncased` for the standard and the non-standard Twitter domain.

that the best performing layer is the same across domains for each task. The F1 scores peak at the same layers for both the domains, across all tasks. In both domains, the F1 for the task of POS tagging peaks at layer 5; for NER and Coref, at layer 10 and 8, respectively. Considering the knowledge discovered by the probes, it can be seen that similar layers are the most important for syntactic and semantic tasks across domains.

Concerning which part of the model encodes syntactic information required for POS and NER, we observe that the middle layers perform the best for both tasks, invariant of domain. This result is consistent with the results reported for non-domain adaptation work (Liu et al., 2019a; Jawahar et al., 2019). For Coref, the upper layers perform better on the task for the source domain. This indicates that the models store the information required for Coref (Liu et al., 2019a), but that the lower layers perform better when it comes to the target domain. We speculate that this is due to the nature of Twitter-coref dataset (target domain). For tasks like coreference resolution, there is a need for the presence of semantic information to identify the co-referring entities. But as tweets are naturally shorter, they contain co-referring entities that are close to each other, and do not require long-range information. This might make it easier for BERT models to use syntactic information from the lower layers to perform well on the target domain dataset.

We note that Merchant et al. (2020) show PLM representations do not experience catastrophic forgetting when fine-tuned on different end tasks such as MNLI, SQuAD and dependency parsing. With the limited capabilities that probes have, the results of this work show that similar information is being encoded for a task in similar layers without fine-tuning on any domain-specific data. This indicates that PLM representations might encode similar linguistic information across domains to begin with, potentially aiding performance on domain-specific end tasks.

5 Limitations

Our analysis using probes for different domains is intended to be an initial exploration of this topic. We inherit all the limitations of the probing classifiers highlighted by recent works (Tenney et al., 2019a; Voita and Titov, 2020; Pimentel et al., 2020). Since probing trains shallow models, there exists a possibility that the performance confounds with the models learning the task rather than being diagnostic about the linguistic power of representations. It also does not indicate that the model uses this information effectively (Hewitt and Manning, 2019) which requires further analysis. We also consider only one target domain — Twitter — and analyze `bert-base-uncased` for our probes. The availability and varying characteristics of the dataset across domains dictates our choice. For example, compared to standard coreference, biomedical text exhibits co-referring terms across sentences in long documents.

6 Related Work

6.1 NLP Robustness

Pretrained language models (PLMs) perform well on a wide range of NLP tasks, but they do not generalize well when the test distribution is different. A robust model must adapt to the shift in distributions (Quionero-Candela et al., 2009) and generalize to out-of-distribution (OOD) examples. (Hendrycks et al., 2020) study the OOD robustness of PLMs, finding that the performance drop is

substantially smaller than their shallow LSTM and CNN counterparts. Much of the literature on PLM robustness use the notion of performance drop in a new target domain (Hendrycks et al., 2020; Tu et al., 2020; Miller et al., 2020). However, analyzing the robustness and invariance of the representations under data from different domains or adversarial examples (Zhu et al., 2020) has not received much attention thus far in domain adaptation.

Concerning the robustness of linguistic information stored in representations, Merchant et al. (2020) analyze the syntactic and semantic information preserved by PLMs, both before and after fine-tuning the models on task-specific data. Similarly, Tamkin et al. (2020) analyze the role of different layers in transfer learning on end tasks. Different from their study, we are interested in the intrinsic invariance of the PLM representations under data from different domains.

6.2 Unsupervised Domain Adaptation

For unsupervised domain adaptation, a popular method is use the adversarial training between a domain and a task classifier (DANN) (Ganin et al., 2016). Compared to DANN, where domain-specific peculiarities are lost, (Bousmalis et al., 2016) introduce domain-specific networks, which where domain-specific and domain-invariant representations are formed in a shared-private network. Another method to obtain invariant representations is to explicitly reduce the domain divergence between different layers of a neural network (Miller et al., 2020; Sun and Saenko, 2016; Shen et al., 2018b,a). For a complete treatment on UDA refer to (Ramponi and Plank, 2020) and for a review on divergence measure refer to (Kashyap et al., 2020). Considering the inherent domain-invariance of representations is thus important for UDA models.

6.3 Probing

As pretrained transformer models provide improvements on end tasks, understanding their internals and knowledge they encode has become increasingly important. For a review on efforts to understand pretrained transformers, see Rogers et al. (2021). Probing is a popular method to understand the linguistic information stored in continuous representations (Conneau et al., 2018). Tenney et al. (2019a,b) use probes to understand the linguistic information that the representations capture.

Recent work has questioned the premise of using a probe. Hewitt and Liang (2019) propose control tasks to ensure that the performance of probe is diagnostic about the linguistic information and is not because of learning the task. (Pimentel et al., 2020) utilize information theory to show that contextual representations contain similar amounts of information as lexical tokens. They suggest that better performing probes are increasingly accurate of detecting linguistic information regardless of their complexity and propose *ease of probing* as an alternative solution. This is similar to minimum description length suggested by Voita and Titov (2020). Contrary to previous works Ravichander et al. (2020a); Elazar et al. (2020) argue that presence of linguistic information does not guarantee its utility for end tasks. In contrast to these works that consider only a single domain, we provide experiments to diagnose cross domain linguistic information using probes.

7 Conclusion

We consider domain robustness from the perspective of domain-invariance of pretrained language model (PLM) representations. We observe that the lower layers of PLMs are generally domain-invariant. We also find that domain variance increases on continuously pretrained (DAPT) models and distilled models (DistilBERT). We have seen that RoBERTa is robust, possibly by virtue of training with more data. Domain adaptation methods using it should be careful in assessing the empirical benefits of their methods. As distillation becomes a mainstay method in NLP for retaining accuracy and saving training and inference costs on large models, considering distillation techniques to retain domain invariance and broadly applicable linguistic properties is of interest to the community.

Considering the inherent domain-invariance of PLM representations at various layers is possibly useful in understanding their performance on out of domain distribution data(Hendrycks et al., 2020) and for domain adaptation in general. For example, since we understand that the lower layers of BERT are domain-invariant compared to higher layers, we can freeze them during domain adaptation (Peters et al., 2019; Shen et al., 2018a) or drop them to make the models smaller and more

efficient (Sajjad et al., 2020). In the future, we will incorporate this information for domain adaptation of models.

We used edge probes (Tenney et al., 2019b) to identify the linguistic information in representations of data from different domains. One has to note that edge probes consider span representations for probing and not the representations of the entire sentence. To answer *Is there any correlation between domain-invariance of a sentence and the amount of linguistic information contained in them*?, we should consider sentence-level probes, similar to (Conneau et al., 2018; Jawahar et al., 2019) But, we are restricted by the lack of sentence-level probing data for different domains. We believe that this is a ripe area for future work.

Acknowledgements

The authors would like to thank Yisong Miao for reading a draft of this paper and providing insightful comments. We would also like to acknowledge the support of the NExT research grant funds, supported by the National Research Foundation, Prime Ministers Office, Singapore under its IRC @ SG Funding Initiative, and to gratefully acknowledge the support of NVIDIA Corporation with the donation of the GeForce GTX Titan X GPU used in this research.

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7747–7763, Online. Association for Computational Linguistics.

Berfin Aktaş, Tatjana Scheffler, and Manfred Stede. 2018. Anaphora resolution for Twitter conversations: An exploratory study. In Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference, pages 1–10, New Orleans, Louisiana. Association for Computational Linguistics.

Berfin Aktaş, Veronika Solopova, Annalena Kohnert, and Manfred Stede. 2020. Adapting coreference resolution to Twitter conversations. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2454–2460, Online. Association for Computational Linguistics.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Timothy Baldwin, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In Proceedings of the Workshop on Noisy User-generated Text, pages 126–135, Beijing, China. Association for Computational Linguistics.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1644–1650, Online. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. Mach. Learn., 79(1-2):151–175.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pages 343–351.

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, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Leon Derczynski, Alan Ritter, S. Clark, and Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In RANLP.

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.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2020. Amnesic probing: Behavioral explanation with amnesic counterfactuals.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. J. Mach. Learn. Res., 17(1):2096–2030.

Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. 2020. Adversarially robust distillation. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The

Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 3996–4003. AAAI Press.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. 2012a. A kernel two-sample test. J. Mach. Learn. Res., 13:723–773.

Arthur Gretton, Bharath K. Sriperumbudur, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, and Kenji Fukumizu. 2012b. Optimal kernel choice for large-scale two-sample tests. In Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States, pages 1214–1222.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342–8360, Online. Association for Computational Linguistics.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2019. Visualizing and understanding the effectiveness of BERT. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4143–4152, Hong Kong, China. Association for Computational Linguistics.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM International Conference on Multimedia, MM '20, page 1122–1131, New York, NY, USA. Association for Computing Machinery.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2744–2751, Online. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. CoRR, abs/1503.02531.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, pages 328–339. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4163–4174, Online. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, and Roger Zimmermann. 2020. Domain divergences: a survey and empirical analysis.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234–1240.

Jianquan Li, Xiaokang Liu, Honghong Zhao, Ruifeng Xu, Min Yang, and Yaohong Jin. 2020. BERT-EMD: Many-to-many layer mapping for BERT compression with earth mover's distance. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3009–3018, Online. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. CoRR, abs/1907.11692.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. 2015. Learning transferable features with deep adaptation networks. In Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, volume 37 of JMLR Workshop and Conference Proceedings, pages 97–105. JMLR.org.

Xiaofei Ma, Peng Xu, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2019. Domain adaptation with BERT-based domain classification and data selection. In Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019), pages 76–83, Hong Kong, China. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 33–44, Online. Association for Computational Linguistics.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. 2020. The effect of natural distribution shift on question answering models. In Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 6905–6916. PMLR.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In Proceedings of the ACL 2010 Conference Short Papers, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 9–14.

Minlong Peng, Qi Zhang, Yu-gang Jiang, and Xuanjing Huang. 2018. Cross-domain sentiment classification with target domain specific information. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2505–2513, Melbourne, Australia. Association for Computational Linguistics.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), pages 7–14, Florence, Italy. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4609–4622, Online. Association for Computational Linguistics.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. In Proceedings of the 13th Conference on Natural Language Processing, KONVENS 2016, Bochum, Germany, September 19-21, 2016, volume 16 of Bochumer Linguistische Arbeitsberichte.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. 2009. Dataset Shift in Machine Learning. The MIT Press.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In Proceedings of the 28th International Conference on Computational Linguistics, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2020a. Probing the probing paradigm: Does probing accuracy entail task relevance?

Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020b. On the systematicity of probing contextualized word representations: The case of hypernymy in BERT. In Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics, pages 88–102, Barcelona, Spain (Online). Association for Computational Linguistics.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet classifiers generalize to ImageNet? In Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 5389–5400. PMLR.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. Transactions of the Association for Computational Linguistics, 8(0):842–866.

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. 2020. Poor man's bert: Smaller and faster transformer models.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR, abs/1910.01108.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. 2018a. Wasserstein distance guided representation learning for domain adaptation. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pages 4058–4065. AAAI Press.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. 2018b. Wasserstein distance guided representation learning for domain adaptation. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pages 4058–4065. AAAI Press.

Baochen Sun and Kate Saenko. 2016. Deep CORAL: correlation alignment for deep domain adaptation. CoRR, abs/1607.01719.

Siqi Sun, Zhe Gan, Yuwei Fang, Yu Cheng, Shuohang Wang, and Jingjing Liu. 2020. Contrastive distillation on intermediate representations for language model compression. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 498–508, Online. Association for Computational Linguistics.

Alex Tamkin, Trisha Singh, Davide Giovanardi, and Noah Goodman. 2020. Investigating transferability in pretrained language models. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1393–1401, Online. Association for Computational Linguistics.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. 2020. Measuring robustness to natural distribution shifts in image classification.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

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. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive representation distillation. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. Transactions of the Association for Computational Linguistics, 8:621–633.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 183–196, Online. Association for Computational Linguistics.

Alex Wang, Ian F. Tenney, Yada Pruksachatkun, Phil Yeres, Jason Phang, Haokun Liu, Phu Mon Htut, , Katherin Yu, Jan Hula, Patrick Xia, Raghu Pappagari, Shuning Jin, R. Thomas McCoy, Roma Patel, Yinghui Huang, Edouard Grave, Najoung Kim, Thibault Févry, Berlin Chen, Nikita Nangia, Anhad Mohananey, Katharina Kann, Shikha Bordia, Nicolas Patry, David Benton, Ellie Pavlick, and Samuel R. Bowman. 2019. `jiant` 1.3: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

Weischedel, Ralph, Palmer, Martha, Marcus, Mitchell, Hovy, Eduard, Pradhan, Sameer, Ramshaw, Lance, Xue, Nianwen, Taylor, Ann, Kaufman, Jeff, Franchini, Michelle, El-Bachouti, Mohammed, Belvin, Robert, and Houston, Ann. 2013. Ontonotes release 5.0.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pages 3320–3328.

Weijie Yu, Chen Xu, Jun Xu, Liang Pang, Xiaopeng Gao, Xiaozhao Wang, and Ji-Rong Wen. 2020. Wasserstein distance regularized sequence representation for text matching in asymmetrical domains. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2985–2994, Online. Association for Computational Linguistics.

Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. 2017. Central moment discrepancy (CMD) for domain-invariant representation learning. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension. CoRR, abs/1810.12885.

Sicheng Zhu, Xiao Zhang, and David Evans. 2020. Learning adversarially robust representations via worst-case mutual information maximization. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research, pages 11609–11618. PMLR.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV '15, page 19–27, USA. IEEE Computer Society.

A Divergence Measures

Maximum Mean Discrepancy (MMD): MMD is a non-parametric method to estimate the distance between distributions based on Reproducing Kernel Hilbert Spaces (RKHS). Given two random variables $X = \{x_1, x_2, ..., x_m\}$ and $Y = \{y_1, y_2, ..., y_n\}$ that are drawn from distributions P and Q, the empirical estimate of the distance between distribution P and Q is given by

$$MMD(X,Y) = \left\| \left(\frac{1}{m}\sum_{i=1}^{m}\phi(x_i) - \frac{1}{n}\sum_{i=1}^{n}\phi(y_i)\right)\right\|_{\mathcal{H}} \quad (1)$$

Here $\phi : \mathcal{X} \rightarrow \mathcal{H}$ are nonlinear mappings or of the samples to a feature representation in a RKHS, called kernels. In this work, we map the contextual word representations of the text to RKHS. Various kernels can be used for this purpose. Some of the kernels are given below.

Rational Quadratic Kernel

$$\phi(x,y) = \left(1 + \frac{1}{2\alpha}(x-y)^T\Theta^{-2}(x-y)\right)^{-\alpha}$$

Energy

$$\phi(x,y) = -\left\|x-y\right\|_2$$

Gaussian

$$\phi(x,y) = exp(-\frac{\|x-y\|_2^2}{\gamma})$$

Laplacian

$$\phi(x,y) = exp(-\frac{\|x-y\|_2}{\sigma})$$

In this work we use a mixture of Gaussian Kernels rather than a single kernel which is known to be more stable than just using a single kernel. The mixture of kernels are given by

$$\mathcal{K} = \left\{ \sum_{i=1}^{m}\lambda_i k_i \mid \sum_{i=1}^{m}\lambda_i = 1 \right\} \quad (2)$$

We set λ_i to be $\frac{1}{m}$. We follow (Long et al., 2015) and use the Gaussian kernel. We calculate a initial value γ_s and set it to the median pairwise distances between two samples, also known as *median heuristic*. For every kernel k_i, the value of γ is set from $2^{-8}\gamma_s$ and $2^8\gamma_s$, increasing it by a multiple of 2.

Correlation Alignment (CORAL): Correlation alignment is the distance between the second-order moment of the source and target samples. If d is the representation dimension, $\|\|_F$ represents Frobenius norm and Cov_S, Cov_T is the covariance matrix of the source and target samples, then CORAL is defined as

$$D_{CORAL} = \frac{1}{4d^2}\left\|Cov_S - Cov_T\right\|_F^2 \quad (3)$$

Central Moment Discrepancy (CMD): Central Moment Discrepancy is another metric that measures the distance between source and target distributions. It not only considers the first moment and second moment, but also other higher-order moments. While MMD operates in a projected space, CMD operates in the representation space. If P and Q are two probability distributions and $X = \{X_1, X_2, ..., X_N\}$ and $Y = \{Y_1, Y_2, ..., Y_N\}$ are random vectors that are independent and identically distributed from P and Q and every component of the vector is bounded by $[a, b]$, CMD is then defined by

$$CMD(P,Q) = \frac{1}{|b-a|}\left\|E(X) - E(Y)\right\|_2 + \sum_{k=2}^{\infty}\frac{1}{|b-a|^k}\left\|c_k(X) - c_k(Y)\right\|_2 \quad (4)$$

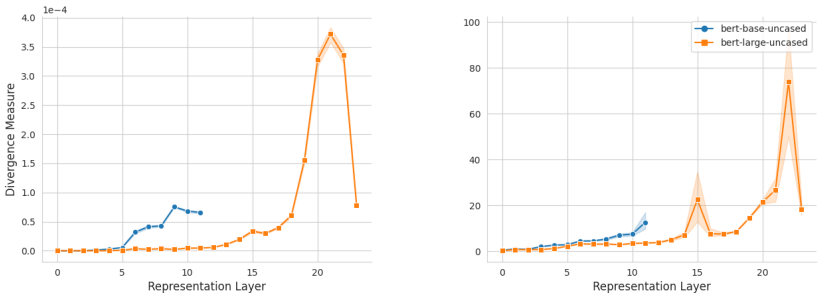
where $E(X)$ is the expectation of X and c_k is the $k-th$ order central moment which is defined as

$$c_k(X) = E\left(\prod_{i=1}^{N}(X_i - E(X_i))^{r_i}\right) \quad (5)$$

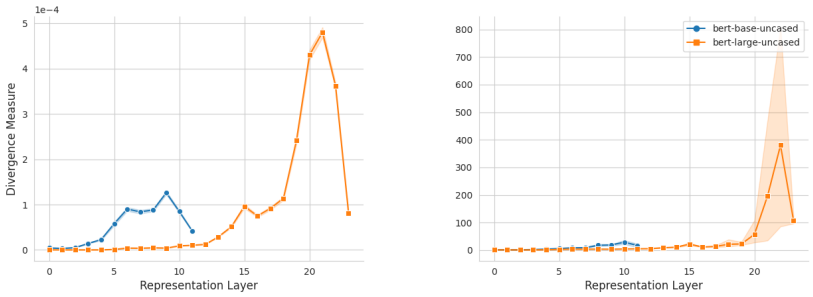
and $r_1 + r_2 + r_N = k$ and $r_1....r_N \geq 0$

B Domain Divergence Plots

Fig. 5 shows the domain divergence measures for `bert-base-uncased` and `bert-large-uncased` models only for greater quality. We consider only the CORAL and CMD divergence measures. Even though both the models are trained on similar amounts of data, with similar training procedures, at comparable layers `bert-large-uncased`-models are more domain-invariant compared to `bert-base-uncased`.



(a) CORAL (b) CMD



(c) CORAL (d) CMD

Figure 5: Comparing divergence of `bert-base-uncased` and `bert-large-uncased` models. **Top:** Divergence for standard and biomedical domain **Bottom**: Divergence for the standard and twitter domain.

C DAPT Twitter Divergence Plots

Divergence plots for DAPT-twitter compared with its `roberta-base` counterpart are shown in Fig. 6. Here we consider the CORAL and CMD divergence measures. The plots show that for DAPT-twitter the divergence measures are either the same or more than their `roberta-base` counterpart which indicates their specialization for a domain.

D DistilBERTvs BERT for twitter

Fig. 7 shows the comparison between DistilBERT – the student network and the `bert-base-uncased` which is the teacher network in knowledge distillation (Hinton et al., 2015). We consider the CORAL and CMD divergence measures. The plots show that at comparable layers DistilBERT divergence measures are strictly more than `bert-base-uncased`.

E PCA Plots for PLM Representations

Figures 8 to 15 show the 2 dimensional PCA plots for representations from different layers for all the encoders – `bert-base-uncased` , `bert-large-uncased` , `roberta-base` , `distilbert-base`. Interspersed representations in the plots indicate more domain-invariance (less domain divergence) between the representations and clearer separation indicates lesser domain divergence (more domain divergence).

F Probing classifiers' training details

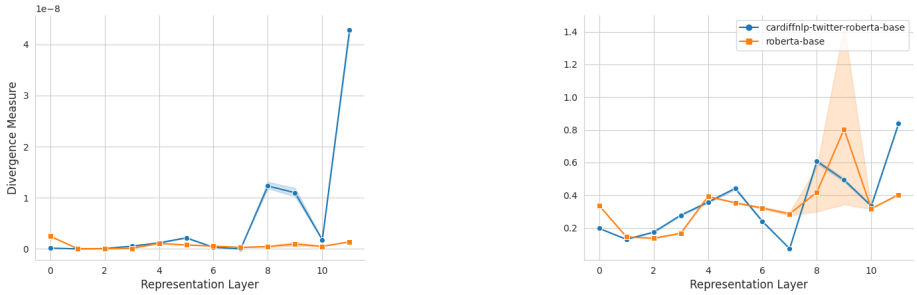
Task	Train	Dev	Test
POS	110,514	15060	11,462
NER	720	150	690
Coref	2255	101	101

Table 2: Train/dev/test split used for the probing experiments in §4 for each task.

.

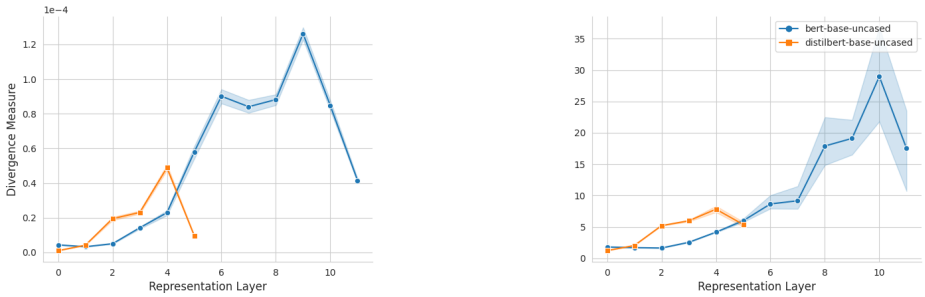
The statistics about the datasets used for training the probes are consolidated in Table 2.

For all the three tasks (POS, NER and coreference resolution) we train the probing classifiers on the source domain for 3 epochs. We use Adam as the optimizer with a learning rate of 1e-4, and a batch size of 32. We also evaluate on the validation dataset every 1000 steps, and halve the learning rate if no improvement is seen in 5 validations. The rest of the hyperparameters are the



(a) CORAL (b) CMD

Figure 6: Comparing divergences for RoBERTa vs DAPT-twitter (Barbieri et al., 2020). We consider the CORAL and CMD divergence measures for standard vs twitter samples. The plots are shown for CORAL and CMD divergence measures.



(a) CORAL (b) CMD

Figure 7: BERT vs DistilBERT models for standard vs twitter domains. CORAL and CMD divergence measures for standard vs twitter samples. Two encoders are considered here. `bert-base-uncased` which is the teacher and `distilbert-base` which is the student of knowledge distillation. The domain-invariance of `distilbert-base` is always larger than it's teacher.

Figure 8: PCA plots for representation of `bert-base-uncased` for the pair of standard and the biomedical domain for different layers. The PCA representations show that representations are interspersed in the lower layers with a progressively clearer separation in the later layers.

same as defined by (Tenney et al., 2019a) in their edge probing experiments.

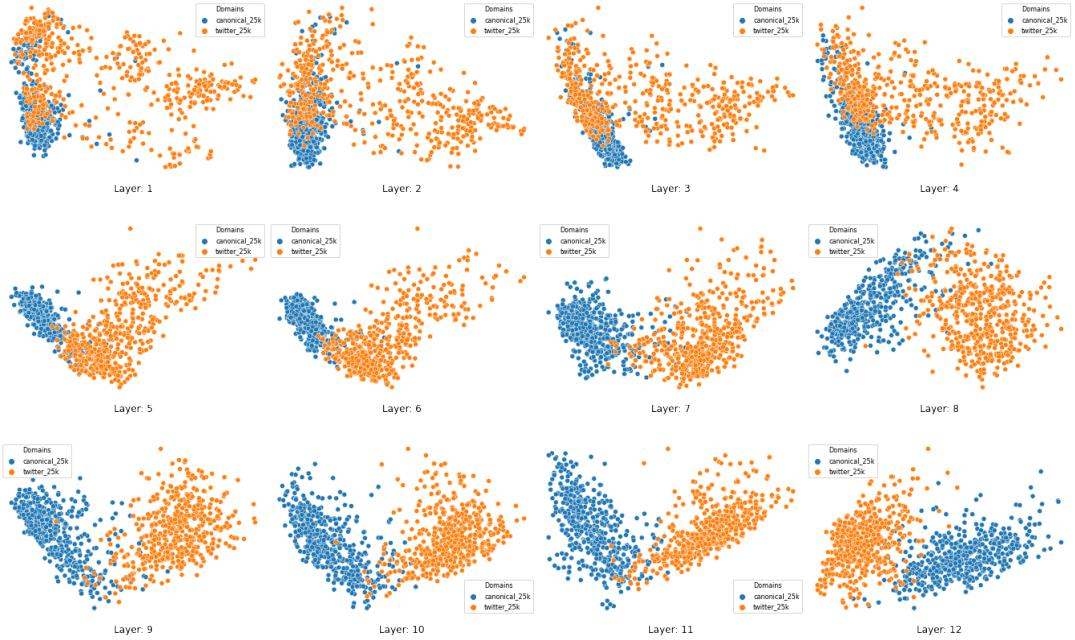


Figure 9: PCA plots for representation of `bert-base-uncased` for the pair of standard and the twitter domain for different layers. The PCA representations show that representations are interspersed in the lower layers with a progressively clearer separation in the later layers.

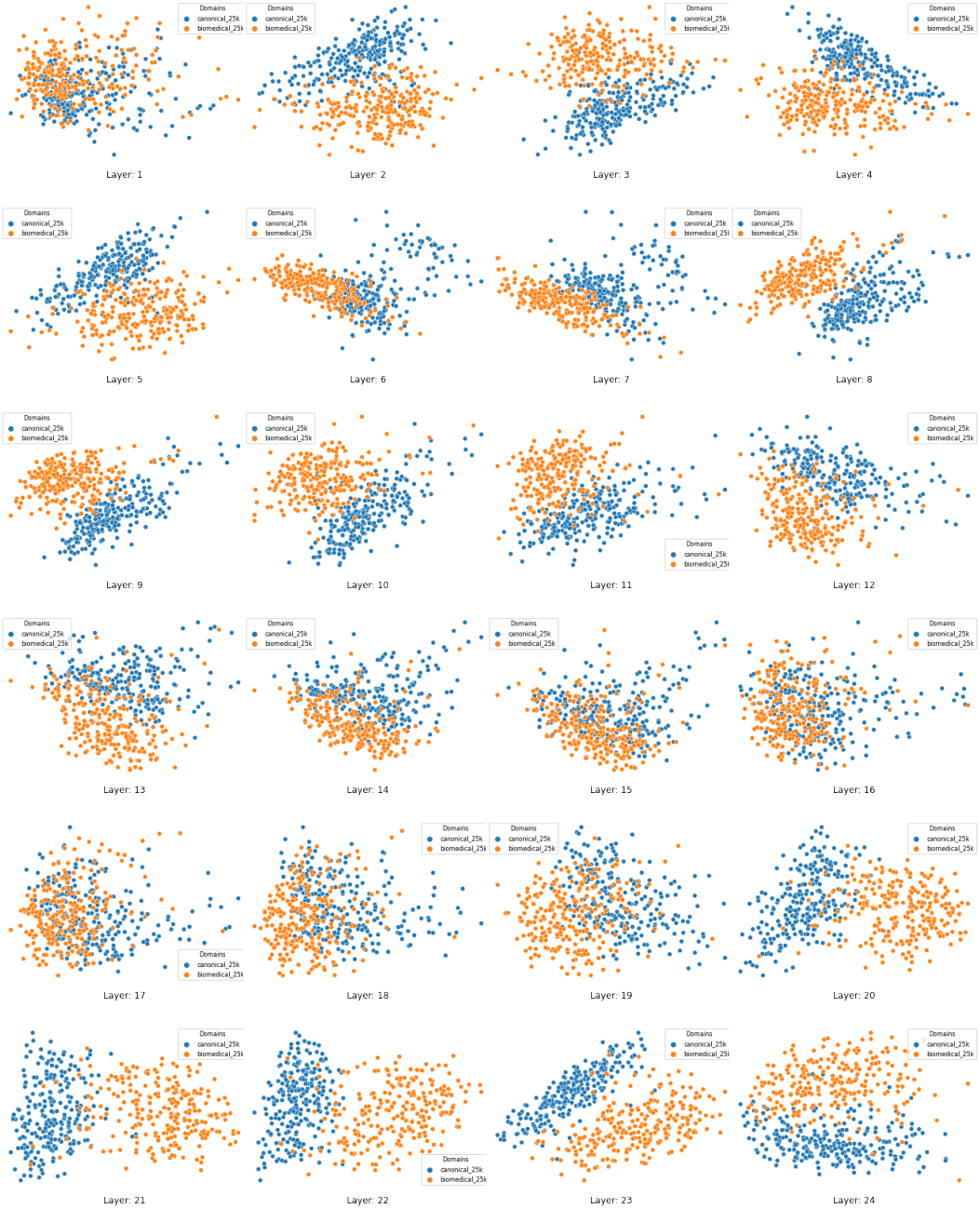


Figure 10: PCA plots for representation of `bert-large-uncased` for the pair of standard and the biomedical domain for different layers. The PCA representations show that representations are interspersed in the lower layers with a progressively clearer separation in the later layers.

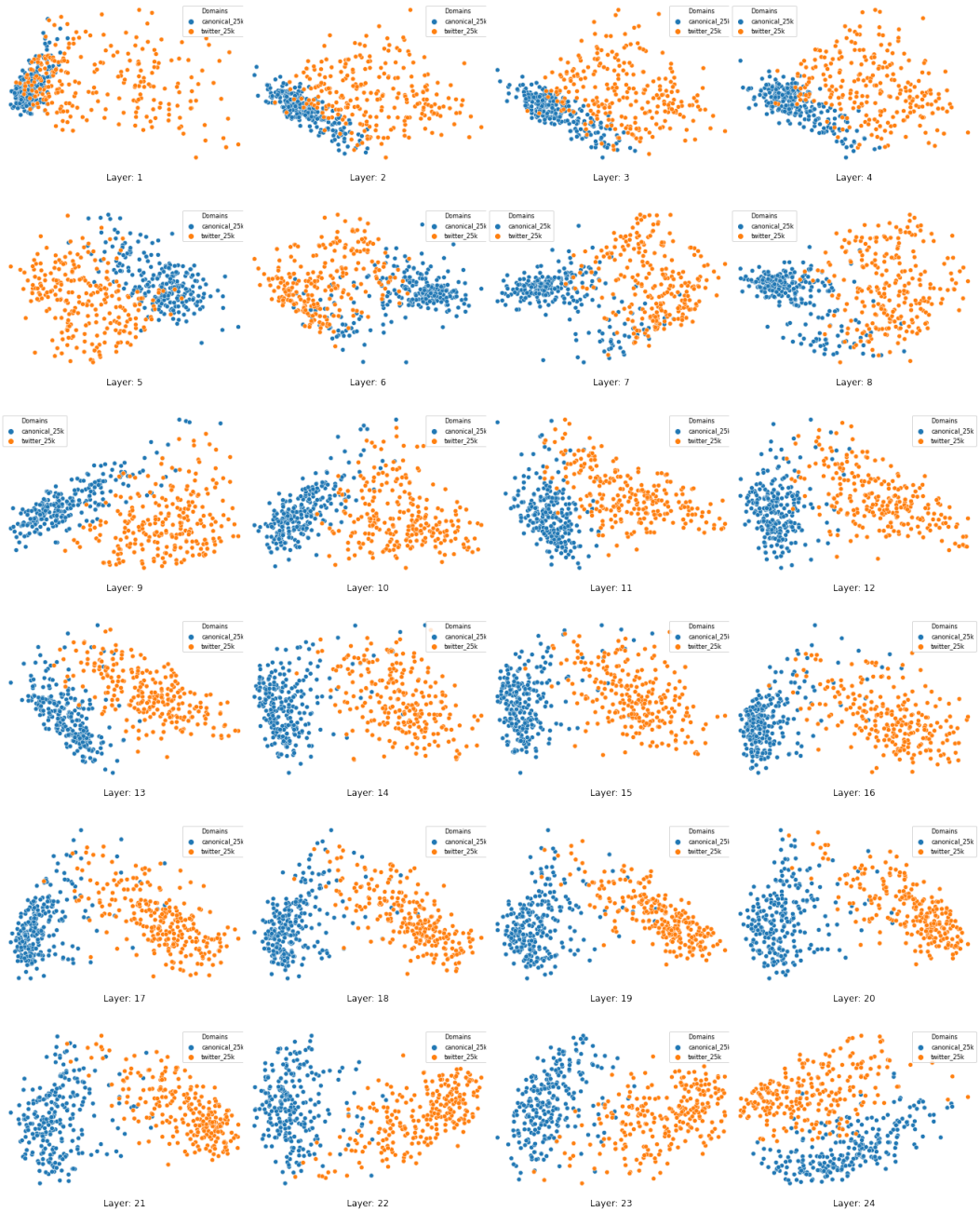


Figure 11: PCA plots for representation of `bert-large-uncased` for the pair of standard and the twitter domain for different layers. The PCA representations show that representations are interspersed in the lower layers with a progressively clearer separation in the later layers.

Figure 12: PCA plots for representation of `roberta-base` for the pair of standard and the biomedical domain for different layers. The PCA representations compared to the `bert-base-uncased` and `bert-large-uncased` models, the representations are interspersed across all the layers

Figure 13: PCA plots for representation of `roberta-base` for the pair of standard and the twitter domain for different layers. The PCA representations compared to the `bert-base-uncased` and `bert-large-uncased` models, the representations are interspersed across all the layers

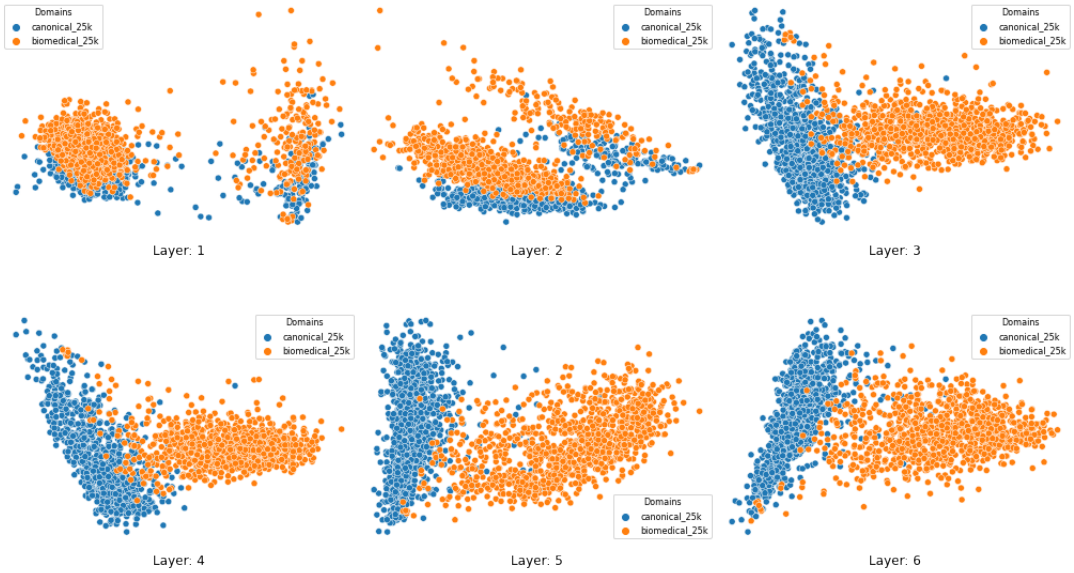


Figure 14: PCA plots for representation of `distilbert-base` for the pair of standard and the biomedical domain for different layers. The lower layers are still interspersed, with a clearer separation in the higher layers. We can see a corresponding increase in the divergence measures for these layers.

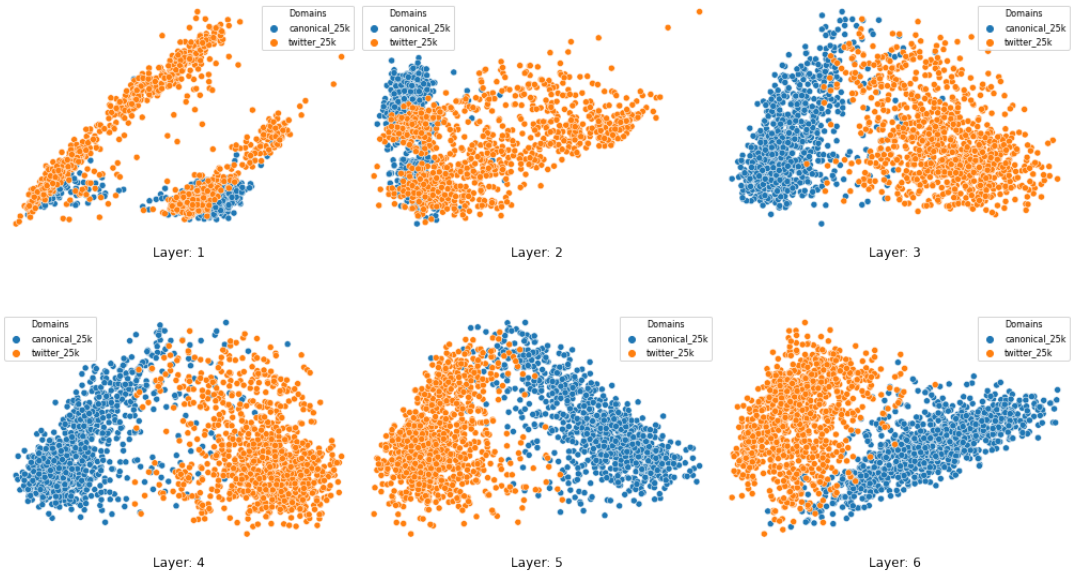


Figure 15: PCA plots for representation of `distilbert-base` for the pair of standard and the biomedical domain for different layers. The lower layers are still interspersed, with a clearer separation in the higher layers. We can see a corresponding increase in the divergence measures for these layers.