Exploring Self-awareness of Domain in Language Models

Tal Ifargan, Ziv Keidar

TECHNION Israel Institute of Technology

Introduction

We investigate the weights of large language models (LLMs) in domain adaptation tasks. We aim to find out how LLMs encode domain information, whether it is shared across domains, and how to take advantage of this knowledge for unsupervised domain adaptation. We conduct experiments on BERT model over different domains, fine-tuning it for the task of sentiment analysis. We show that models encode information about the domain throughout all the layers. Inspired by this result, we try different methods for unsupervised domain adaptation.

Data

The Amazon Reviews dataset for Sentiment Analysis

5 chosen domains


Automotives


Electronics

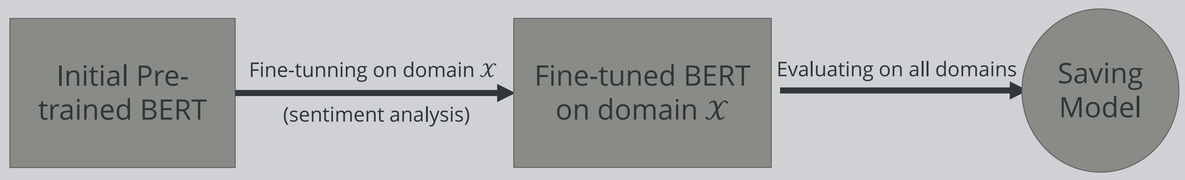

Movies and TV


Pet-supplies


Sports and Outdoors

Method

Running The following experiment 50 times for each chosen domain $\mathcal{X}$:

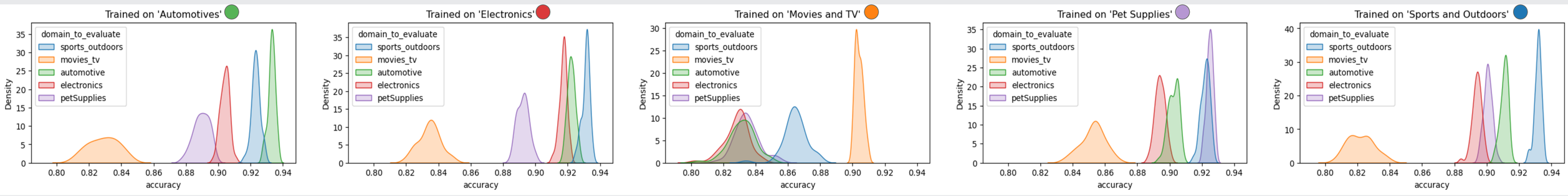


- In each experiment, we sample 100K training examples using a **different seed.**
- The initial pre-trained BERT model is **fixed** for all experiments.
- Evaluation is done using a sample of 10K examples from the target domain.

Post training model inventory:

- 5 Domains
- 50 fine-tuned models per domain
- Total of 250 BERT models

Cross-Domain Evaluation

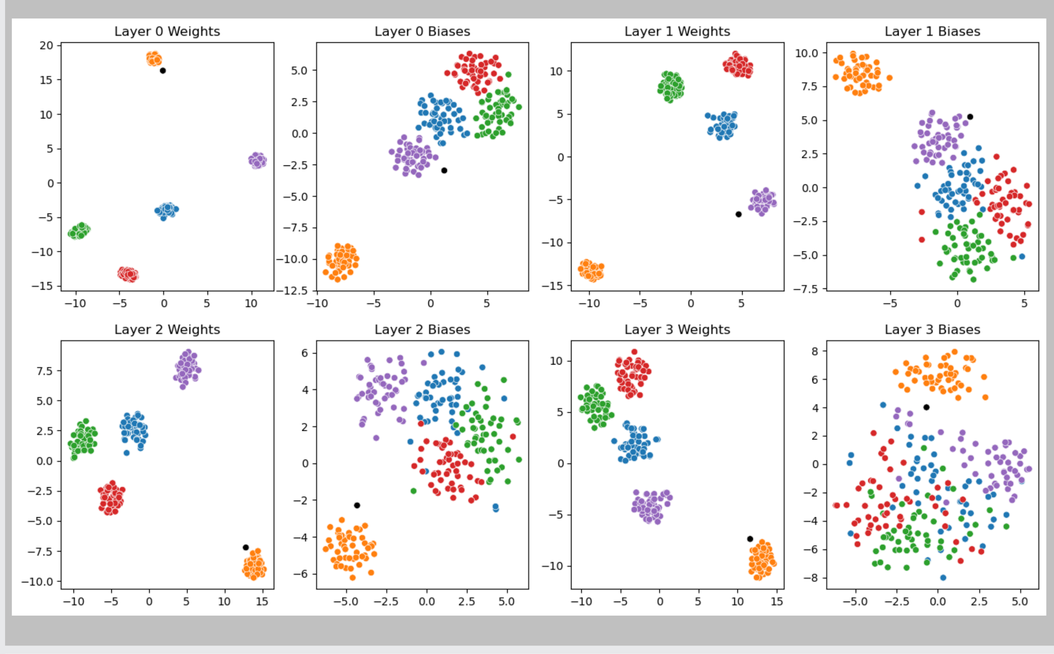


* Similar results obtained using F1 score metric for evaluation.

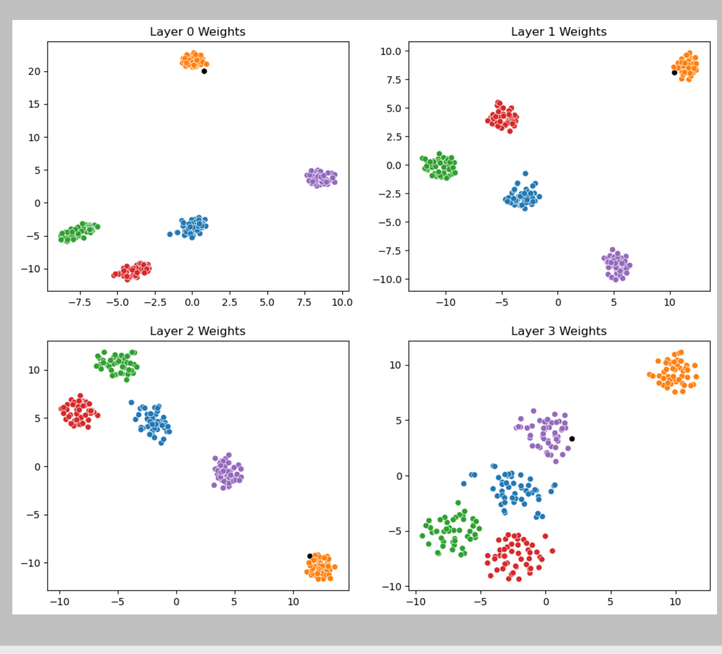
Weight Exploration - Results

Using our model inventory, we used TSNE to project different model components to a 2D space:

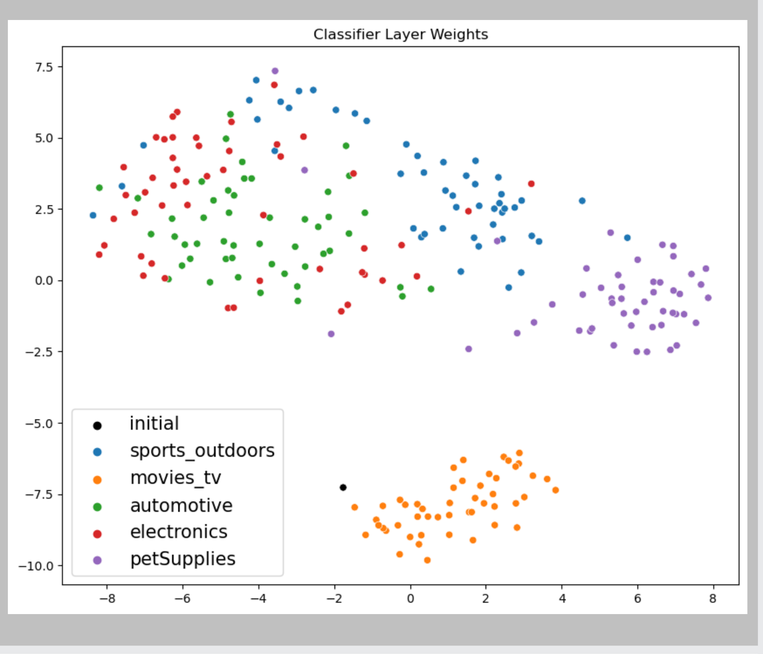
Attention Value



Linear Weights



Classifier Weights



Advanced Part – Towards UDA

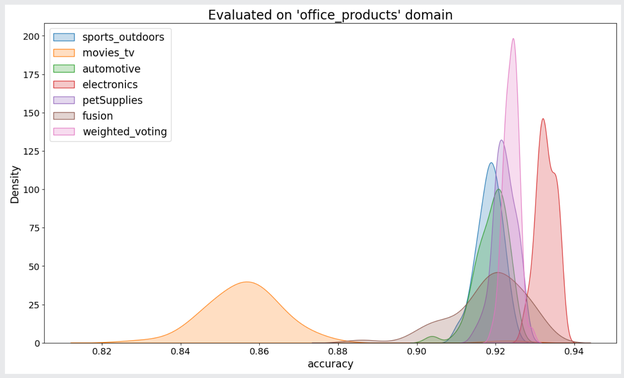
New Distance Metric

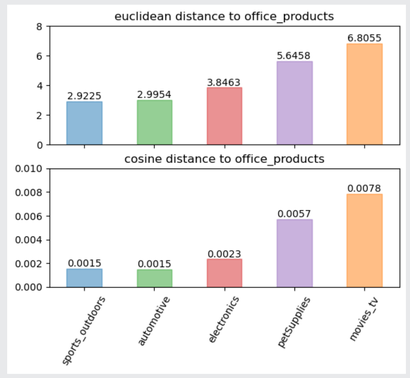
- Based on the data only
- For each domain, an “average” contextual embedding is calculated
- The distance is the Euclidean distance between the domains

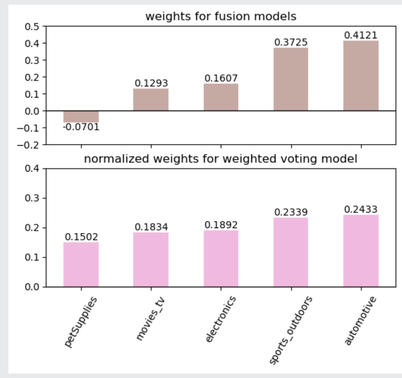
Suggestions for UDA

- Find distance of target domain from all known domains and:
  - Choose closest domain model as classifier for target domain (zero shot)
  - Create weighted weights ensemble of models
  - Create weighted voting ensemble of models

Results - Unseen Domain







Conclusion

We learned that models **are self-aware about their domain**, they encode domain knowledge throughout the whole layers, especially in the more shallow layers. We also suggested several method for UDA that were based on the fact that the model do encode information about their domain.