Understanding Domain Learning in Language Models Through Subpopulation Analysis

Zheng Zhao Yftah Ziser Shay B. Cohen

School of Informatics, University of Edinburgh

Overview

- Probing is a way to test the properties of neural representations by training additional classifiers. However, probing random presentations shows they sometimes encode important information about a task, pointing to a difficulty of using them [2].
- We suggest to study neural representations by creating an analog to control/treatment trials. An **experimental model** is trained on the general data, and a **control model** is trained from a subset of the data that satisfies a specific property.
- The comparison of the representations from the control and experimental model is done using **Singular Vector Canonical Correlation Analysis** [1], a method to calculate correlation between two sets of vectors that are linearly projected into two spaces to maximize such correlation.
- We use this methodology to test properties of neural networks with respect to **lexical domains** that exist in the data. We investigate **how different domains are encoded** in modern neural network architectures.

Methodology

We first train two kinds of models: an experimental model E, trained using data with a mixture of domains , and control models C_i, trained using a single domain i. Then we obtain two set of latent representations, from examples fed to E and C_i. Finally, we calculate SVCCA between these two representations.

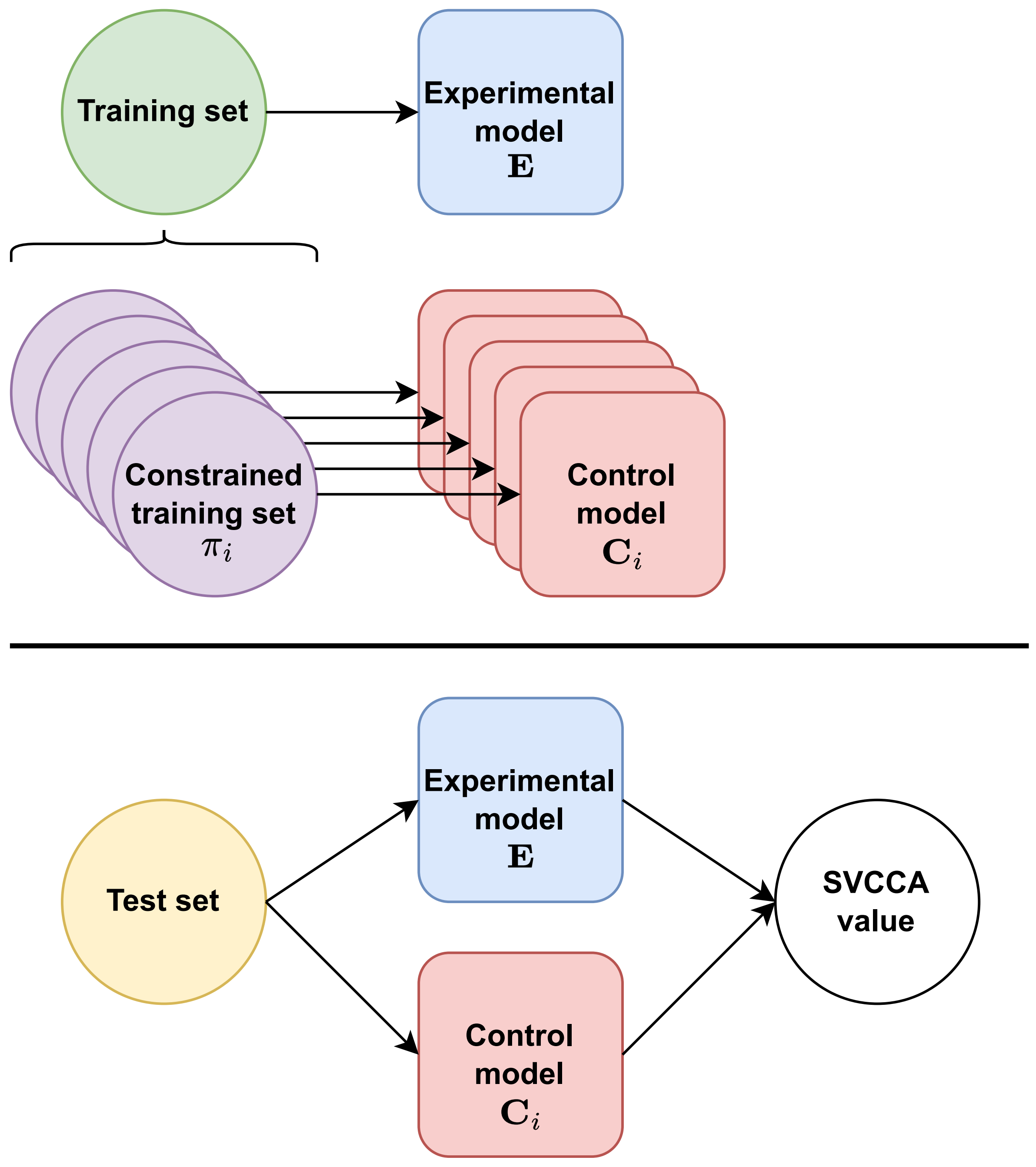

Training set
Experimental model $\mathbf{E}$
Constrained training set π_i
Control model $\mathbf{C}_i$
Test set
Experimental model $\mathbf{E}$
Control model $\mathbf{C}_i$
SVCCA value

Figure 1. A diagram illustration for our proposed pipeline.

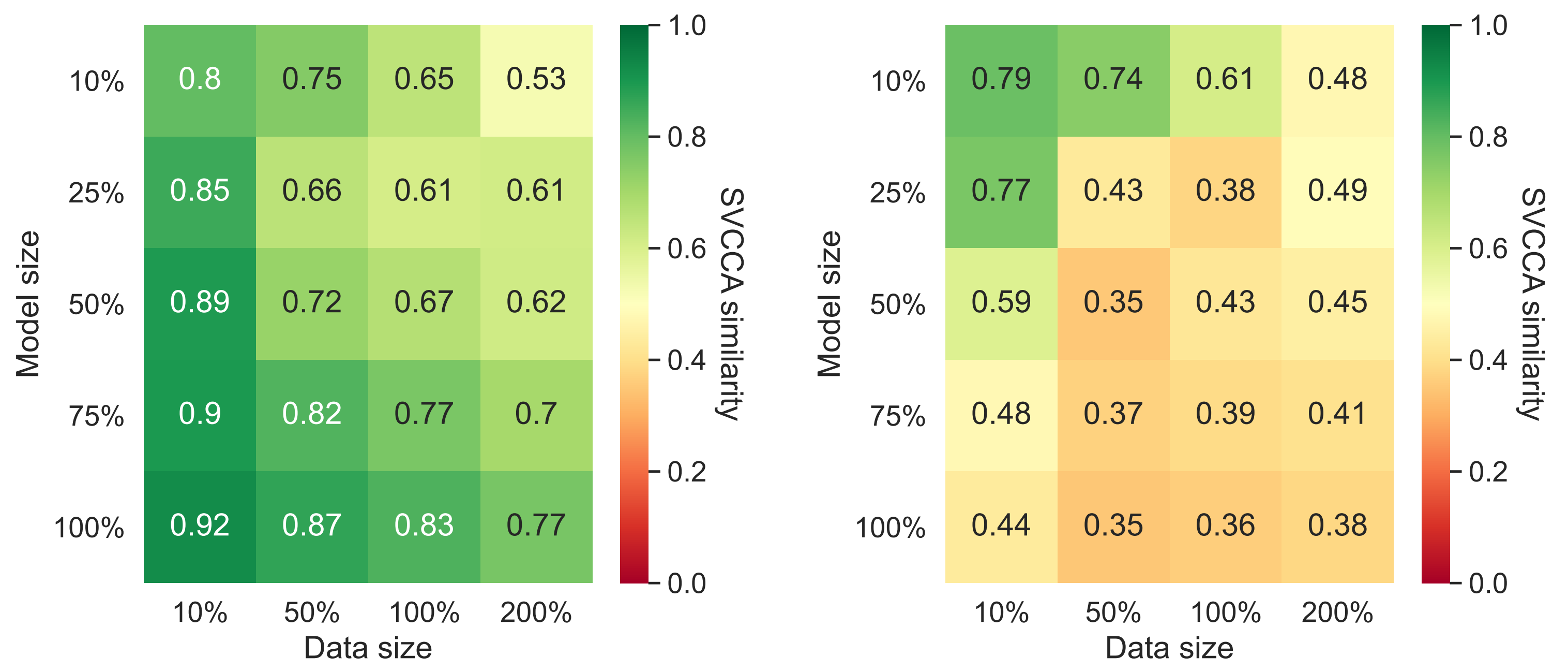
Experimental Setup

- **DATA:** the **Amazon Reviews** dataset and the **WikiSum** dataset. For Amazon Reviews, we pick the top five domains by review counts. For WikiSum we concatenate the document and summary together. To study the effect of data size on model representation, we create different data splits: **10%**, **50%**, **100%**, and **200%** splits.
- **TASK:** masked language modelling (MLM).
- **MODEL:** BERT$_{\text{BASE}}$ model trained from scratch. We also experiment with a reduced model capacity of **75%**, **50%**, **25%**, and **10%** by reducing the dimension of the hidden layers.

Main Findings

1. As the capacity of the experimental model increases, more domain-specific information is stored in the embedding layer (ℓ_0), and less in the final layer (ℓ_{12}).

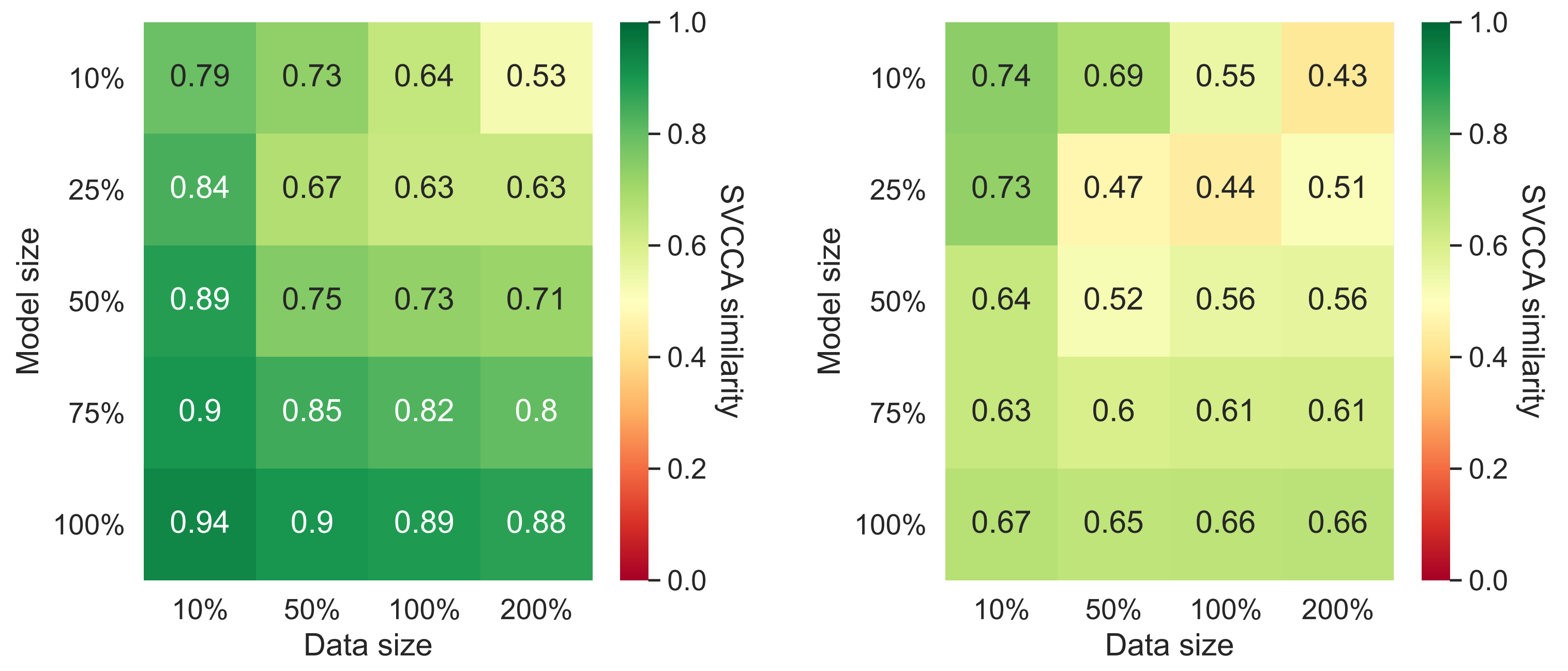
2. Increasing the data size results in less domain-specific information stored for both layers.



(a) ℓ_0: all words (b) ℓ_{12}: all words

Figure 2. The SVCCA scores between E and C_{Books} for different data sizes and model capacities.

3. As the capacity of the experimental model increases, it stores more domain information for domain-specific words, for both the embedding layer and the final layer.

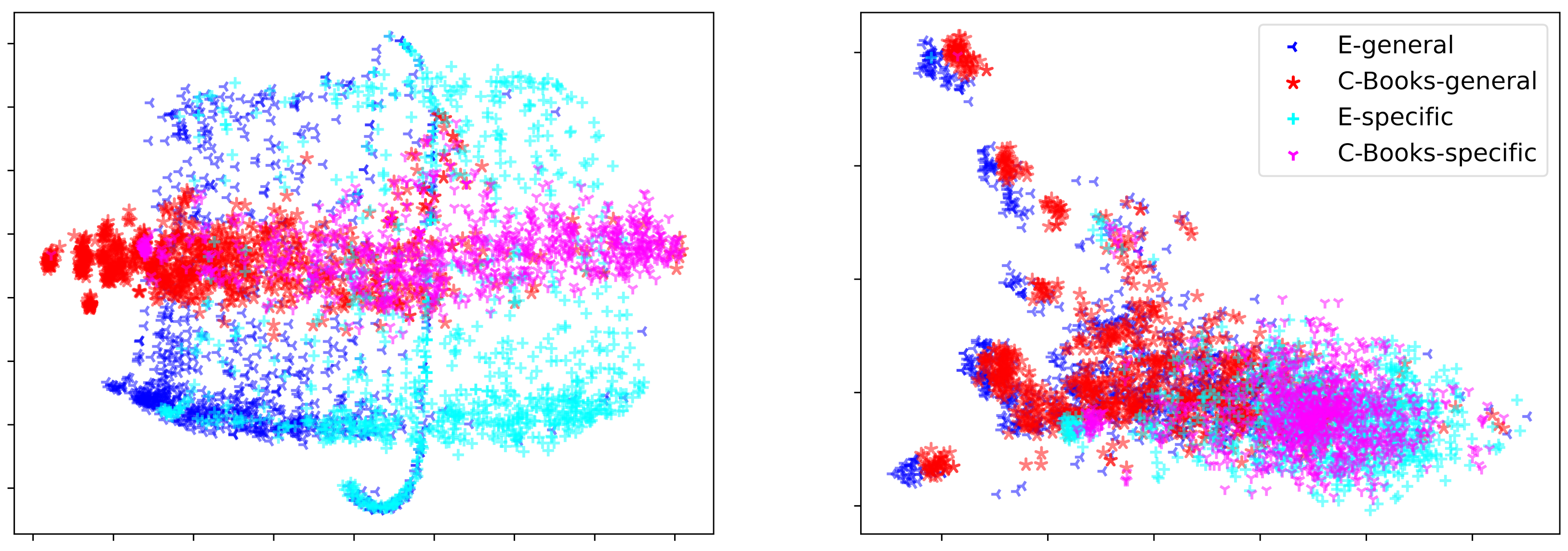


(a) ℓ_0: domain-specific words (b) ℓ_{12}: domain-specific words

Figure 3. The SVCCA score between E and C_{Books} for the domain-specific subset of tokens.

We also validate our main findings using other domains in Amazon Reviews, and WikiSum. We observe that the trend in SVCCA scores across different scenarios is generally consistent.

Figure 4 provides an example visualization of our subpopulation analysis tool. As model capacity increases, the embedding representations (ℓ_0) from E and C_{Books} models are more aligned, supporting *Finding 1*.



(a) ℓ_0: m=10% (b) ℓ_0: m=100%

Figure 4. An example visualization of general words' and domain-specific words' embedding representations for E (≺/+) and C_{Books} (⋆/Y). m denotes model capacity.

References

[1] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems 30*, pages 6076–6085. Curran Associates, Inc., 2017.

[2] Naomi Saphra and Adam Lopez. Understanding learning dynamics of language models with SVCCA. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3257–3267, Minneapolis, Minnesota, June 2019.