Isotropy in the Contextual Embedding Space: Clusters and Manifolds

Xingyu Cai, Jiaji Huang, Yuchen Bian, Kenneth Church

Baidu Research, 1195 Bordeaux Dr, Sunnyvale, CA 94089, USA
{xingyucai,huangjiaji,yuchenbian,kennethchurch}@baidu.com

Isotropy in the Contextual Embedding Space: Clusters and Manifolds

- Motivation
  - Recent studies show that the contextual embedding space for deep language models, e.g. BERT, is strongly anisotropic [1] [2].
  - We know that Isotropy often makes the space more effectively utilized and more robust to perturbations (no extreme directions that lead to high condition number).
  - It is counter-intuitive and not clear why those contextual embedding models perform remarkably well, given their anisotropic embeddings bring all the vectors close together, hard to distinguish one from another.

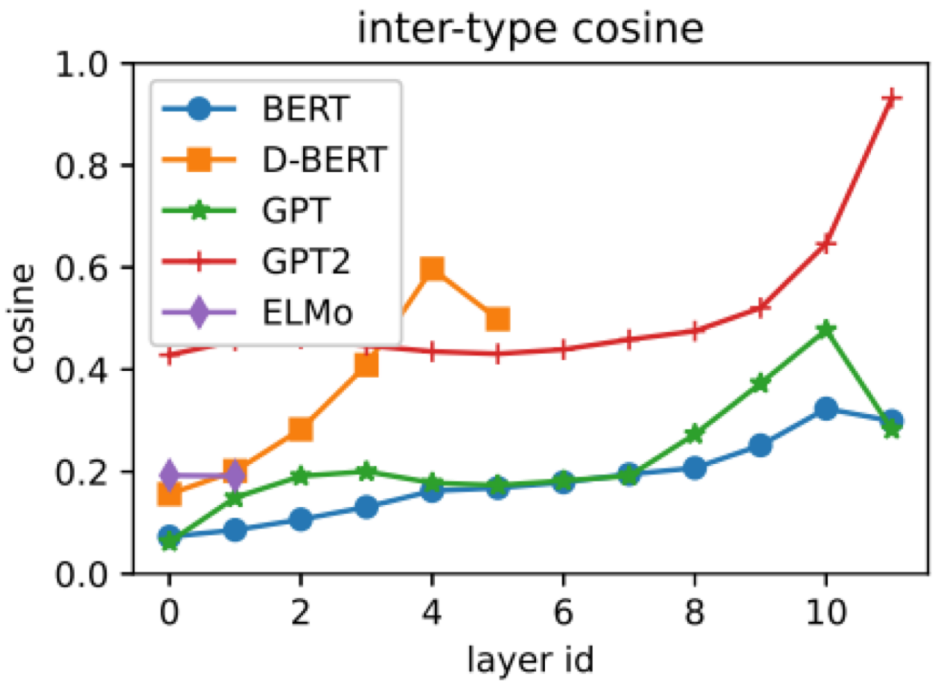
[1] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. EMNLP 2019.
[2] Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. Representation degeneration problem in training natural language generation models. ICLR 2019.

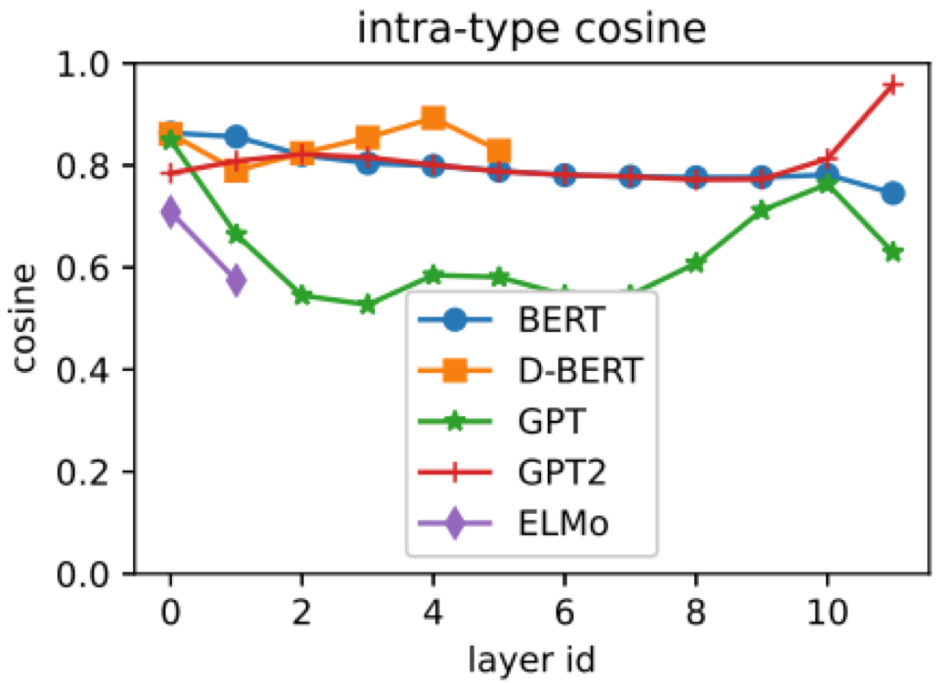
Isotropy in the Contextual Embedding Space: Clusters and Manifolds

- Notation
  - Let V be vocabulary, t_i be a type in V. Let $\phi_k(t_i)$ be one contextual embedding of t_i, we define ***inter-type*** and ***intra-type*** cosine similarity as:
  - $S_{inter} = E_{i \neq j}[\cos\left(\phi(t_i), \phi\left(t_j\right)\right)]$
  - $S_{intra} = E_i[E_{k \neq l}[\cos(\phi_k(t_i), \phi_l(t_i))]]$
  - ***inter-type*** metric describes the similarity between different types, the ***intra-type*** one measures similarity between same type's embedding instances.
- Models
  - BERT, DistilBERT, GPT, GPT2 and ELMo.
- Dataset
  - PTB and WikiText-2.

Isotropy in the Contextual Embedding Space: Clusters and Manifolds

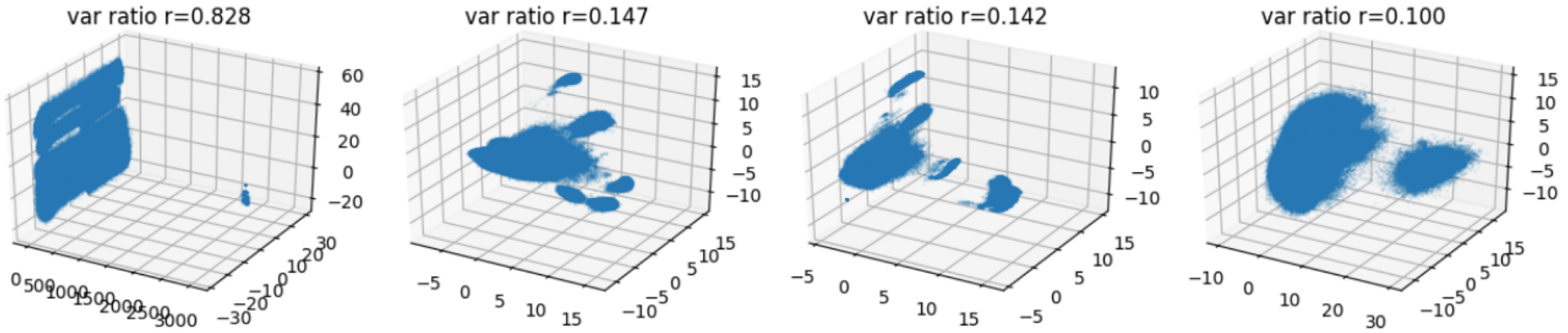
- Initial Look at Anisotropy
  - Both S-inter and S-intra are high across all the layers and all the models, indicating strong anisotropy.





Isotropy in the Contextual Embedding Space: Clusters and Manifolds

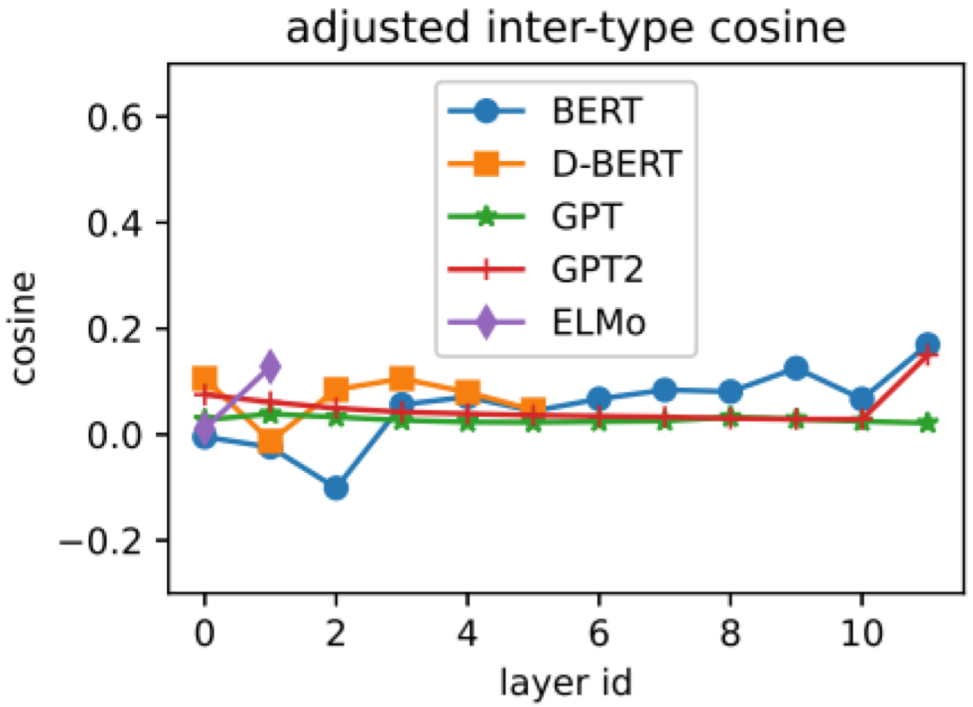
- Isolated Clusters in Space
  - The space is dominated by isolated clusters

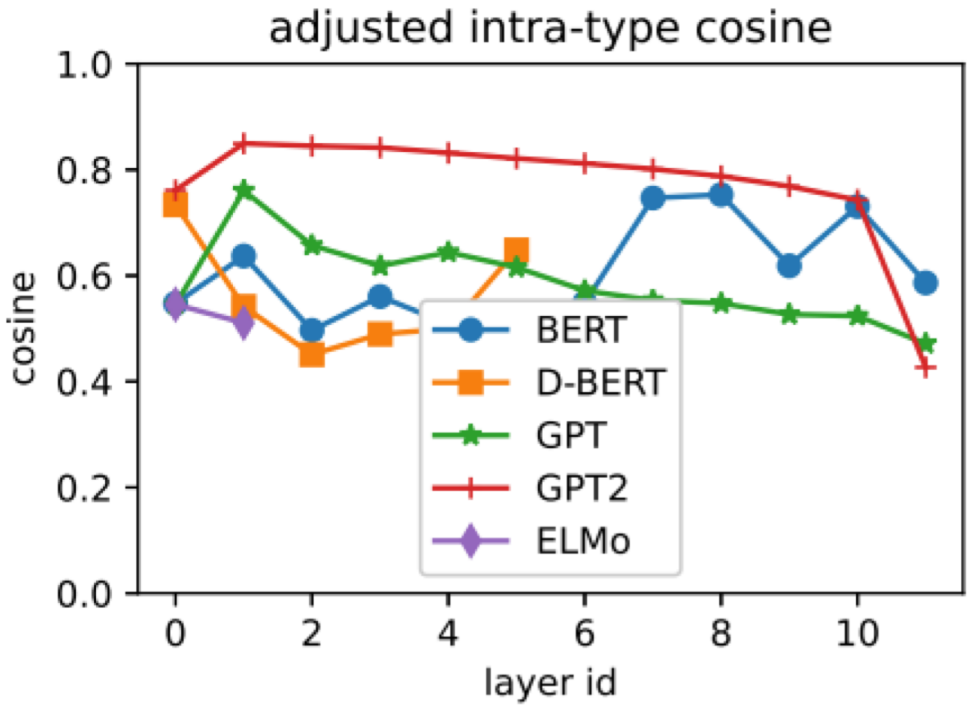


(a) GPT2 layer 6 (b) BERT layer 6 (c) D-BERT layer 1 (d) ELMo layer 1

Isotropy in the Contextual Embedding Space: Clusters and Manifolds

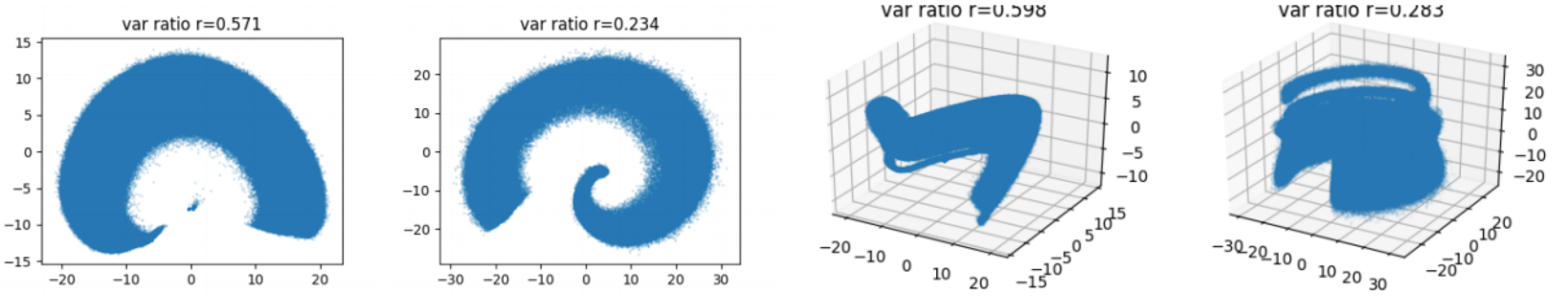
- To reveal isotropy, we do the following:
  1. Perform clustering using K-Means, to isolate different clusters.
  2. Within each cluster, we subtract the mean value from the embedding vectors.
  3. Finally, we recalculate the adjusted S-inter and S-intra.



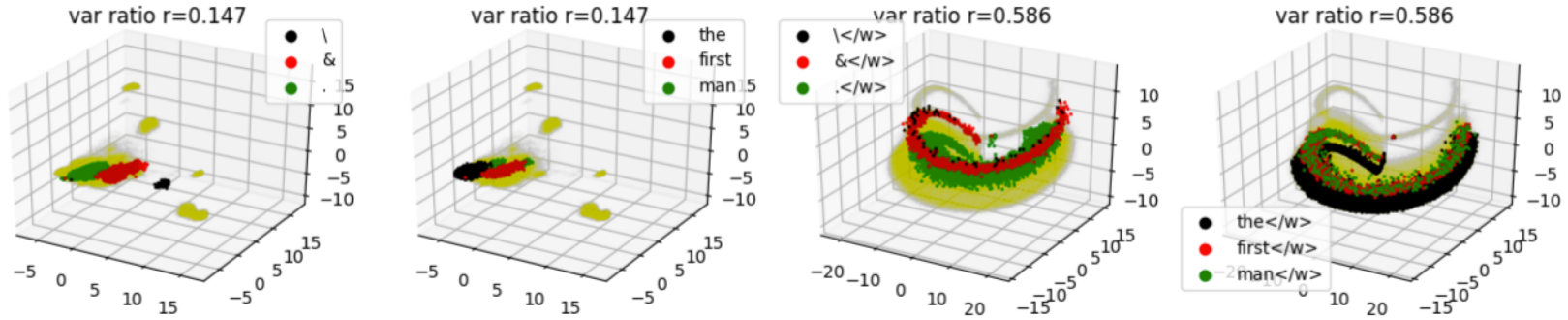


Isotropy in the Contextual Embedding Space: Clusters and Manifolds

- Swiss Roll Manifold of GPT/GPT2



(a) GPT layer 2 (b) GPT2 layer 2 (c) GPT layer 2 3-D view (d) GPT2 layer 2 3-D view



(a) BERT, symbol tokens (b) BERT, word tokens (c) GPT, symbol tokens (d) GPT, word tokens

Isotropy in the Contextual Embedding Space: Clusters and Manifolds

- Manifold Local Intrinsic Dimension

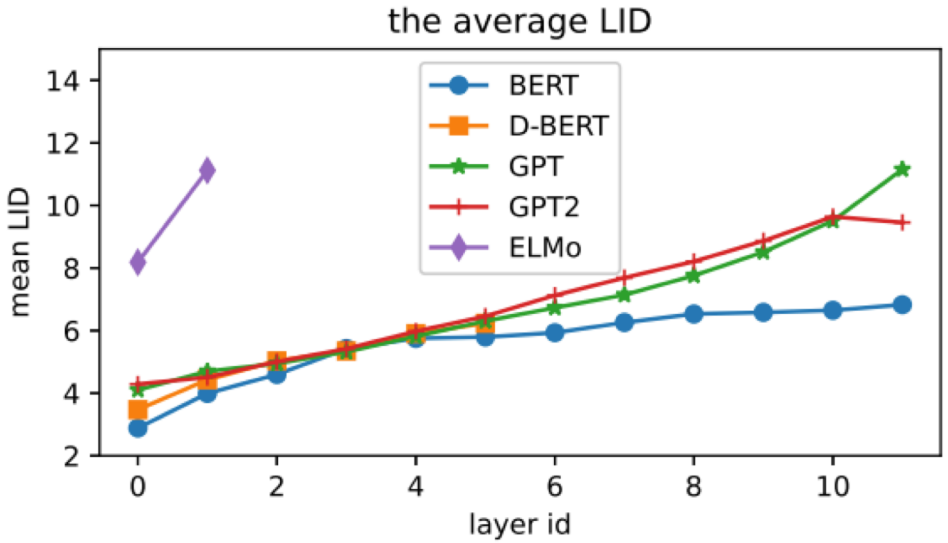


Figure 9: The average LID using Euclidean distance. ELMo's original embdding dimension is 1024, larger than other models' 768.

Table 3: A comparison of LIDs (using cosine similarity) among contextual and static embedding spaces.

	Model	n	m	avg LID
Contxt Embeds	BERT	1.19 M	768	5.6
	D-BERT	1.19 M	768	7.3
	GPT	0.96 M	768	6.8
	GPT2	1.09 M	768	7.0
	ELMo	0.88 M	1024	9.1
Static Embeds	GloVe	1.18 M	100	18.0
	GloVe-2M	2.20 M	300	26.1
	GNEWS	3.00 M	300	21.1

Isotropy in the Contextual Embedding Space: Clusters and Manifolds

- Conclusions
  - *We suggest that the anisotropy in contextual embedding space is a global view, being largely misled by distinct clusters resided in the space.*
  - *Our analysis show that it is more constructive to isolate the clusters and transform the space to reveal the isotropy. From this view, within the clusters, the spaces of different models all have nearly perfect isotropy that could explain the large model capacity.*
  - *Our visualization demonstrates a low-dimensional Swiss Roll manifold for GPT and GPT2 embeddings, that has not been reported before. The tokens and word frequencies are presented to qualitatively show the manifold structure.*
  - *We propose to use the approximate LID to quantitatively measure the local subspace, and compared with static embedding spaces. The results show smaller LID values for the contextual embedding models, which can be seen as a local anisotropy in the space.*
  - *We hope this research could help the design of new language models with better interpretability.*

Thank You