Isotropy in the Contextual Embedding Space: Clusters and Manifolds

Xingyu Cai, Jiaji Huang

Yuchen Bian, Kenneth Church

Paper Summary

- Anisotropy in contextual embedding space is largely misled by isolated clusters in the space.
- We suggest clustering and look at within-cluster embeddings to reveal isotropy
- We found low-dimensional manifolds exist in contextual embedding spaces, and calculate their Local Intrinsic Dimension (LID). Their intrinsic dimension is much smaller than original embedding dimension
- The above findings could lead to better language model design, e.g. dimension reductions, and better interpretability.

Notation, Models and Dataset

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:

$$S_{inter} = E_{i\neq j}[\cos\left(\phi(t_i), \phi(t_j)\right)]$$

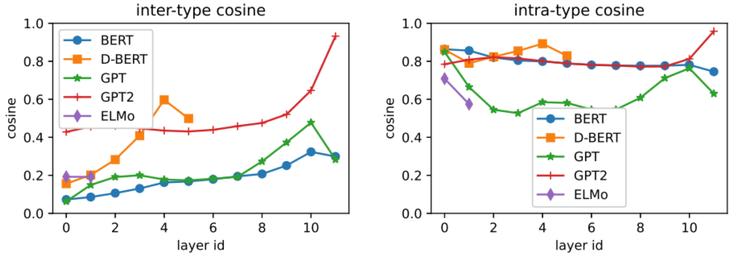
$$S_{intra} = E_i[E_{k\neq l}[\cos(\phi_k(t_i), \phi_l(t_i))]]$$

inter-type metric describes the similarity between different types, the ***intra-type*** one measures similarity between same type's embedding instances.

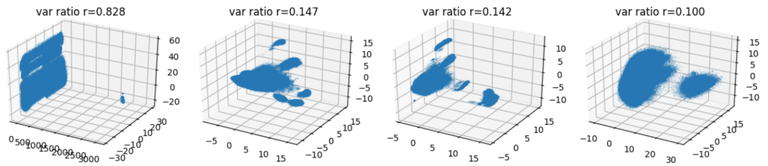
We investigate BERT, DistilBERT, GPT, GPT2 and ELMo models. Their contextual embedding in different layers are studied. We use PTB and WikiText-2 corpus to generate embeddings.

Anisotropy by Global View

- Both S-inter and S-intra are high across all the layers and all the models.
- S-inter tends to increase with layer, in contrast with S-intra which in general decreases but with fluctuations.



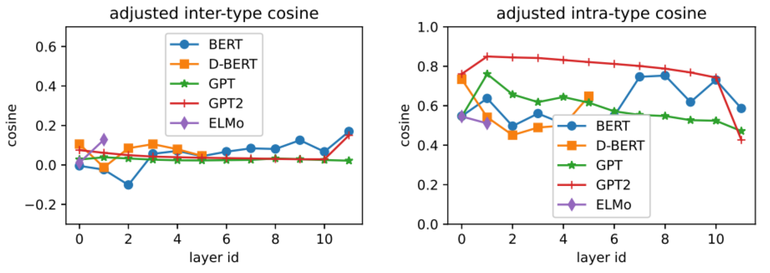
Disconnected Clusters Dominate Space



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

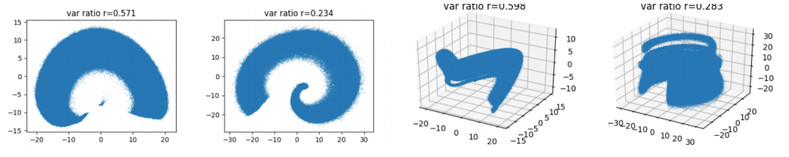
Isotropy Exists within Clusters

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.

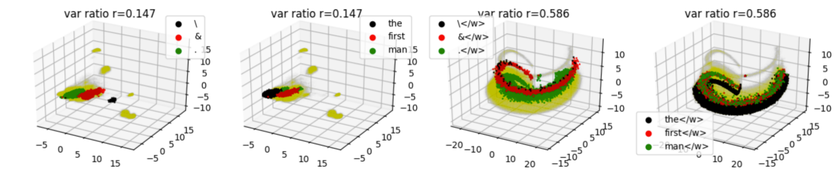


- For the adjusted inter-type cosine (the left plot), all models are having consistent near-zero S-inter. This means nearly perfect isotropy exists within each cluster, in each layer of all the models. The last layer of GPT2 and BERT has slightly worse isotropic behavior, nevertheless, general inter-type isotropy stays across all layers. This reveals the distinguishable embedding vectors.
- The general decreasing trend of intra-type cosine (the right plot) shows that the multiple instances for the same type/word, is slowly spreading over the layers. This is consistent with the un-centered intra-type cosine shown before.

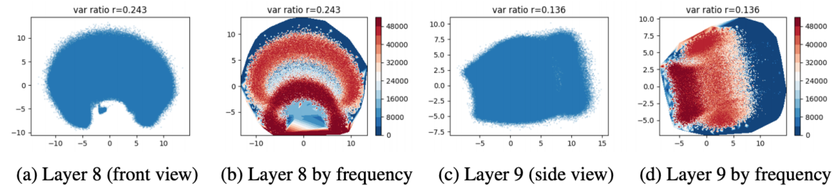
Swiss Roll Manifold of GPT/GPT2



(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



(a) Layer 8 (front view) (b) Layer 8 by frequency (c) Layer 9 (side view) (d) Layer 9 by frequency

- GPT has manifold structure, s.t. vectors are along the spiral band. BERT's space is closer to a Euclidean space as similar vectors are in concentrated clusters.
- Word frequency heatmap in GPT layer 8 and 9 are shown above. Red is high frequency, blue is low.

Low-dimensional Manifolds

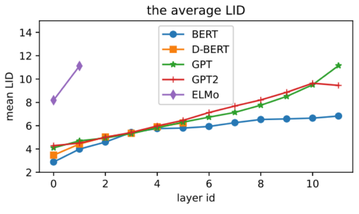


Figure 9: The average LID using Euclidean distance. ELMo's original embdding dimension is 1024, larger than other models' 768.

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

- The mean LIDs for all the models in all the layers are below 12, indicating the manifold is low dimensional. It is much smaller than original dimension (768 or 1024).
- The LID increases in deeper layers. As layer goes deeper, each token embedding is collecting information from context by adding their embeddings (and non-linear transforms concatenated). This could explain the spreading / expanding of the local subspace, and therefore the LID increases.
- The static embedding spaces generally have higher LID than the contextual ones. This means that the data points are more isotropic in the static embeddings, possibly due to their large vocabulary.