IUST
Iran University of
Science and Technology

Isotropy In Embedding Space

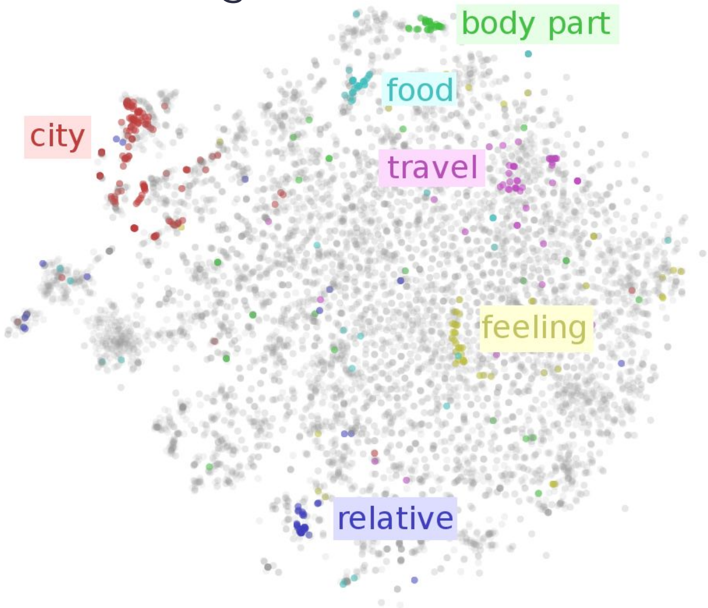
By Sara Rajaee
May, 2021

Tehran Institute for Advanced Studies

Contextual Word Embeddings

➔ Considering context in word representations

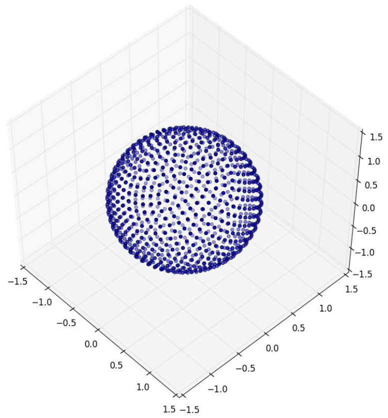
➔ Carrying different semantic and syntactic knowledge



1. https://ruder.io/word-embeddings-1/

Isotropy

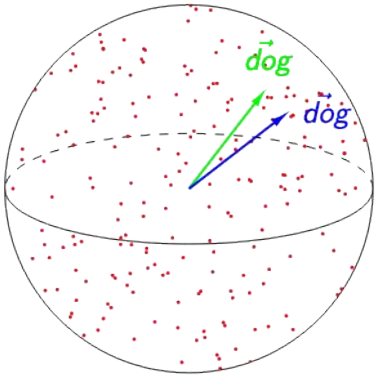
- ➔ Uniform distribution of data points (e.g. word embeddings)
- ➔ Equal elongations respect to different directions

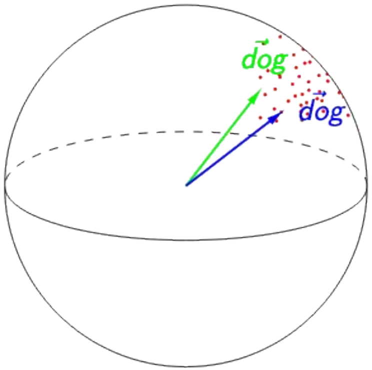


Why Isotropy is Important?

In anisotropic embedding space:

- ➔ Randomly sampled words have high cosine similarity
- ➔ Longer convergence time
- ➔ Word representations power is limited





http://ai.stanford.edu/blog/contextual/

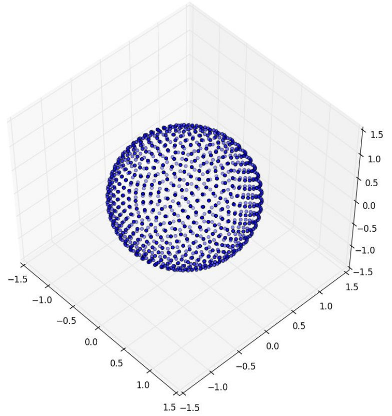
Quantizing Isotropy

- ➔ Using Principal Components (PCs) to find **dominant directions**
- ➔ The more be isotropic the embedding space, I(W) is closer to one

$$I(\mathcal{W}) = \frac{min_{u \in U} Z(u)}{max_{u \in U} Z(u)}$$

where

$$F(u) = \sum_{i=1}^{N} e^{u^T w_i}$$



01

Isotropy in pre-trained LM models

Selected pre-trained models

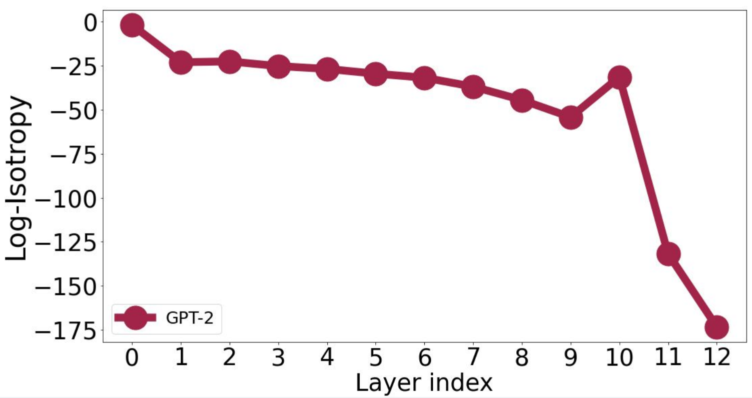
01
GPT-2
Unidirectional Language Model

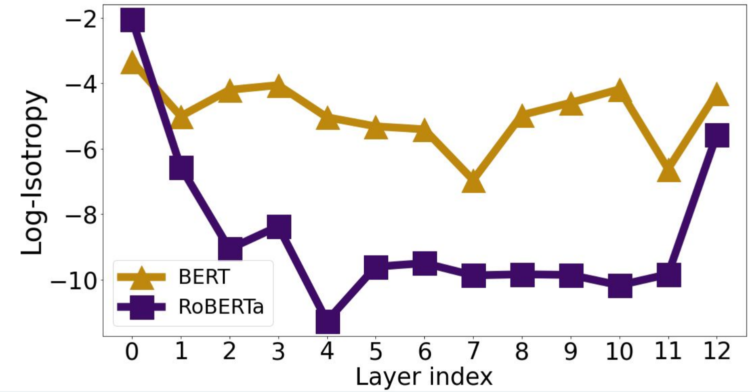
02
BERT
Bidirectional Masked Language model

03
RoBERTa
Bidirectional + additional training data

Global Approach

- Isotropy of GPT-2 consistently decreases in upper layers.
- The last layer is the most isotropic layer in BERT and RoBERTa





Local Approach

- Local approach shows BERT and RoBERTa are almost isotropic in a local view
- GPT-2 is still extremely **anisotropic**

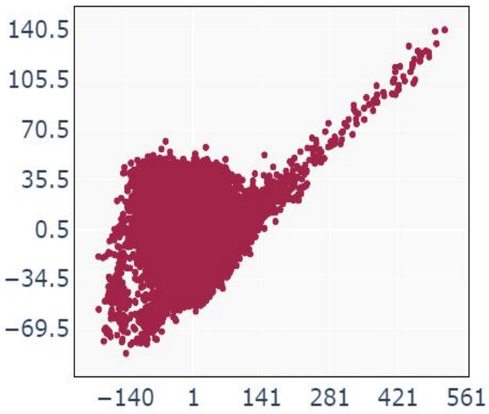
	GPT-2	BERT	RoBERTa
Baseline	5.02E-174	5.05E-05	2.70E-06
$k = 1$	2.49E-220	0.010	0.015
$k = 3$	9.42E-66	0.040	0.290
$k = 6$	**1.40E-41**	0.125	0.453
$k = 9$	1.18E-41	0.131	0.545
$k = 20$	4.06E-47	**0.262**	**0.603**

Table 2: CWRs isotropy after clustering and making zero-mean each cluster separately. The results are reported for the different number of clusters (k) on STS-B dev set.

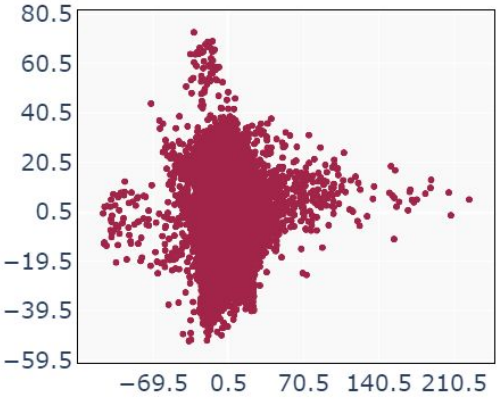
Cosine Similarity

A superficial alternative

- Many research use cosine similarity as a measurement for isotropy
- **Exceptional** cases where cosine similarity does **not** work (near zero cosine similarity in anisotropic space)



(a)



(b)

Geometry of GPT-2 embedding space a)before b)after locally making zero-mean on STS-B dev set

02

A local approach for improving isotropy

3-Step method

Clustering

Apply k-means clustering to word representations

Zero-mean

Making zero-mean each cluster separately

Remove Dominant Directions

Using PCA to find dominant directions

03

Experiments

Target Tasks

- Semantic Textual Similarity(STS)
- Recognizing Textual Entailment(RTE)
- The Corpus of Linguistic Acceptability(CoLA)
- The Stanford Sentiment Treebank(SST-2)
- The Microsoft Research Paraphrase Corpus(MRPC)
- Word-in-Context(WiC)
- Boolean Questions(BoolQ)

Settings

Regression task

Use cosine similarity of sentence embeddings as score

Classification task

Train an MLP on top of BERT, while its weights are frozen

Semantic Textual Similarity

	Model	STS 2012	STS 2013	STS 2014	STS 2015	STS 2016	SICK-R	STS-B
Baseline	GPT-2	26.49	30.25	35.74	41.25	46.40	45.05	24.8
	BERT	42.87	59.21	59.75	62.85	63.74	58.69	47.4
	RoBERTa	33.09	56.44	46.76	55.44	60.88	61.28	56.0
Global method	GPT-2	51.42	69.71	55.91	60.35	62.12	59.22	55.7
	BERT	53.66	68.66	60.34	63.73	69.47	63.64	65.1
	RoBERTa	51.48	71.20	59.64	66.72	68.14	65.44	67.7
Our approach	GPT-2	**52.40**	**72.71**	**59.23**	**62.19**	**64.26**	**59.51**	**62.3**
	BERT	**58.34**	**75.65**	**63.55**	**64.37**	**69.63**	**63.75**	**66.0**
	RoBERTa	**54.87**	**76.70**	**64.18**	**67.05**	**69.28**	**66.93**	**71.4**

Table 2: Performance of pre-trained models (baseline), after the global method, and after our local cluster-based approach on different datasets in the STS benchmark, according to Spearman's ρ correlation percentage.

Classification Tasks

	RTE	CoLA	SST-2	MRPC	WiC	BoolQ	Average
Baseline	54.44	38.0	81.4	70.26	60.07	64.7	61.47
Our approach	**56.5**	**40.7**	**82.5**	**72.41**	**61.07**	**66.4**	**63.26**

Table 4: Performance of our proposed method compared to CWRs (Baseline) using pre-trained BERT on different classification tasks. Numbers are reported based on Matthew's correlation for CoLA and accuracy for the rest of them.

04

Analyses

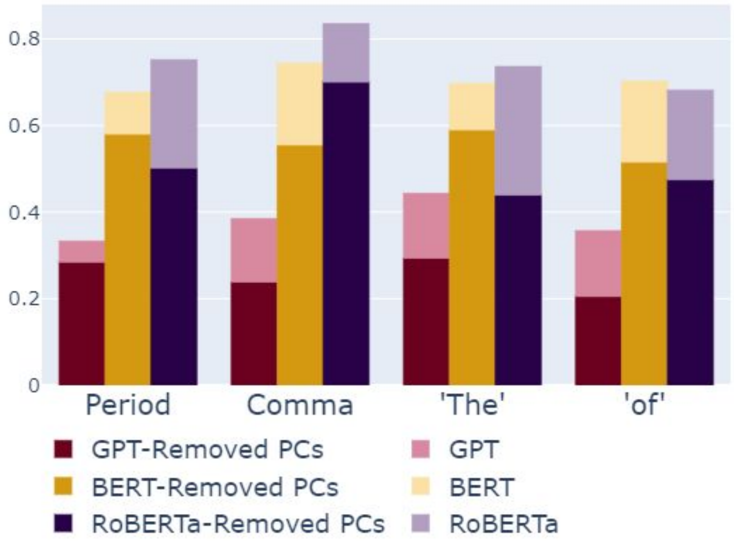
Linguistic Knowledge

Punctuations and Stop Words

local dominant directions carry **structural** and **syntactical** information about the sentences they appear.

- ★ *A man is crying.*
- ★ *A woman is dancing.*

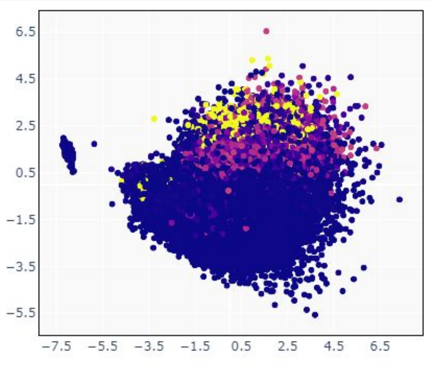
- Use a dataset consists of groups in which sentences are structurally and syntactically similar but have no semantic similarity.
- pick 200 different structural groups
- Find the percentage of each representation's nearest neighbors that are in a same group



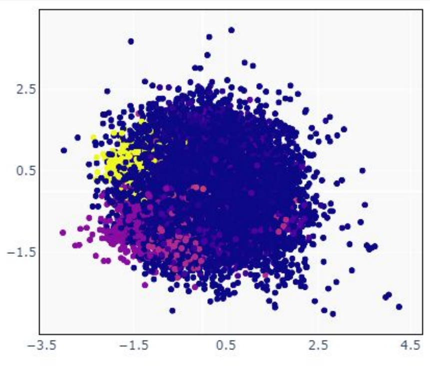
Linguistic Knowledge

Word frequency

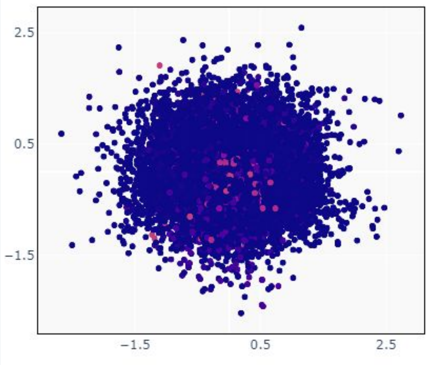
- ★ CWRs are biased toward their frequency
- ★ Parts of removed PCs encode frequency information
- ★ The proposed method can overcome frequency bias



Baseline



Global method



Our proposed method

RoBERTa's CWRs visualization using PCA on STS-B dev set. Points color indicates their frequency calculated based on Wikipedia dump; the lighter point, the more frequent.

Linguistic Knowledge

Verb Tense

- verb representations are distributed based on their **tense**, not their semantic similarity
- Using SemCor, for a randomly sampled verb's representation, we calculate its distance to other representations in three categories, including:
  - Representations with the same tense and same meaning
  - Representations with the same tense but different meaning
  - Representations with different tense and the same meaning.

Verb Tense

	Base line				Removed PCs			
Model	**ST - SM**	**ST - DM**	**DT - SM**	**Isotropy**	**ST - SM**	**ST - DM**	**DT - SM**	**Isotropy**
GPT-2	48.82	48.19	50.86	2.26E-05	9.32	9.53	9.49	0.172
BERT	13.44	14.24	14.87	2.24E-05	10.31	10.50	10.32	0.319
RoBERTa	5.89	6.31	6.86	1.22E-06	4.78	5.00	4.89	0.73

Table 5: The mean l_2-norm for randomly sampled verbs. For each sampled verb, the mean of its distance is calculated to all other verbs that have the Same Tense and Same Meaning (ST-SM), Same Tense and Different Meaning (ST-DM), and Different Tense and Same Meaning (DT-SM).

Convergence time

★ Isotropy decreases the convergence time

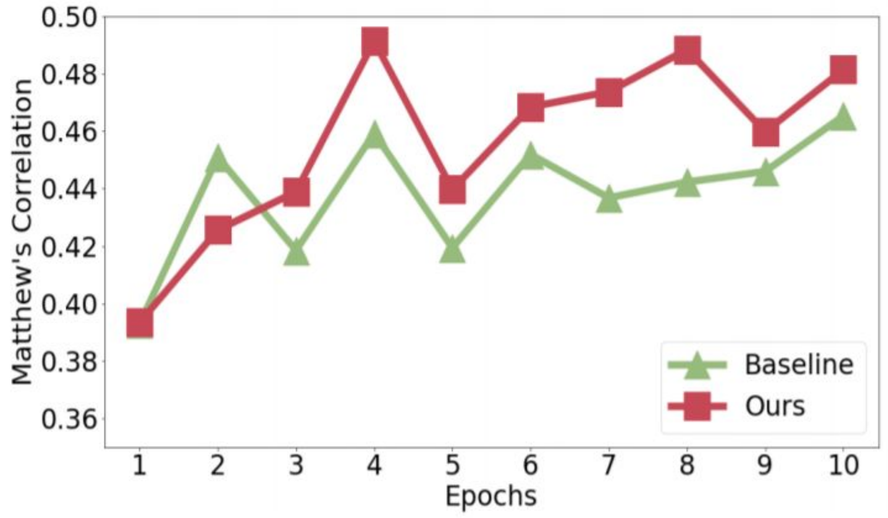


Figure 4: Convergence time in CoLA.

05

Analyzing the effect of fine-tuning on isotropy of embedding space

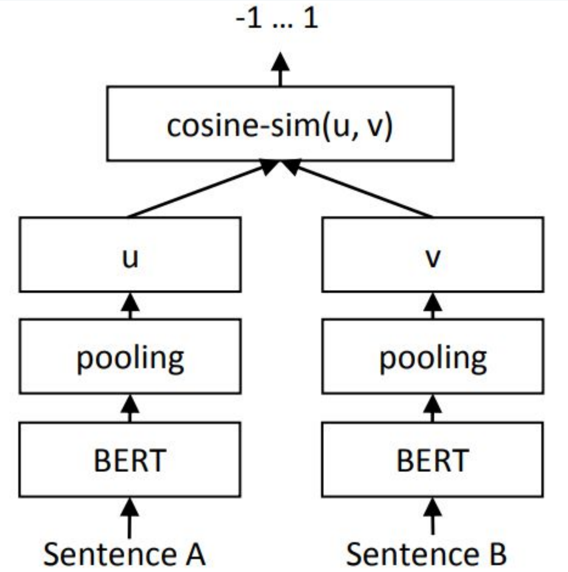
Fine-tuning

- ★ Adding a simple classification layer on top of the pre-trained model
- ★ Training the pre-trained layers and the classifier **jointly**
- ★ Significantly improves the performance

Questions?

- ★ Can we attribute the enhancement achieved during fine-tuning to improving isotropy?
- ★ Increasing isotropy can lead to further improvement?

Methodology



Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks

Target Task

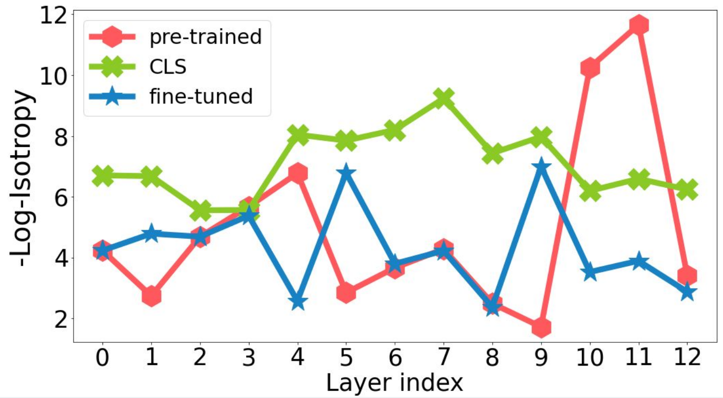
Semantic Textual Similarity

Setups

- ★ BERT base
- ★ Fine-tuning with CLS
- ★ Fine-tuning with mean-pooling using Siames Network

Evaluating Isotropy

★ Fine-tuned BERT embedding space is still extremely anisotropic
★ the [CLS] tokens are much more anisotropic compared to representations



Improving Isotropy

Settings

Baseline

Fine-tuned CWRs

Zero-mean

Making all CWRs zero-mean

Clustering+ZM

After clustering, making each cluster zero-mean separately

Global Approach

Removing dominant directions globally

Local Approach*

After clustering, eliminating dominant directions in each cluster

Results

1. Increasing Isotropy in the fine-tuned embedding space **hurts** the performance.

	Baseline	Zero-mean	Clustering+ZM	Global Approach	Local Approach
			Performance		
pre-trained	54.09	57.00	64.25	71.40	75.29
fine-tuned	85.70	85.76	80.66	82.86	63.06
			Isotropy		
pre-trained	1.35E-5	2.16E-6	0.23	0.29	0.67
fine-tuned	1.05E-3	5.08E-4	0.04	0.10	0.46

Pre-trained and fine-tuned CWRs performance on STS-B dev set. Performance results are base on Spearman Correlation. Isotropy is calculated using I(W).

Results

2. the clustered structure of the embedding space changes during fine-tuning.

	Baseline	Zero-mean	Clustering+ZM	Global Approach	Local Approach
			Performance		
pre-trained	54.09	57.00	64.25	71.40	75.29
fine-tuned	85.70	85.76	80.66	82.86	63.06
			Isotropy		
pre-trained	1.35E-5	2.16E-6	0.23	0.29	0.67
fine-tuned	1.05E-3	5.08E-4	0.04	0.10	0.46

Pre-trained and fine-tuned CWRs performance on STS-B dev set. Performance results are base on Spearman Correlation. Isotropy is calculated using I(W).

Results

3. The number of harsh dominant directions significantly increases during fine-tunning

	Baseline	Zero-mean	Clustering+ZM	Global Approach	Local Approach
			Performance		
pre-trained	54.09	57.00	64.25	71.40	75.29
fine-tuned	85.70	85.76	80.66	82.86	63.06
			Isotropy		
pre-trained	1.35E-5	2.16E-6	0.23	0.29	0.67
fine-tuned	1.05E-3	5.08E-4	0.04	0.10	0.46

Pre-trained and fine-tuned CWRs performance on STS-B dev set. Performance results are base on Spearman Correlation. Isotropy is calculated using I(W).

06

Isotropy in Multilingual Embedding Space

Multilingual models

mBERT

- ★ A **single** language model pre-trained on the concatenation of monolingual Wikipedia corpora from **104** languages without any supervision

Settings

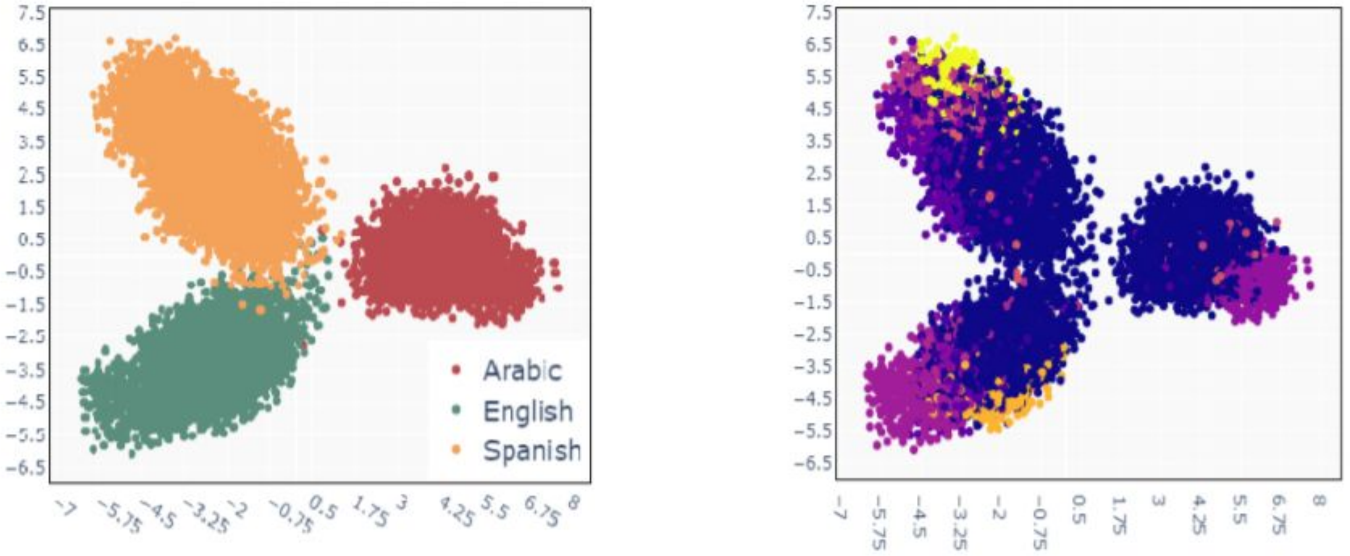
- ★ Considering Arabic, English, and Spanish
- ★ Multilingual STS as the target task (cross- and mono-lingual tracks)

Probing Isotropy

	Arabic	English	Spanish
mBERT	6.21E-03	3.91E-04	1.27E-04

Table 1: CWRs isotropy for Arabic, English, and Spanish on multilingual STS, reporting based on $I(W)$.

Embedding Distribution



(a) Distribution

(b) Word Frequency

a Brief Conclusion

Anisotropic embeddings

Embedding spaces are highly anisotropic in all languages

Elongation from center

All clusters start from the origin of the shared embedding space

Similar structure

In the frequency case, different languages have similar structures.

Questions?

★ Do different languages encode similar linguistic knowledge in their dominant directions?

★ Can improving isotropy lead to performance enhancement in a multilingual setting?

Results

	Ar-Ar	Ar-En	Es-Es	Es-En	Es-En-WMT	En-En
Baseline	46.93	14.63	63.92	20.21	20.21	58.94
Individual	61.00	35.14	73.56	46.86	13.95	70.50
Zero-shot	55.83	34.65	70.15	48.24	16.27	-

Table 2: Spearman's rank correlation ρ between cosine similarity of sentence embeddings and gold labels on multi- and cross-lingual STS datasets using mBERT. The performance is reported as $\rho \times 100$. Applying the cluster-based method can improve the performance on the multi- and cross-lingual datasets in both Individual and Zero-shot settings.

Results

	Ar-Ar	Ar-En	Es-Es	Es-En	Es-En-WMT	En-En
Baseline	6.21E-03	4.38E-04	1.27E-04	6.66E-05	2.21E-03	3.91E-04
Individual	0.820	0.833	0.834	0.841	0.752	0.893
Zero-shot	0.015	0.197	0.084	0.153	0.020	0.893

Table 3: Isotropy of CWRs on multi- and cross-lingual STS datasets calculating based on $I(W)$; higher value more isotropic embedding space.

Wrapping up

We showed:

- ★ Pre-trained LM models are **highly anisotropic.**
- ★ Cosine similarity is an **inappropriate** metric for isotropy.
- ★ Our local approach can consistently improve performance on different tasks.
- ★ Removing **tense-based distribution** of verbs, **structural knowledge** encoded in punctuations and stop words, and **frequency bias** are parts of our approach's results.
- ★ Isotropic embedding space decreases convergence time in deep models.

Wrapping up

We showed:

★ Fine-tuning does not improve isotropy
★ Clustered structure of CWRs changed from pre-training to fine-tuning
★ Essential knowledge has been encoded in dominant directions
★ The number of dominant directions has been increased during fine-tuning

Wrapping up

We showed:

★ Multilingual CWRs lack isotropy
★ Distribution of CWRs is similar in different languages
★ Dominant directions encode similar knowledge
★ Our local approach can improve CWRs performance in Individual and Zero-shot settings.

Thanks
for your attention

Questions?