Iran University of Science & Technology

Isotropicity of Semantic Spaces

Sara Rajaee

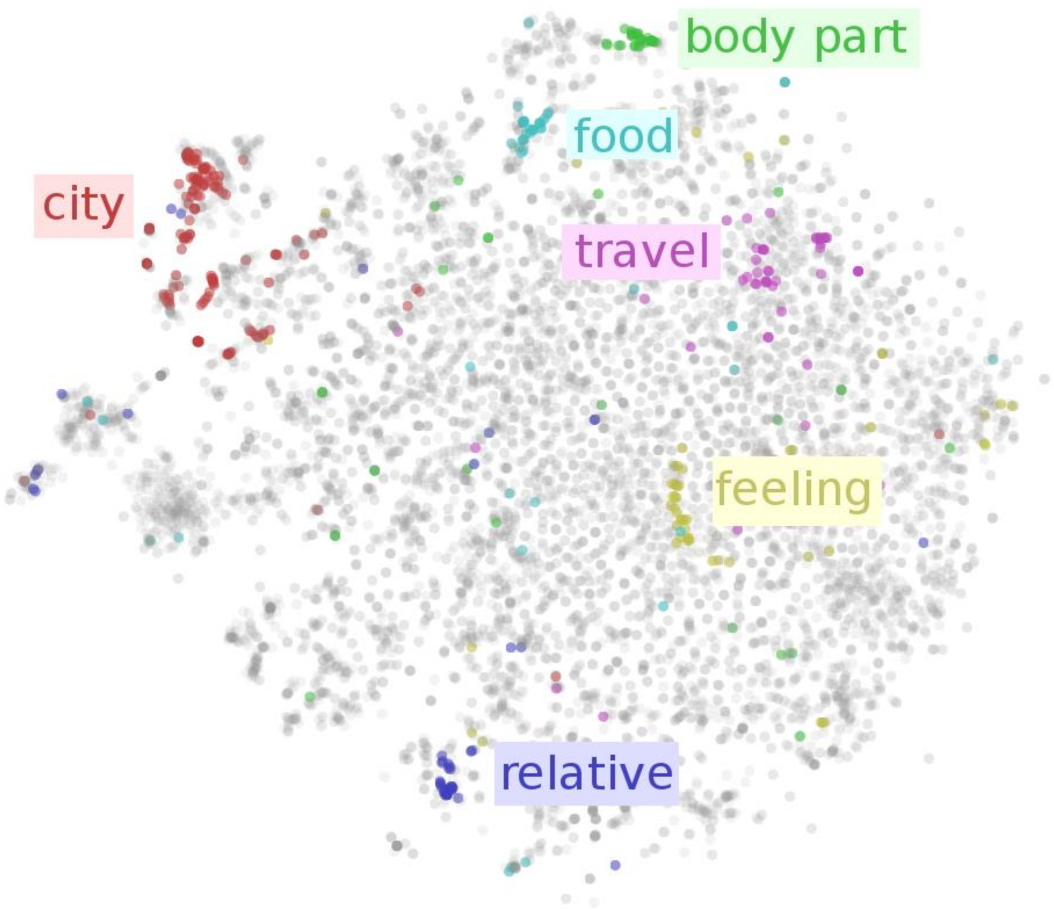
May 2022

Tehran Institute for Advanced Studies

Contextual Embedding Space

Representing linguistic knowledge

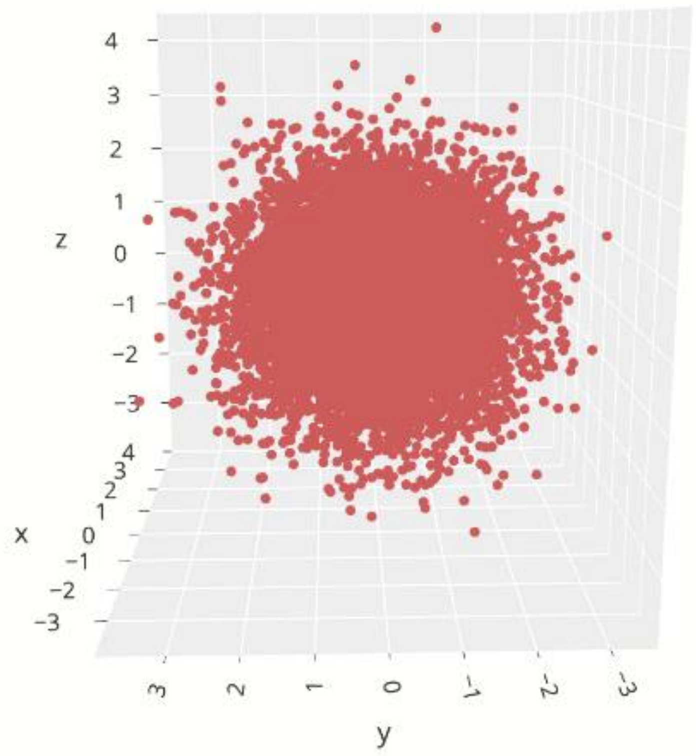
Based on context, CWRs encode different level of linguistic knowledge.



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

Isotropy

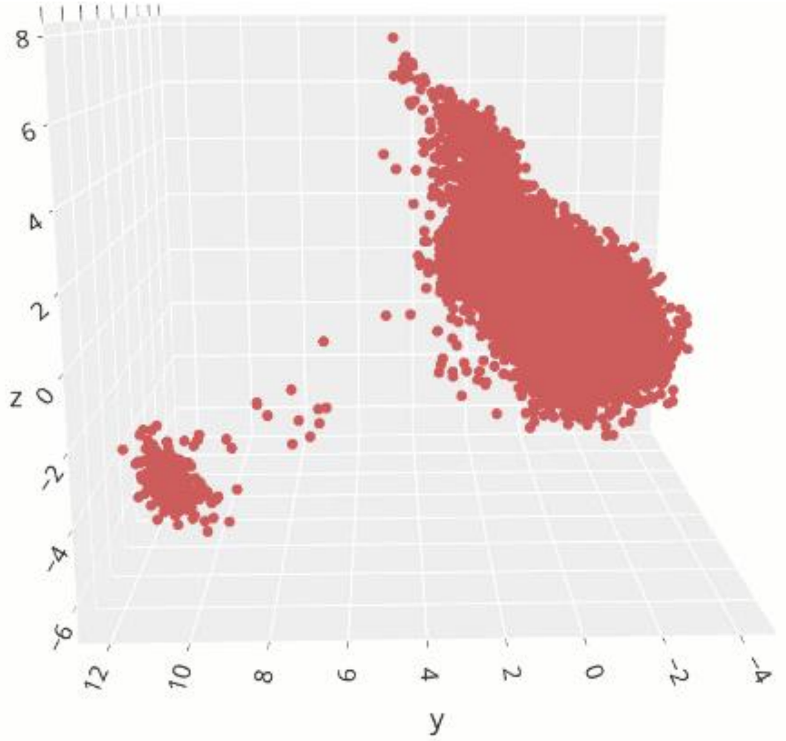
- Uniform distribution
- Equal elongations concerning different directions



Why Isotropy is Important?

In anisotropic embedding space:

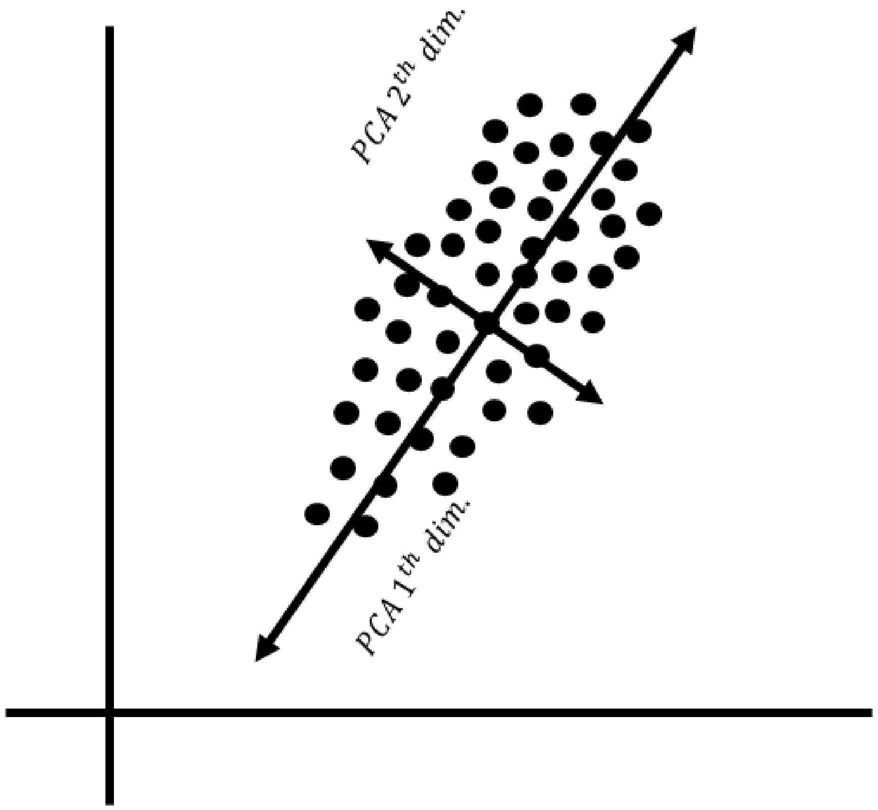
- High cosine similarity between random embeddings
- Word representations power is limited
- Longer convergence time



Measuring Isotropy

Review - PCA

PCs represent the directions explaining a maximal amount of variance.



Measuring Isotropy

PC-based metric

$W: Embedding\ matrix$
$w_i: i^{th}\ word's\ embedding$
$U = \{eignevectors\ of\ W^T W\}$

$$F(u) = \sum_{i=1}^{M} \exp(u^T w_i)$$

$$I_{PC}(W) = \frac{\min_{u \in U} F(u)}{\max_{u \in U} F(u)}$$

Measuring Isotropy

Cosine Similarity

Average cosine similarity between randomly sampled word embeddings.

$$I_{Cos}(W) = \frac{1}{N} \sum_{i=1, x_i \neq y_i}^{N} Cos(x_i, y_i)$$

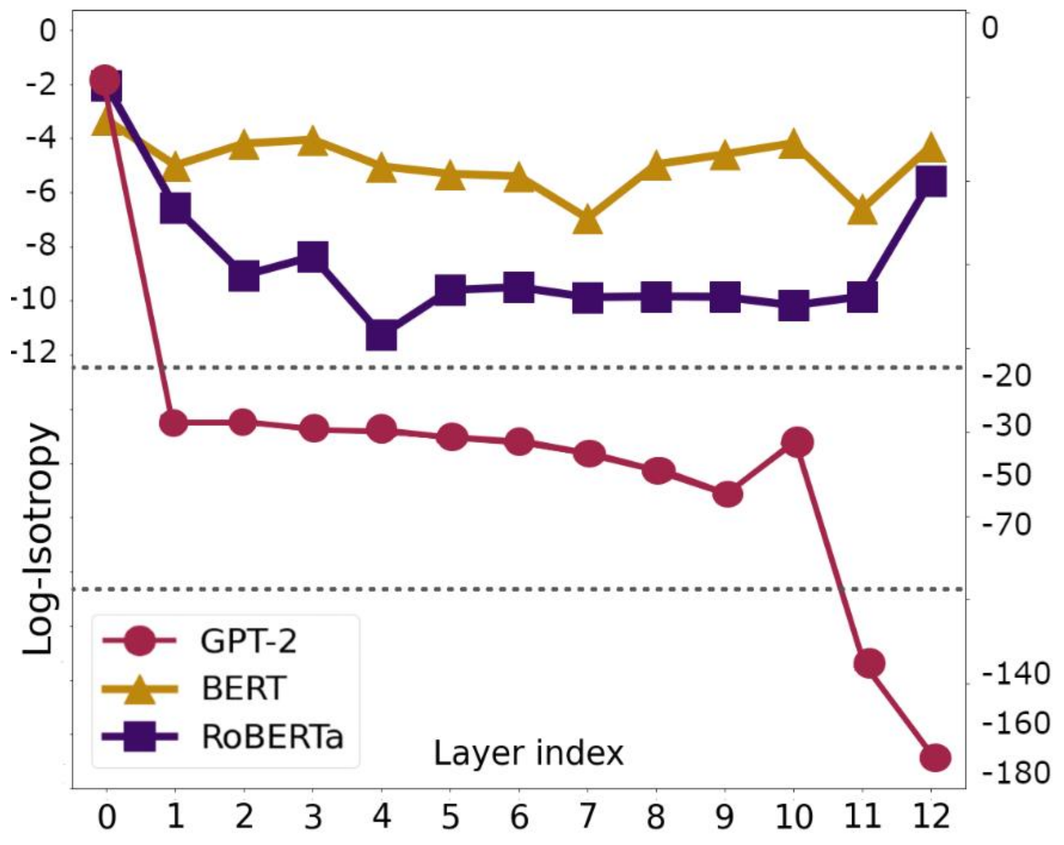
Language models

- GPT-2
  - Unidirectional language model
- BERT
  - Bidirectional encoders
  - Masked language modeling
- RoBERTa
  - More training data (compared to BERT)

Probing Isotropy

Global assessment

- All contextualized models are anisotropic in all layers
- GPT-2 has exceedingly anisotropic embedding space (except in the input layer)



Probing Isotropy

Local assessment

- Contextualized models are more isotropic.
- GPT-2 is still anisotropic.

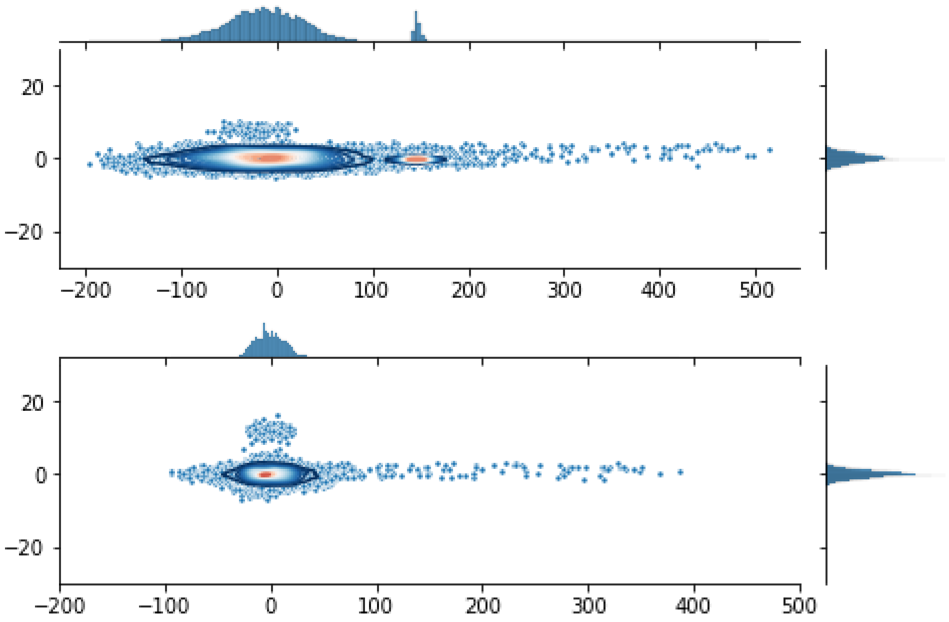
	GPT-2	**BERT**	**RoBERTa**
Baseline	5.02E-174	5.05E-05	2.70E-06
$k = 1$	2.49E-220	0.010	0.015
$k = 3$	9.42E-66	0.040	0.290
$k = 6$	**1.40E-41**	0.125	0.453
$k = 9$	1.18E-41	0.131	0.545
$k = 20$	4.06E-47	**0.262**	**0.603**

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.

Probing Isotropy

A superficial metric

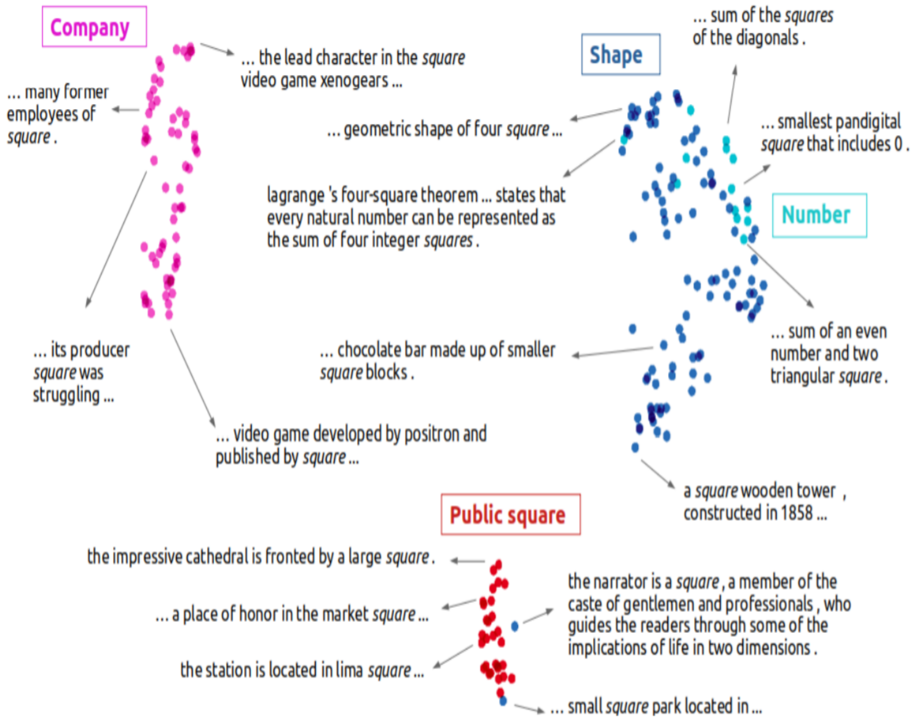
- **Exceptional** cases where cosine similarity does **not** work (near zero cosine similarity in anisotropic space)



Cluster-based approach

Motivation

- Clustered structure in contextual embedding space

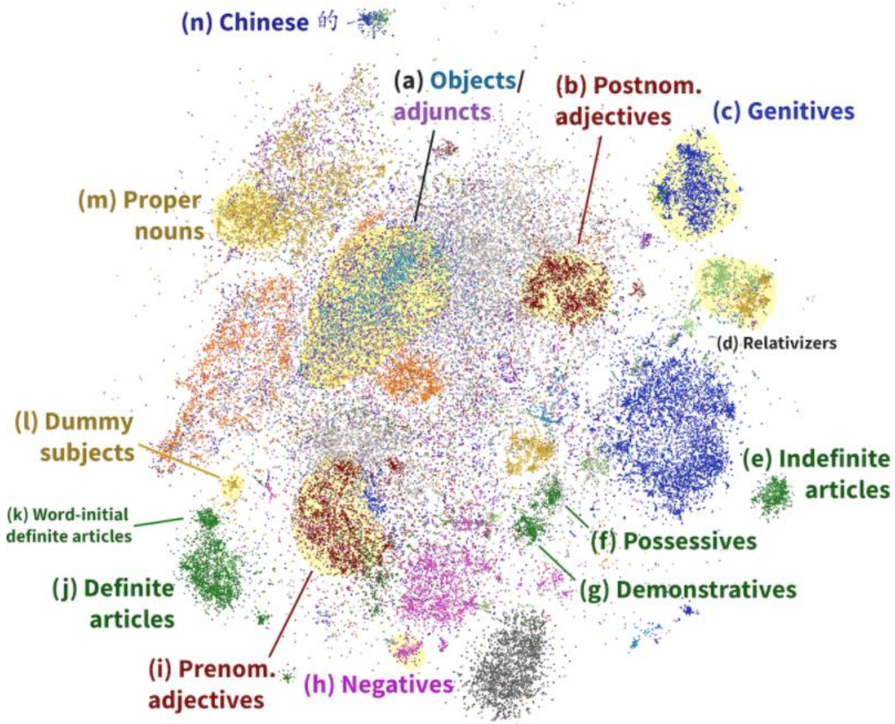


http://ethanachi.com/multilingual-probing.html
(Michael et al., 2020; Loureiro et al., 2021)

Cluster-based approach

Motivation

- Clustered structure in contextual embedding space

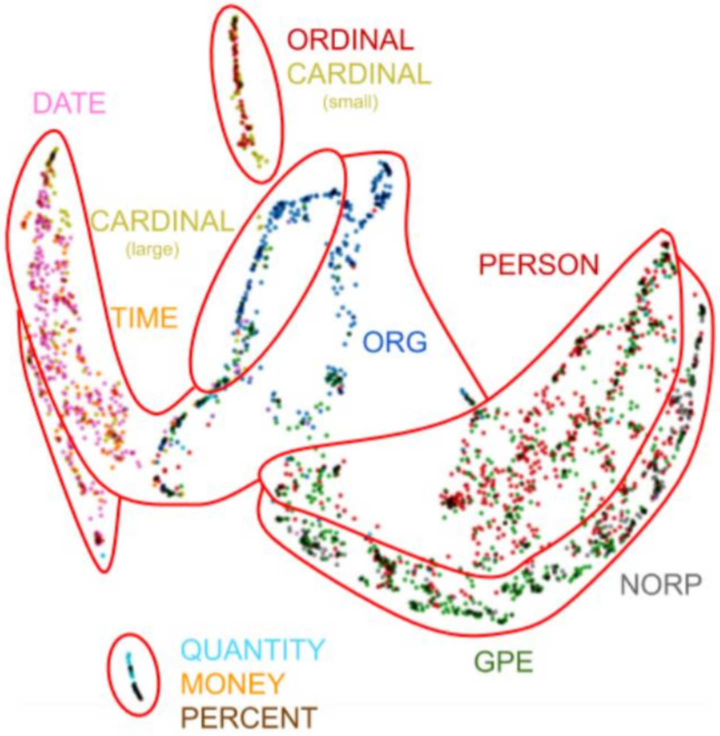


http://ethanachi.com/multilingual-probing.html
(Michael et al., 2020; Loureiro et al., 2021)

Cluster-based approach

Motivation

- Clustered structure in contextual embedding space

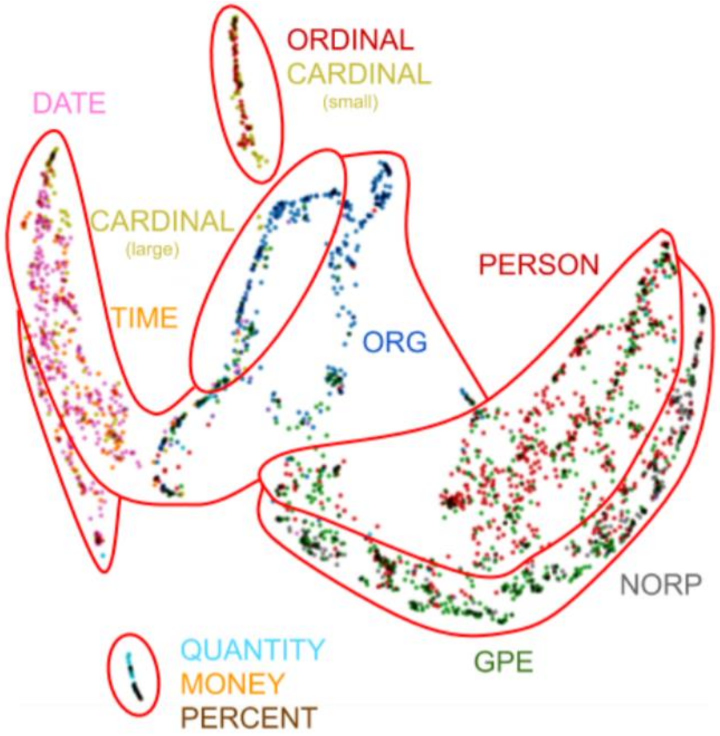


http://ethanachi.com/multilingual-probing.html
(Michael et al., 2020; Loureiro et al., 2021)

Cluster-based approach

Motivation

- Clustered structure in contextual embedding space
- Higher isotropy in the local view



http://ethanachi.com/multilingual-probing.html
(Michael et al., 2020; Loureiro et al., 2021)

Cluster-based approach

1

Clustering embeddings using k-means.

2

Making each cluster zero-mean separately

3

Removing few top dominant directions calculated using PCA in each cluster individually

Setups

- 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)

Target tasks

Setups

- Regression task
  - Using the cosine similarity of the sentence embeddings as score
- Classification task
  - Training an MLP on top of BERT, while its weights are frozen

Experiments

Model	STS 2012	STS 2013	STS 2014	STS 2015	STS 2016	SICK-R	STS-B
				Baseline			
GPT-2	1.4E-178	1.0E-170	1.4E-172	2.9E-177	6.0E-174	9.9E-140	2.6E-105
BERT	3.1E-05	1.9E-04	2.6E-04	3.7E-07	2.8E-04	4.2E-05	1.1E-04
RoBERTa	3.1E-06	3.1E-07	3.8E-06	3.8E-06	3.5E-06	3.7E-07	2.9E-06
				Global approach			
GPT-2	0.57	0.40	0.05	0.12	0.60	0.57	0.51
BERT	0.48	0.41	0.55	0.72	0.65	0.63	0.58
RoBERTa	0.67	0.87	0.87	0.84	0.85	0.90	0.88
				Cluster-based approach			
GPT-2	**0.71**	**0.74**	**0.47**	**0.74**	**0.74**	**0.78**	**0.70**
BERT	**0.68**	**0.61**	**0.77**	**0.81**	**0.75**	**0.82**	**0.73**
RoBERTa	**0.89**	**0.91**	**0.93**	**0.92**	**0.89**	**0.94**	**0.90**

Semantic Textual Similarity - Isotropy

Experiments

	Model	STS 2012	STS 2013	STS 2014	STS 2015	STS 2016	SICK-R	STS-B
	GPT-2	26.49	30.25	35.74	41.25	46.40	45.05	24.8
Baseline	BERT-base	42.87	59.21	59.75	62.85	63.74	58.69	47.4
	RoBERTa-base	33.09	56.44	46.76	55.44	60.88	61.28	56.0
	GPT-2	51.42	69.71	55.91	60.35	62.12	59.22	55.7
Global approach	BERT-base	54.62	70.39	60.34	63.73	69.37	63.68	65.5
	RoBERTa-base	51.59	73.57	60.70	66.72	**69.34**	65.82	70.1
	GPT-2	**52.40**	**72.71**	**59.23**	**62.19**	**64.26**	**59.51**	**62.3**
Cluster-based approach	BERT-base	**58.34**	**75.65**	**63.55**	**64.37**	**69.63**	**63.75**	**66.0**
	RoBERTa-base	**54.87**	**76.70**	**64.18**	**67.05**	69.28	**66.93**	**71.4**

Semantic Textual Similarity

Experiments

	RTE	CoLA	SST-2	MRPC	WiC	BoolQ	*Average*
Baseline	54.4	38.0	80.1	70.2	60.0	64.7	61.2
Global approach	56.2	38.8	80.2	72.1	60.7	64.9	62.1
Cluster-based approach	**56.5**	**40.7**	**82.5**	**72.4**	**61.0**	**66.4**	**63.2**

Classification tasks using BERT

Analysis

Investigating linguistic knowledge

Analysis

Punctuations and Stopwords

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

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

Analysis

Punctuations and Stopwords

- 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 the same group.

Analysis

Punctuations and Stopwords

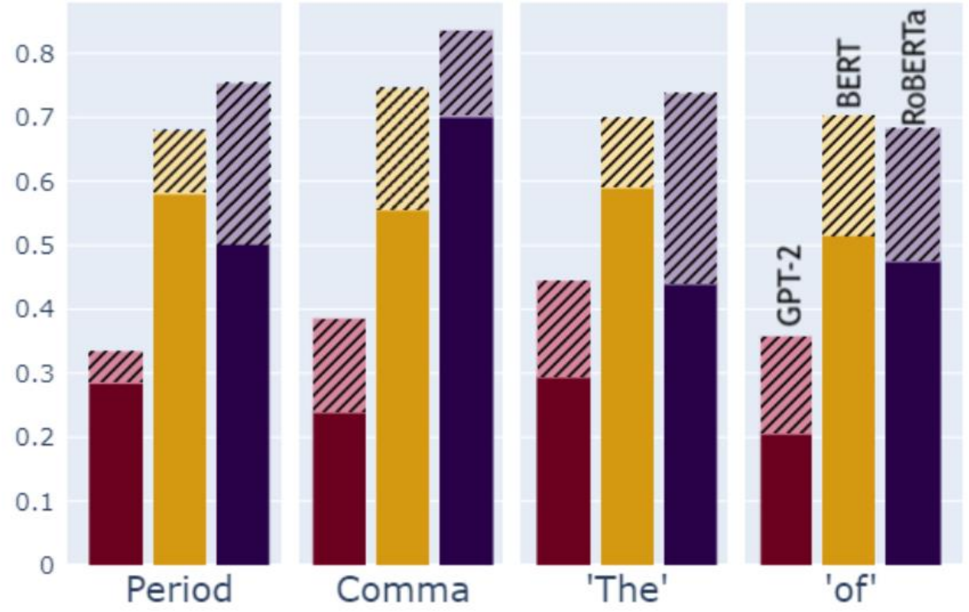
Original	shapley participated in the “ great debate ” with heber d .
1	morris put in the “ heroic speech ” with heber energy .
2	hall met in the “ ninth season ” with walton moore .
3	patel helped in the “ double coup ” with ibn salem .
4	chu sent in the “ universal text ” with u z .
5	smith exhibited in the “ red year ” with william james .

(Ravfogel et al., 2020)

Analysis

Punctuations and Stopwords

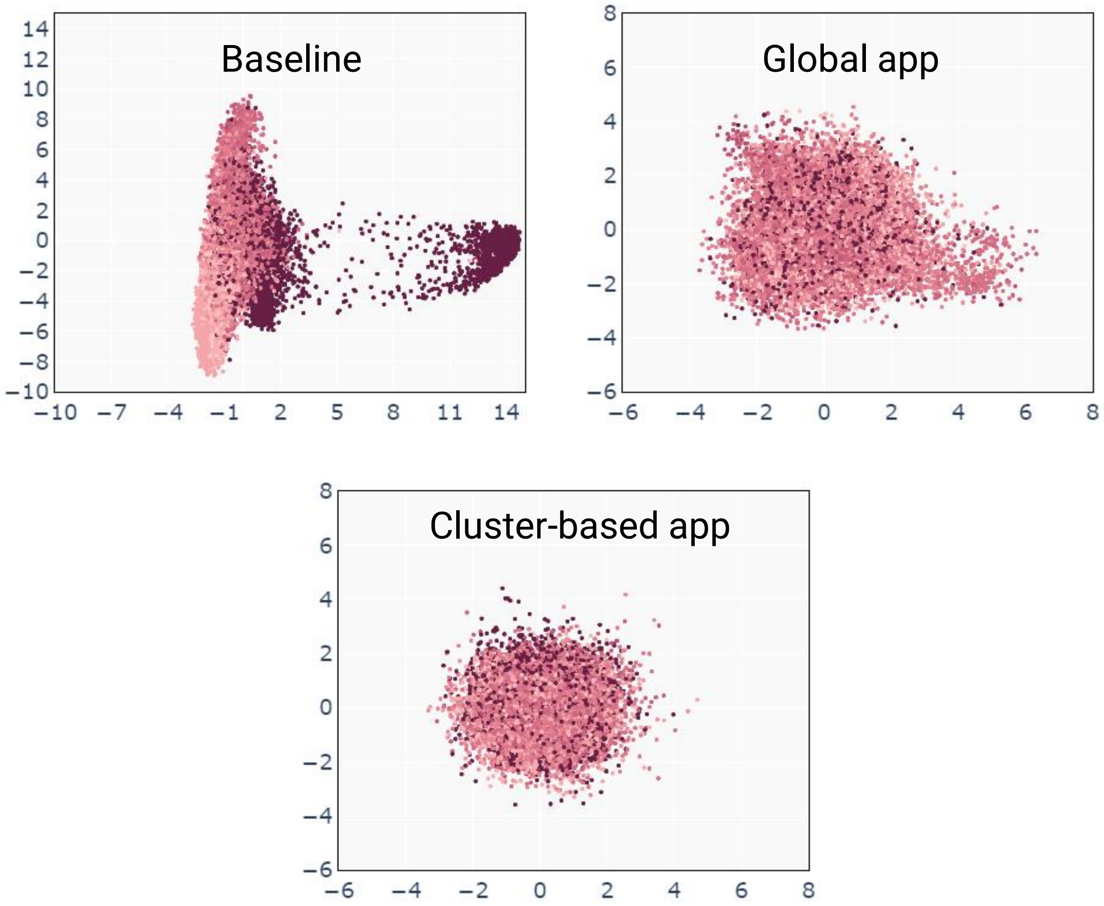
The percentage of nearest neighbours that share similar structural and syntactic knowledge, before (lighter, pattern-filled) and after removing dominant directions in pre-trained CWRs.



Analysis

Word frequency

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

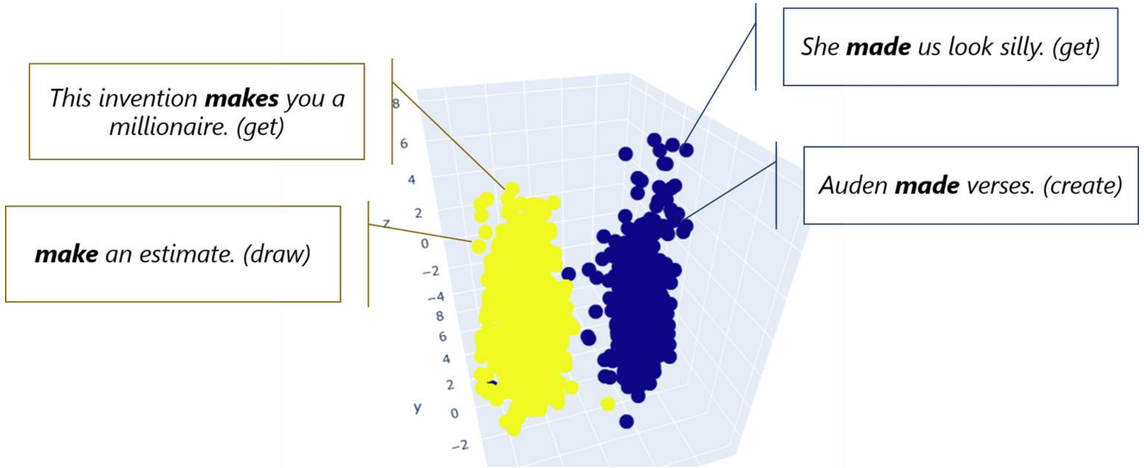


BERT's CWRs visualization using PCA on STS-B dev set. Points color indicates their frequency calculated based on Wikipedia dump; the lighter point, the more frequent.

Analysis

Verb Tense

- Verb representations are distributed based on their tense, not their semantic similarity



Distribution of "make" and "made" in BERT's embedding space

Analysis

Verb Tense

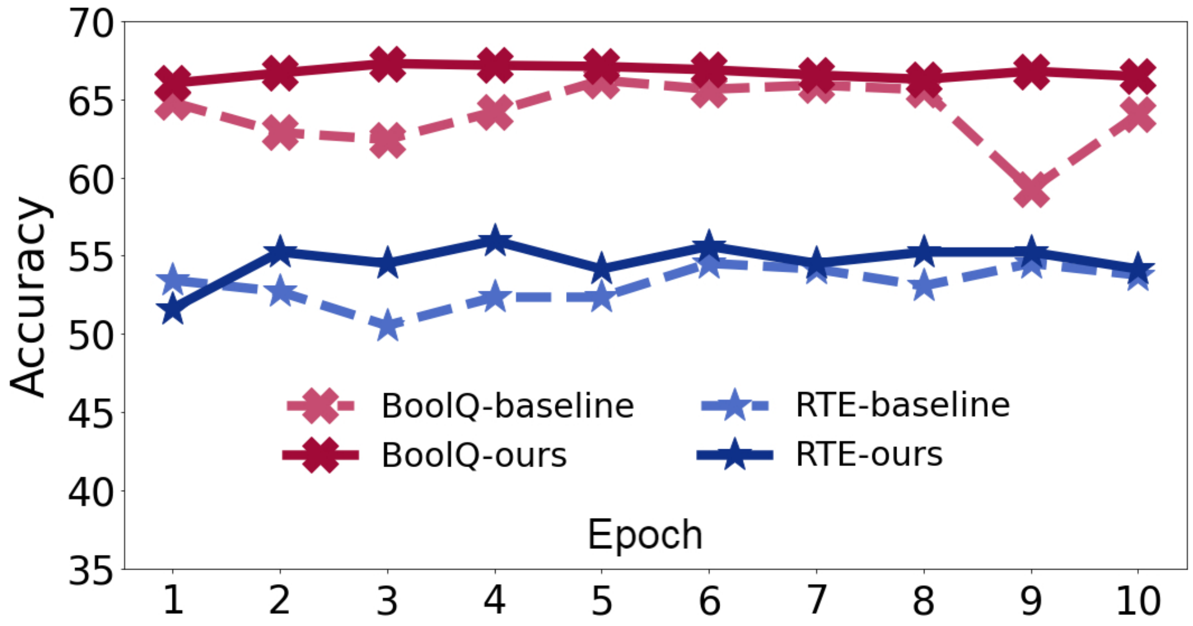
	Baseline				Removed PCs			
Model	**ST-SM**	**ST-DM**	**DT-SM**	*Isotropy*	**ST-SM**	**ST-DM**	**DT-SM**	*Isotropy*
GPT-2	48.82	48.19	50.86	2.26E-05	9.32	9.53	9.49	0.17
BERT	13.44	14.24	14.87	2.24E-05	10.31	10.50	10.32	0.32
RoBERTa	5.89	6.31	6.86	1.22E-06	4.78	5.00	4.89	0.73

Table 4: The mean Euclidean distance of a sample occurrence of a verb to all other occurrences of the same verb with the Same-Tense and the Same-Meaning (ST-SM), the Same-Tense but Different-Meaning (ST-DM), and a Different-Tense but the Same-Meaning (DT-SM). Semantically, it is desirable for DT-SM to be lower than ST-DM.

Analysis

Convergence time

- Isotropy decreases the convergence time



Summary

- Pre-trained LM models are highly **anisotropic**.
- Cosine similarity is an **inappropriate** metric for isotropy.
- Our cluster-based approach can consistently **improve** performance on different tasks.
- Discarded directions encode tense information, structural frequency knowledge

Fine-tuning LMs

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

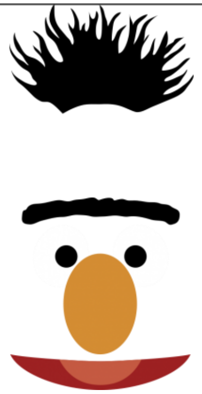
Fine-tuning

- Adding a simple classification layer on top of the pre-trained model
- Training the pre-trained layers and the classifier jointly
- Leading to performance improvement

Questions

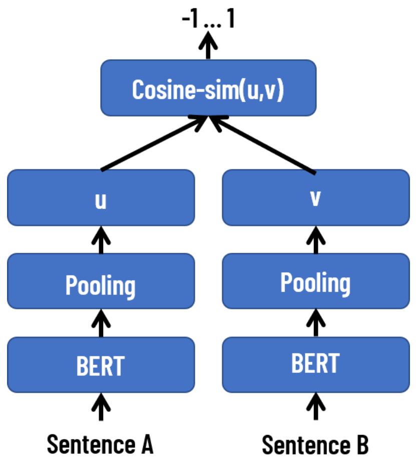
- How does isotropy change during fine-tuning?
- Does isotropy enhancement lead to performance improvement?
- How does the distribution of CWRs change upon fine-tuning?

Setups



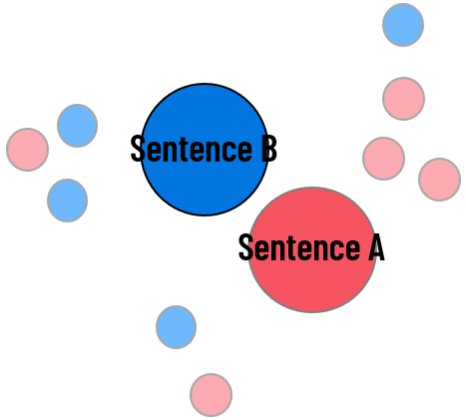
BERT and RoBERTa

Using BERT and RoBERTa-base, which have 12 attention heads and 768 dimensions.



Siamese Architecture

Fine-tuning two pre-trained models simultaneously.

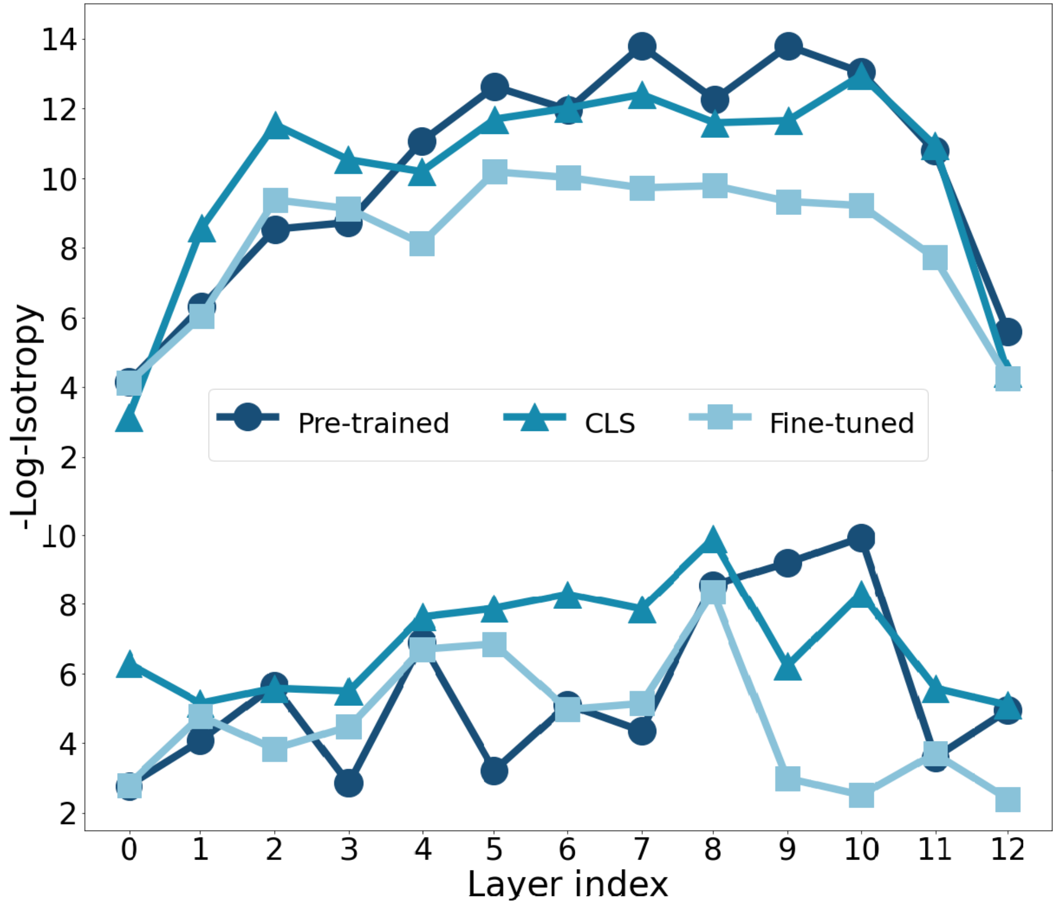


Semantic Textual Similarity

Evaluating the similarity of two sentences in a pair using the STS benchmark dataset.

How does isotropy change during fine-tuning?

The embedding space of fine-tuned models is highly anisotropic.



Isotropy Enhancement

Zero-mean	Clustering+ ZM	Global app	Cluster-based app
Making all representations zero-mean.	Clustering embeddings and making each cluster zero-mean separately.	Discarding a few top dominant directions calculated using PCA.	Clustering embeddings and discarding a few top dominant directions in each cluster.

Results

	Baseline		Zero-mean		Clustering+ZM		Global		Cluster-based	
	Perf.	Isotropy	Perf.	Isotropy	Perf.	Isotropy	Perf.	Isotropy	Perf.	Isotropy
Pre-trained†	54.14	1.1E-5	59.70	1.1E-4	67.73	0.31	69.20	0.59	**74.01**	**0.83**
Fine-tuned†	84.41	4.1E-3	**84.94**	6.6E-3	80.10	0.11	82.14	0.22	64.43	**0.60**
Pre-trained‡	33.99	2.5E-6	37.66	8.3E-2	60.32	0.69	65.99	0.86	**73.86**	**0.95**
Fine-tuned‡	81.08	3.3E-4	**81.34**	6.1E-3	76.03	0.05	79.71	0.18	60.96	**0.28**

Spearman correlation performance and isotropy.

Results

	Baseline		Global App.				Cluster-based App.			
			100 least dir.		700 least dir.		100 least dir.		700 least dir.	
	Perf.	Isotropy	Perf.	Isotropy	Perf.	Isotropy	Perf.	Isotropy	Perf.	Isotropy
BERT	84.41	4.1E-3	84.93	2.2E-3	82.93	2.2E-3	77.87	0.10	75.10	0.16
RoBERTa	81.08	3.3E-4	81.66	3.2E-4	78.59	1.4E-2	73.19	0.13	71.39	0.13

Spearman correlation performance and isotropy.

Summary

- The fine-tuned embedding space is **anisotropic**.
- **Increasing** isotropy **hurts** the performance.
- High **sensitivity** to a few top dominant directions
- The clustered structure of CWRs has been **faded**.

Multilingual Embedding Space

Probing Isotropy in Multilingual Space

Setups

Models

- English BERT
- mBERT

Languages

- English
- Spanish
- Arabic
- Turkish
- Sundanese
- Swahili

Data

- Wikipedia articles

Probing Isotropy

	BERT	mBERT					
	En	En	Es	Ar	Tr	Su	Sw
$I_{Cos}(\mathcal{W})$	0.34	0.24	0.27	0.27	0.25	0.25	0.27
$I_{PC}(\mathcal{W})$	2.4E-5	6.4E-5	5.0E-5	1.6E-5	2.5E-4	1.2E-4	7.8E-5

Isotropy of BERT and mBERT.

Sensitivity to Rogue Dimensions

Cosine Similarity

$$x = (x_1, x_2, \dots, x_d)$$

$$y = (y_1, y_2, \dots, y_d)$$

$$Cos(x, y) = \frac{\sum_{i=1}^{d} x_i y_i}{\|x\| \|y\|}$$

	$I_{Cos}(\mathcal{W})$	**First**	**Second**	**Third**
BERT	0.34	0.385	0.005	0.005
English	0.24	0.041	0.029	0.020
Spanish	0.27	0.033	0.029	0.018
Arabic	0.27	0.033	0.025	0.022
Turkish	0.25	0.036	0.024	0.024
Sundanese	0.25	0.036	0.016	0.016
Swahili	0.27	0.025	0.018	0.014

The contribution of top-three dimensions to the expected cosine similarity in BERT and mBERT models.

Outlier Dimensions

Outliers are the specific dimensions with consistently high value across all representations.

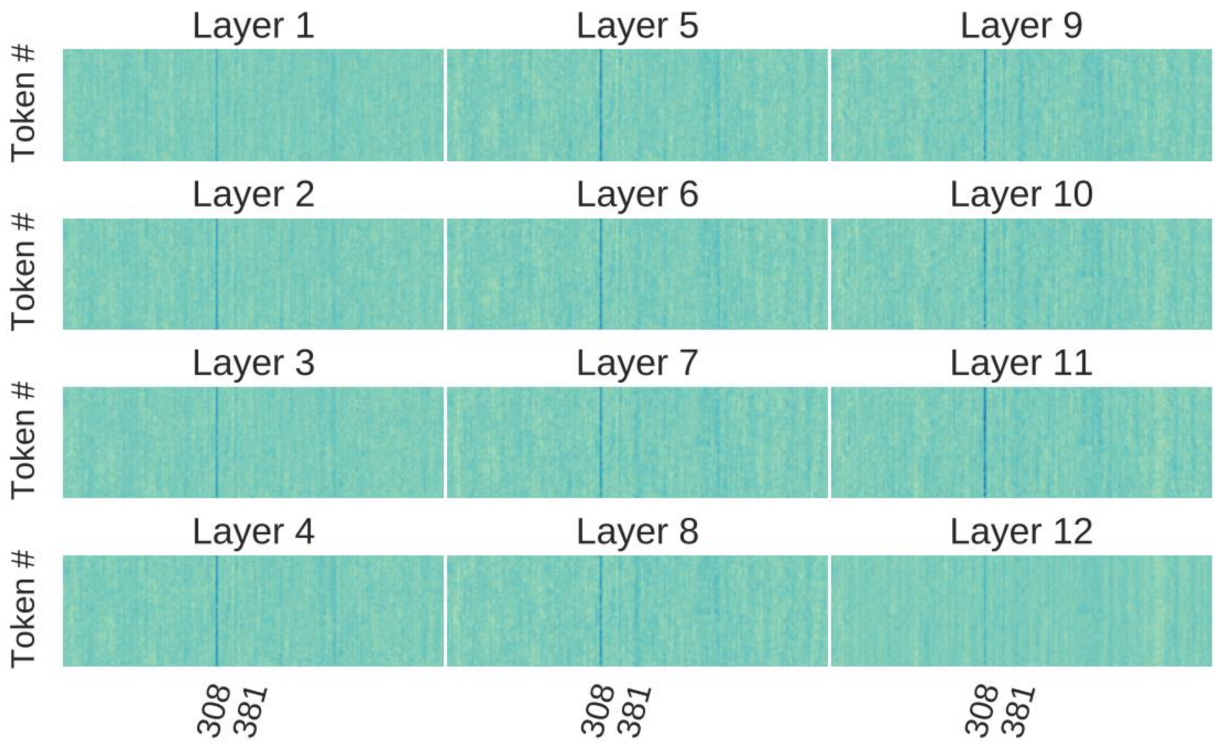


Figure 3: Outlier LayerNorm features 308, 381 in BERT-base-uncased (randomly sampled input).

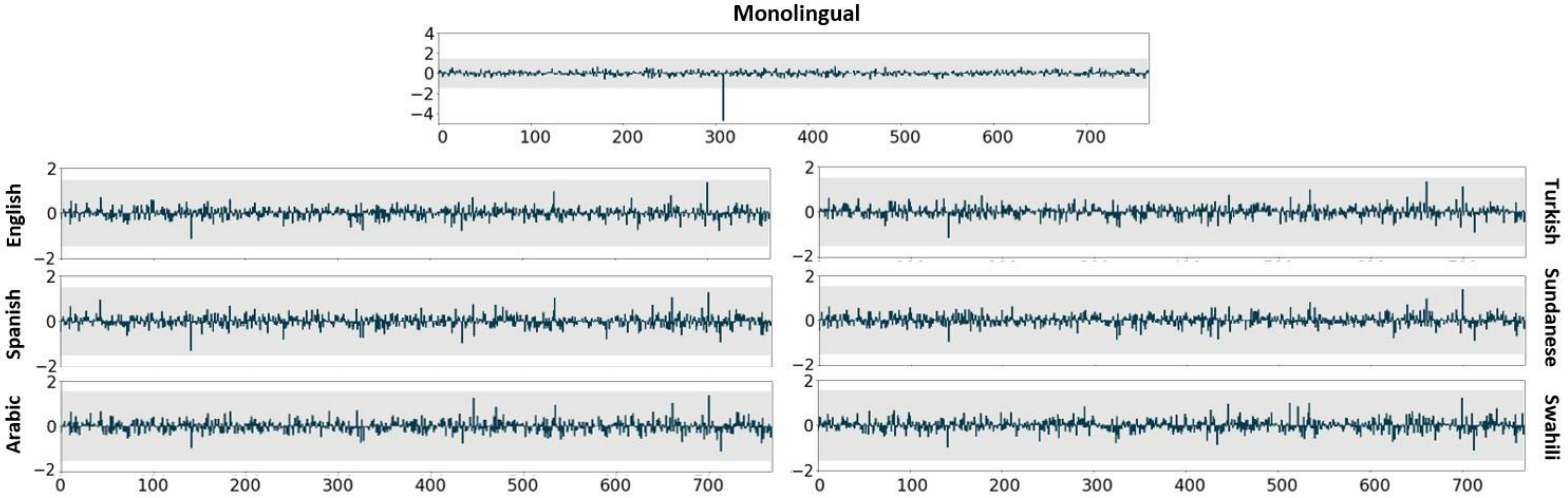
(Kovaleva, et.al, 2021)

Outlier Dimensions

		MRPC	STS-B	MNLI	MNLI-mm	COLA	SST-2	QQP	QNLI	RTE
	Baseline (full model)	87.2	88.8	84.1	84.2	56.8	92.5	89.8	90.6	61.7
Post-ft	Non-outlier†	+0.3	-0.1	-0.2	-0.1	+0.2	0	-0.1	0	-0.4
	Outlier-`308`	-10.5	-23.4	-2.2	-1.8	-2.16	-0.6	-1.0	-1.9	-7.2
	Outlier-`381`	-4.6	-4.4	-13.7	-13.0	-22.2	-3.4	-10.8	-7.3	-5.0
	Random non-outlier pair‡	-1.1	0.0	+0.3	+0.2	-0.5	+0.1	+0.1	0	+0.5
	Outliers `308` + `381`	**-8.6**	**-44.1**	**-27.9**	**-27.2**	**-32.3**	**-20.8**	**-13.0**	**-12.2**	**-10.0**

Performance of BERT-based on GLUE.

Outlier Dimensions



The average of representations.

Isotropy Enhancement

	Ar-Ar	Ar-En	Es-Es	Es-En	Es-En-WMT	Tr-En	En-En
Baseline	51.76 *(8E-5)*	10.61 *(1E-4)*	64.15 *(3E-5)*	31.26 *(5E-4)*	11.39 *(1E-4)*	17.78 *(1E-4)*	60.82 *(2E-6)*
Individual	64.26 *(0.60)*	23.10 *(0.57)*	70.88 *(0.54)*	46.23 *(0.50)*	13.47 *(0.50)*	25.59 *(0.55)*	71.99 *(0.54)*
Zero-shot	52.76 *(6E-5)*	19.36 *(0.04)*	65.69 *(8E-4)*	43.82 *(0.09)*	13.68 *(8E-3)*	19.89 *(0.03)*	-

STS performance and Isotropy.

THANK YOU!