Improving Automated CMG

Megha Chakraborty
Mentors: Venkatesh Vinayakarao (CMI) and Monika Gupta (IBM)

Chennai Mathematical Institute
IBM Global Remote Mentoring (GRM) Program

February 23, 2022

Isotropy in Contextual Word Representations

- *Isotropy*: vectors are uniformly distributed in all directions
- *Anisotropy*: vectors have high cosine similarities leading to a narrow cone structure of the embedding space.
- Lacking isotropy affects:
  - Optimization (accuracy, convergence time)
  - Expressiveness of embedding space
- Improving isotropy can lead to performance improvements

Isotropy in Contextual Embedding Space

A Cluster-based Approach for Improving Isotropy in Contextual Embedding Space [4]

- Existing techniques mostly employ a global assessment to study isotropy
- Local assessment could be more accurate due to clustered structure of Contextual Word Representations(CWRs)

Measuring Isotropy

- From papers [1] and [3]
- Partition function $F(u) = \sum_i^N e^{u^T w_i}$
  - u: unit vector
  - w_i: embedding for ith word $\in W$
  - $W \in \mathrm{R}^{N \times D}$: embedding matrix
  - N: size of vocabulary
  - D: size of embedding
- $F(u)$ can be approximated using a constant for isotropic embedding spaces

Measuring Isotropy

Define $I(W) = \frac{min_{u \in U} F(u)}{max_{u \in U} F(u)}$ close to 1 for isoptropic spaces, where U is set of eigenvectors of $W^T W$

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

This could be what we are looking for to measure isotropy. Further reading of papers [1] and [3] required.

This definition of isotropy has been used in [4]'s implementation. See Sara Rajaee's GitHub

Measuring Isotropy of CC2Vec

- taking only the non-zero columns of W (458 out of 500), we get the isotropy score as: $1.1080678e-05$
- considering entire embedding space of feature vectors, we get the isotropy score as: $2.0109392e-09$

Both these numbers indicate that CC2Vec is extremely anisotropic in nature.

The Next Step

- ‘Why we need to improve isotropy of CC2Vec?’
- ‘When does isotropy matter?’
- ‘Does isotropy matter for CC2Vec?’

A Slight Digression

We have been discussing contextual embedding spaces. Is CC2Vec also contextual?
The answer is: YES!
Reason: Where there is attention, there is context.

Isotropy and Context

Kawin Ethayarajh talks about contextuality of Contextualized Word Representations in his 2019 paper, 'How Contextual are Contextualized Word Representations?: Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings' [2]

Key Points from the Paper

- In all layers of all three models, the CWRs of all words are not isotropic.
- Given that isotropy has both theoritical and empirical benefits for static embeddings, the extent of anisotropy in CWRs is surprising.
- Upper layers produce more context-specific representations and are also more anisotropic.
- *"... suggests that high anisotropy in CWRs is inherent to, or at least a by-product of, the process of contextualization."*

Conclusion and Future Work

Conclusion:

- Increased context specificty is always accompanied by increased anisotropy.

Future Work:

- Given that isotropy has benefits for static embeddings, it may also have benefits for CWRs.

The Next Step for Us

Now, we understand that:

- CC2Vec is contextual and anisotropic.
- Reducing anisotropy of CC2Vec will improve performance for CMG task.
- We have a way of measuring isotropy and we have a performance measure of CMG task as well to compare results of our experiments.

The road ahead:

- We add topics derived from commit messages along with the code changes as input.

Verb-Based Topic Modeling on Commit Messages

- The idea is to introduce topics derived from commit messages in the data set.
- Influence the feature vectors based on the derived topics.
- Theoretically, this would distribute the vectors more uniformly than before.
- Thus, we expect to get a more isotropic embedding space.

Commit Messages Topics As Input

- Observation: most topic words associated with commit messages are “verbs”
- Using Spacy, we extract a list of verbs present in each commit message
- This list of verbs is now representative of topics associated with a code patch
- We then append the list of verbs to the corresponding code change to create the new input
- The following link shows a csv for the new train data set: New code changes csv
- We use the final column, ‘NewCC’ of the aforementioned csv as the new input for CC2Vec

Results

Our experiments have shown that adding topics to the input have improved the isotropy of the embedding space formed as well as given a better results in average bleu score.

- Meausuring Isotropy
- Average Bleu Scores

Thank You

References I

[1] Sanjeev Arora et al. "A latent variable model approach to pmi-based word embeddings". In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 385–399.

[2] Kawin Ethayarajh. "How contextual are contextualized word representations". In: *Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. ArXiv, abs/1909.00512 v1* (2019).

[3] Jiaqi Mu, Suma Bhat, and Pramod Viswanath. "All-but-the-top: Simple and effective postprocessing for word representations". In: *arXiv preprint arXiv:1702.01417* (2017).

[4] Sara Rajaee and Mohammad Taher Pilehvar. "A Cluster-based Approach for Improving Isotropy in Contextual Embedding Space". In: *arXiv preprint arXiv:2106.01183* (2021).