Analysing the Importance of LMs Embeddings' Components on Probing Linguistic Tasks

Pavel Bartenev Bair Mikhailov Kseniia Petrushina Julia Sergeeva
Daniil Shlenskii

Skoltech

March 22, 2024

Presentation Overview

1. Background
   - Language modeling
   - Outlier dimensions
   - Probing tasks

2. Research
   - Problem statement
   - Experiment pipeline
   - Results

Language modeling

- One of the most successful approaches to language modeling is the Transformer architecture
- Encoder produces embeddings – vector representations of the text
- The quality of these embeddings is crucial for solving language-related problems

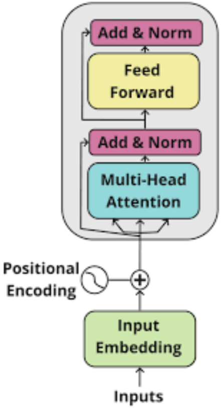


Figure: Transformer encoder

Outlier dimensions

- Important components of transformer embeddings
- Turning them off highly degrade model's language modeling performance

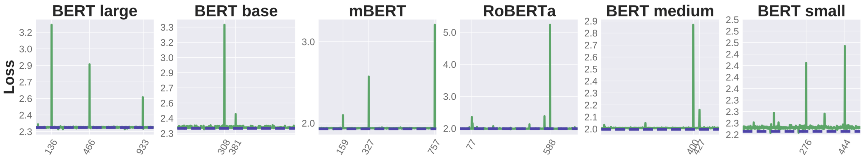
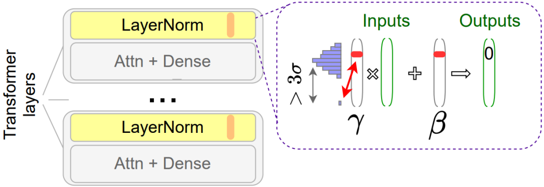


Figure: MLM loss against a turned of dimension

Outlier dimensions. How are they defined?

- Compute mean and std of output LayerNorm weights and biases of all dimensions among all the layers
- For each component on each layer, determine whether it is further than three standard deviations from the mean (for weight and bias)
- If the component deviates greatly from the average on a certain number of encoder layers, then it's called an outlier dimension



Probing tasks

Simple probing tasks are used to discover syntactic and semantic information contained in embeddings:

1. Sentence Length
2. Word Content
3. Bigram Shift
4. Tree Depth
5. Top Constituent
6. Tense
7. Subjects Number
8. Objects Number
9. Semantic Odd Man Out
10. Coordination Inversion

Problem statement

Outlier dimensions have high influence on language modeling tasks. However, it is not well known why this is the case and what information these components contain.

Our **goal** is to figure out if those components contain important information about syntax and semantics.

Experiment pipeline

1. Find outlier dimensions
2. Obtain the vector representations of the probing tasks from ‘roberta-base‘
3. Obtain feature importances:
   - Logistic Regression parameters
   - SHAP for an MLP
   - Gradient Boosting
   - Test accuracy on single features
4. Conduct a comparative analysis of the components

Results. Classification accuracy

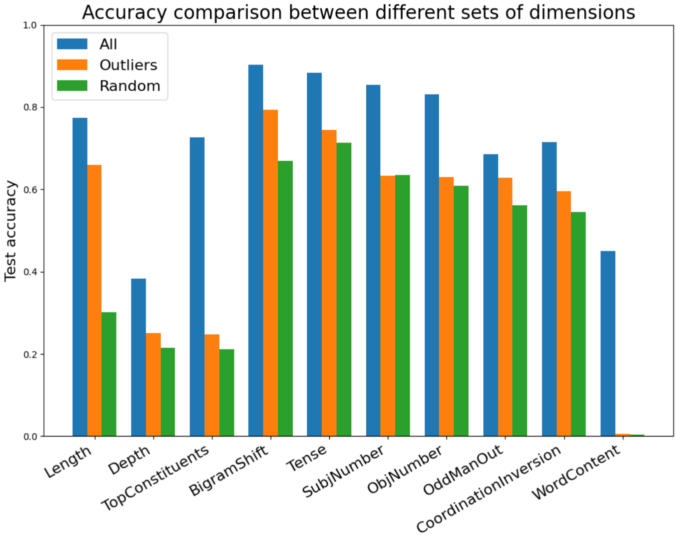
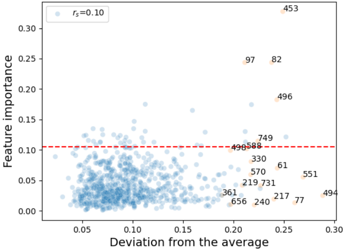


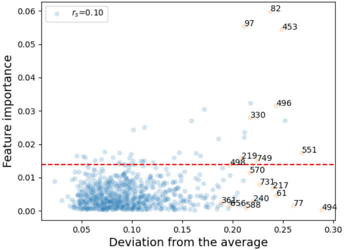
Figure: Test accuracy of logistic regression using different features

Results. BigramShift

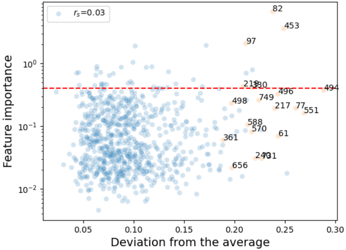
Logistic regression



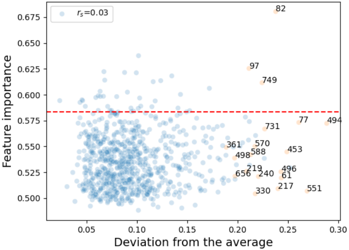
SHAP



Gradient boosting



Test accuracy



Results. Tasks intersection

Top-k Group	50	100	150	200
Surface	61 / 1	61 / 8	61 / 25	61 97 / 47
Syntactic	-	0 / 4	217 / 10	217 61 551 / 27
Semantic	-	-	0 / 4	97 / 7
General	-	-	-	-

Table: Common features in the top-k important features in task groups

Conclusion

- Outlier dimensions perform better than random features on probing tasks
- Several outlier dimensions with high feature importance for each task
- Few distinctive outlier dimensions with syntactic or semantic information