Meeting Formal Linguistics and Optimisation

Lucas Weber, Jaap Jumelet,
Elia Bruni & Dieuwke Hupkes

upf. · Institute for Logic, Language and Computation · Universität Osnabrück · Facebook AI

MAIN CONTRIBUTIONS

1. We introduce *‚similarity probing'*, a method to obtain similarity-estimates of linguistic phenomena from language models (LMs) in an easy-to-apply and computationally-cheap way.

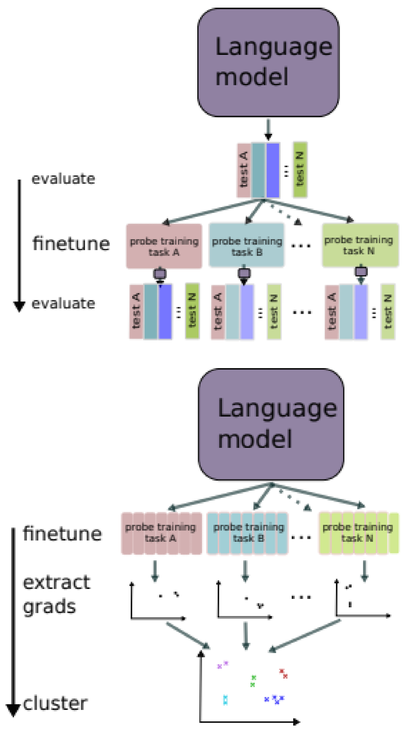
2. We show that linguistically similar sentences evoke similar gradients in language models.

BACKGROUND

1. Weber et al. 2021 consider LMs multi-task learners (MTL), optimizing many different linguistic phenomena at once. They establish a connection between similarity of phenomena and the transfer-learning across them.

2. Yu et al. (2020) show that different tasks interfere as their gradients are pointing in conflicting directions (a.o.).

METHODS



1. Similarity-probing

2. Gradient clustering

1. *Similarity probing* reveals linguistic structure in language models.

&

2. LM-gradients contain information about linguistic phenomena.

DATASET

BLiMP-dataset: minimal pairs of 13 linguistic phenomena, divisible into 67 linguistic paradigms (Warstadt et al., 2020)

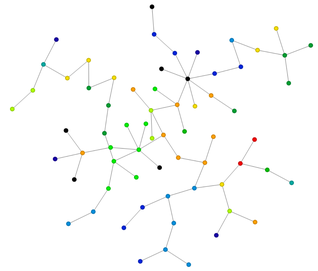
MODELS

LSTM-LM following Gulordava et al. (2018)
Fairseq-transformerLM (Ott et al. 2019)

RESULTS

1. Similarity space:

Maximum Spanning Arborescence of linguistic similarity in a LSTM-LM.

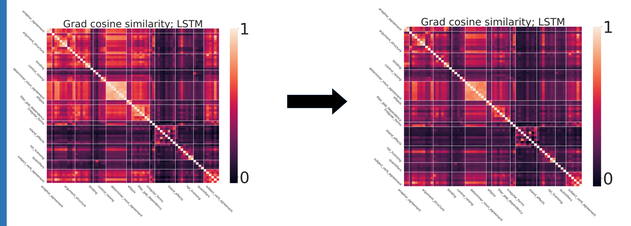


Scan for labeled interactive graph

2. Gradients:

2.1 Gradient similarities

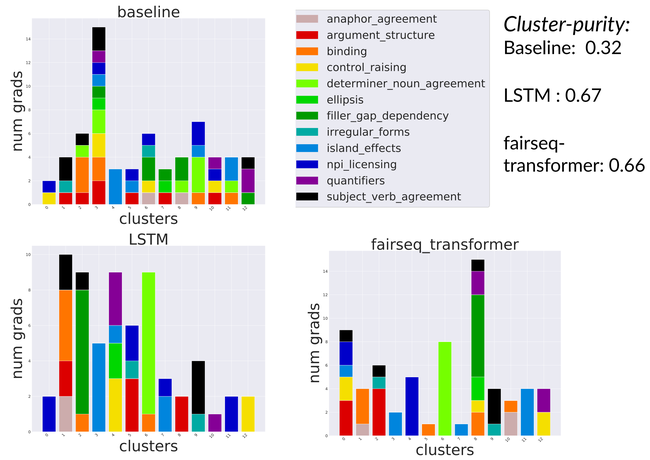
All gradient similarities are > 0 → There are no gradient conflicts in LMs! Further, gradient similarities go towards 0 with increasing pretraining.



pretrain: ½ epoch pretrain: 30 epochs

2.2 Gradient clustering

Gradients of different instances of the same phenomenon cluster together.



Cluster-purity:
Baseline: 0.32

LSTM : 0.67

fairseq-transformer: 0.66

FURTHER QUESTIONS

1. What connects phenomena that have similar gradients?
2. How can we use linguistic knowledge to improve model performance?
3. Why are gradients in implicit and explicit MTL so substantially different? What does that mean for interference in both types of settings?