Neural Language Models are not Born Equal to Fit Brain Data, but Training Helps

Alexandre PASQUIOU

Yair LAKRETZ

John HALE

Bertrand THIRION

Christophe PALLIER

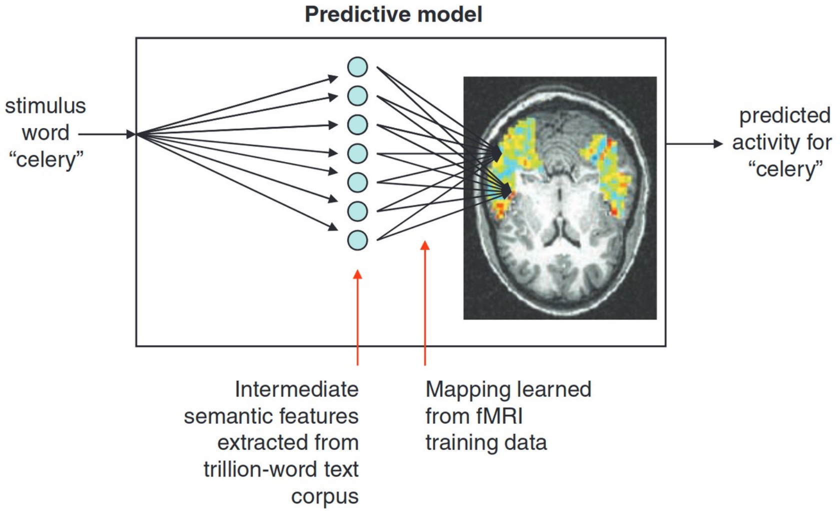
INSERM, INRIA, CEA, Neurospin, Gif-sur-Yvette, France
Dept. of Linguistics, U. of Georgia, Athens,USA.

Using Neural Language Models to explore language processing in the human brain

Trained on large corpora, NLMs acquire some linguistics knowledge, notably some aspects of the semantics of words or sentences.

The activation patterns in these networks correlates with brain activations.

(Mitchell et al., 2008, *Science)*


Predictive model
stimulus word "celery"
predicted activity for "celery"
Intermediate semantic features extracted from trillion-word text corpus
Mapping learned from fMRI training data

Questions

- Which characteristics of neural language models (model type, number of layers) affect the prediction of brain data?
- Is model Perplexity a good predictor of a model’s ability to fit brain data?
- What is the effect of the size of the training corpus on brain predictability ?

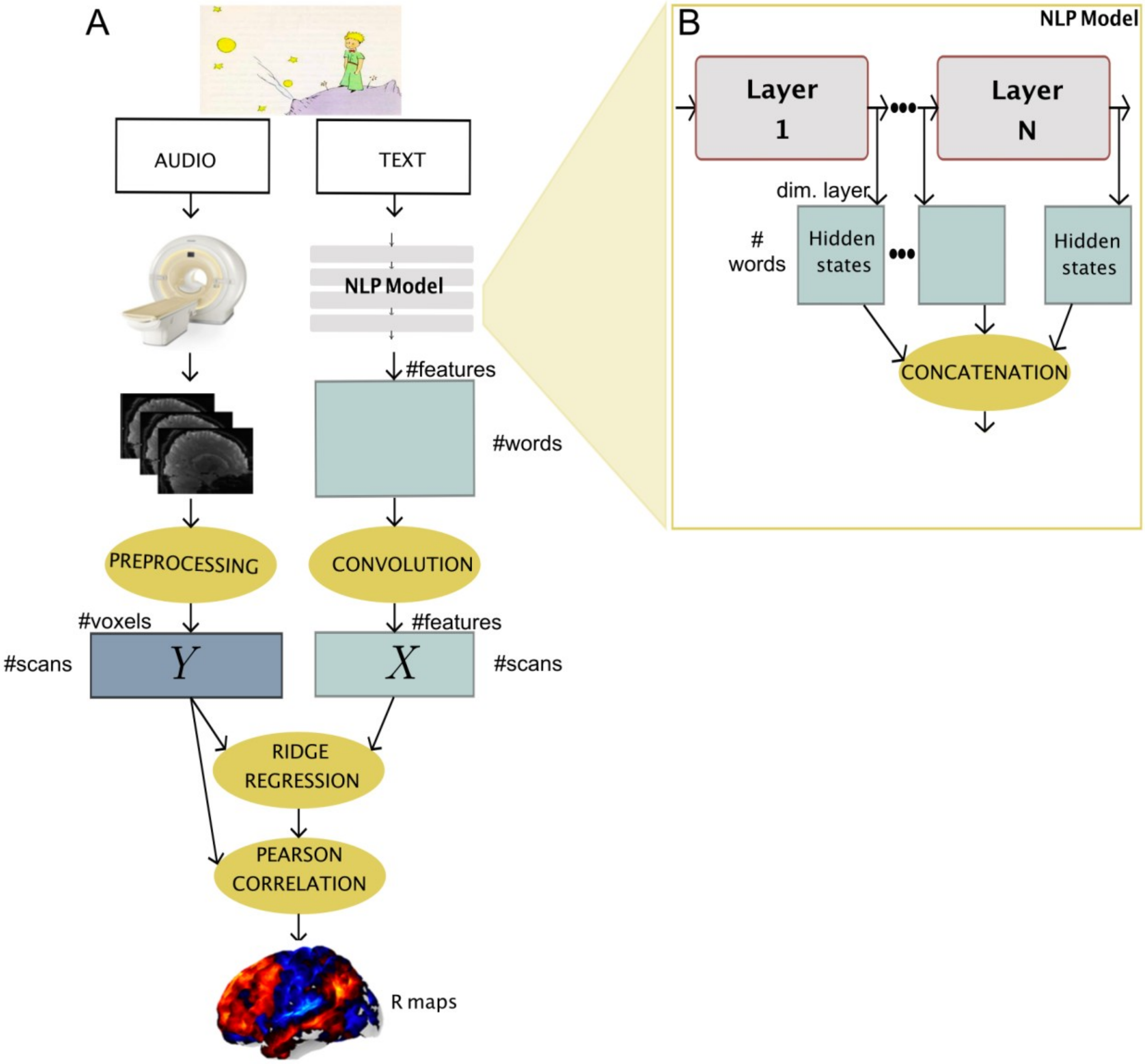
Comparing NLM activations with brain activations

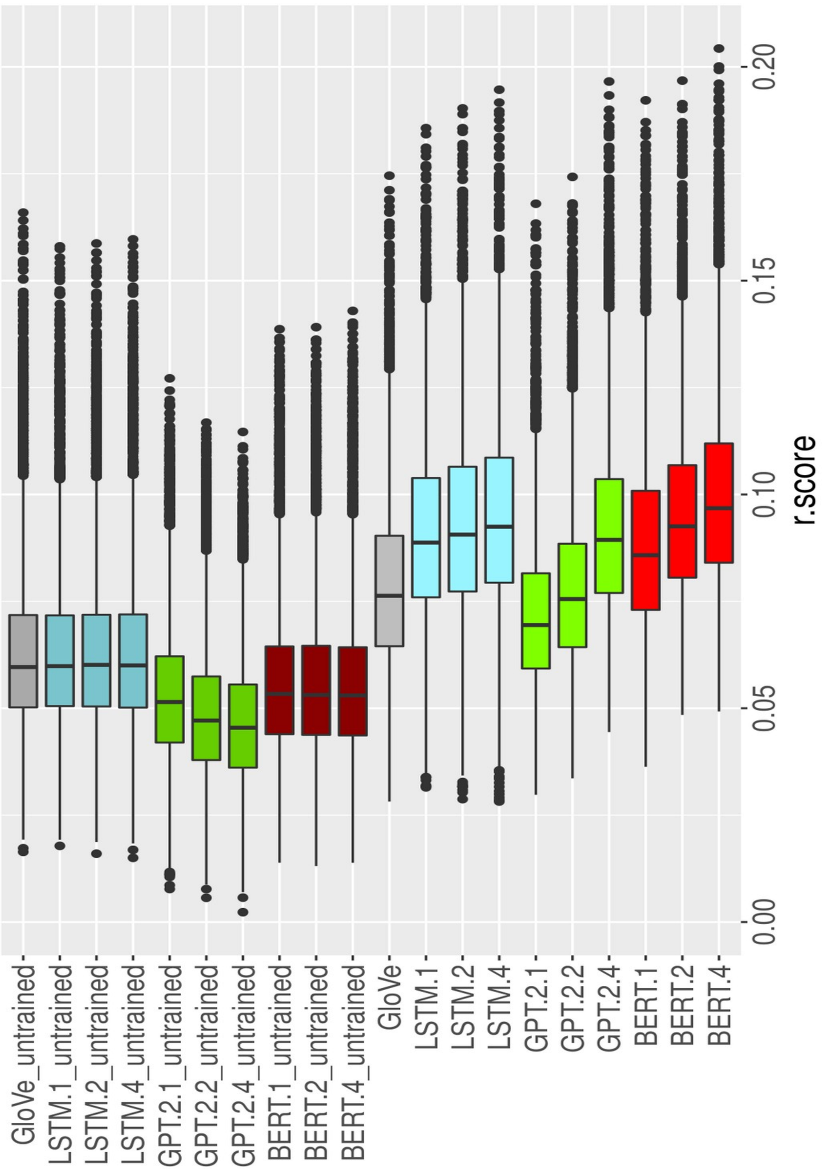
Human participants (N=51) were presented with an auditory version of **The Little Prince** story while their brain activity was recorded with fMRI.

Neural models were presented with a text transcription and the entire state of the network was recorded for each word.

After several pre-processing steps, the two signals were aligned using a regression model.

Finally, cross-validated correlation coefficient between models' predictions and fMRI time-series to produce **R maps**.





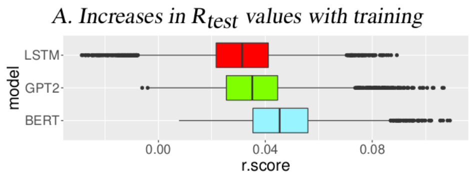
Models performance

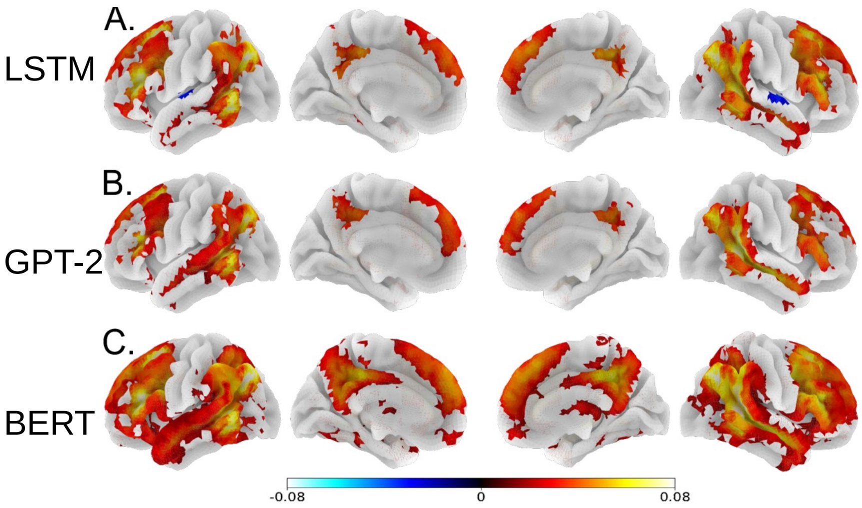
- Untrained models achieve significant brain scores, with untrained LSTM and Glove performing better than transfomers (GPT-2 and BERT)

  The fitting process captures the similarities in brain responses to words that repeat in both the train and test sets.

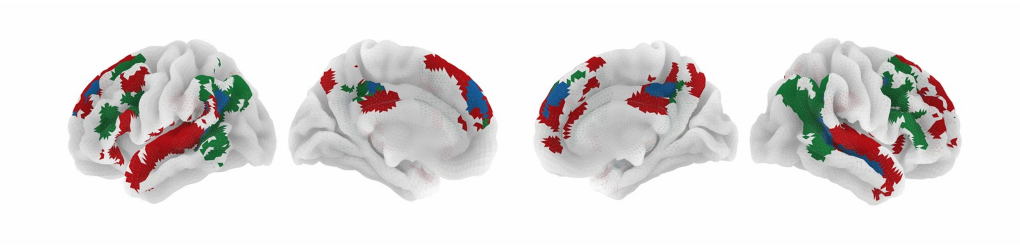
- Transformers benefit more from training than LSTMs. The more layers they have, the better their fitting performance.

Regions where training improves brain score





Training significantly improves brain scores in the same brain regions regardless of models.

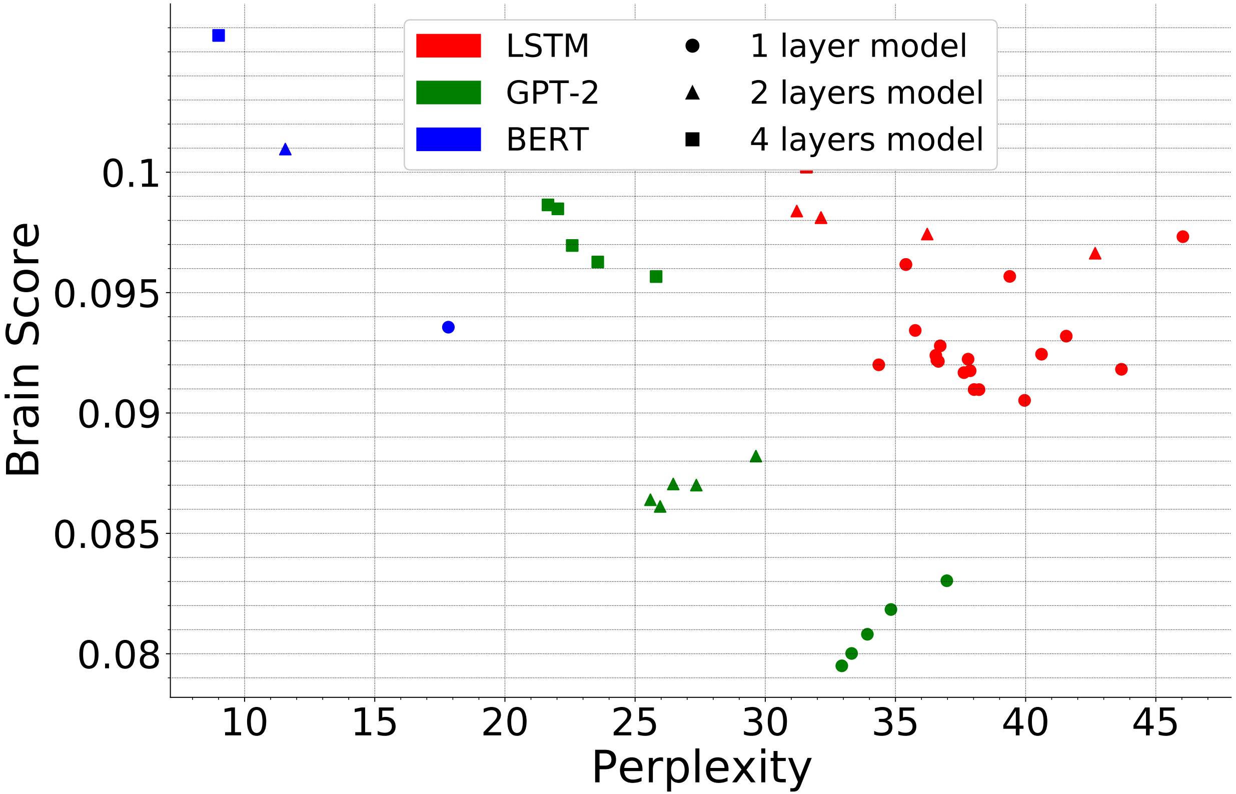


In Green, regions showing a increase of R with training, across models.

In Red: regions showing significant effects for untrained models

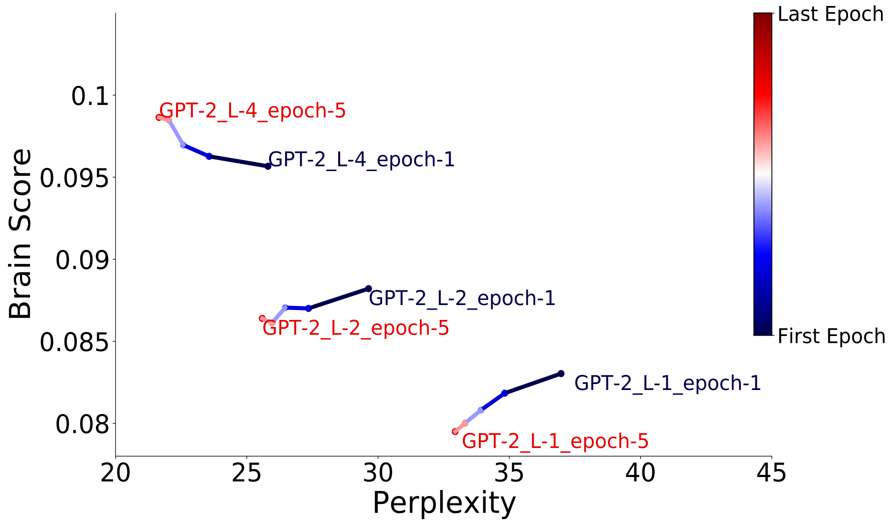
In Blue: their overlap (18%).

Relationship between perplexity and brain score

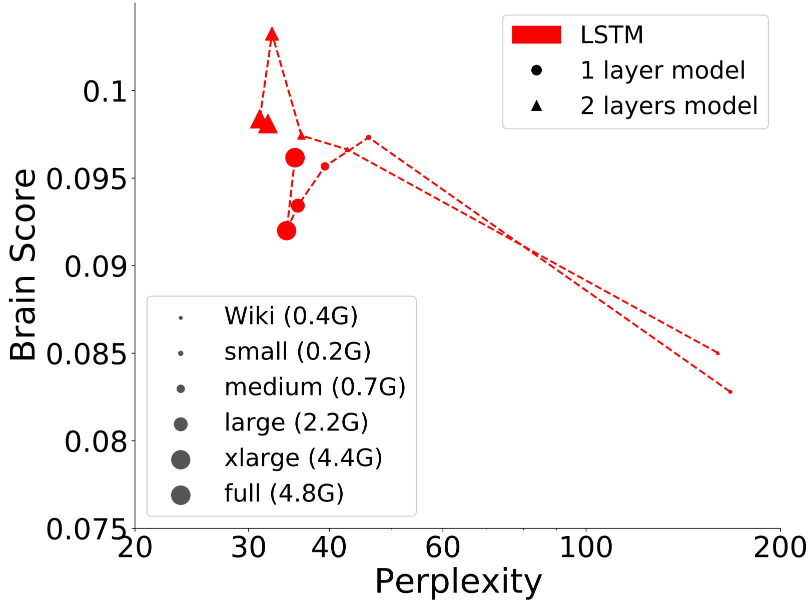


Perplexity (contd.)

Effect of training epoch



Effect of training dataset size



Conclusions

- Untrained models achieve significant brain scores (The fitting process captures the similarities in brain responses to words that repeat in both the train and test sets).
- Transformers benefit more from training than LSTMs. The more layers they have, the better their fitting performance.
- Training has an effect on brain scores in the same set of brain regions, consistently across models.
- Perplexity is not an efficient predictor of brain score.
- The size of the training corpus significantly affects brain scores. Off-the-shelf models, trained on small datasets, might lack statistical power to capture brain activations.

References

SCAN ME

Caucheteux, C., Gramfort, A., and King, J.-R. “Disentangling Syntax and Semantics in the Brain with Deep Networks.” In ICML 2021 - 38th International Conference on Machine Learning, France, 2021. https://hal.archives-ouvertes.fr/hal-03361421.

Hale, John T., Luca Campanelli, Jixing Li, Shohini Bhattasali, Christophe Pallier, and Jonathan R. Brennan. 2021. “Neurocomputational Models of Language Processing, 2021, *Annual Reviews in Linguistics*, 8, 427-446. https://doi.org/10.1146/annurev-linguistics-051421-020803.

Huth, Alexander G., Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. 2016. “Natural Speech Reveals the Semantic Maps That Tile Human Cerebral Cortex.” *Nature* 532 (7600): 453–58. https://doi.org/10.1038/nature17637.

Mitchell, T. M., S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. 2008. “Predicting Human Brain Activity Associated with the Meanings of Nouns.” *Science* 320 (5880): 1191–95. https://doi.org/10.1126/science.1152876.

Schrimpf, Martin, Jonas Kubilius, Michael J. Lee, N. Apurva Ratan Murty, Robert Ajemian, and James J. DiCarlo. 2020. “Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence.” *Neuron*, https://doi.org/10.1016/j.neuron.2020.07.040.

Toneva, Mariya, and Leila Wehbe. 2019. “Interpreting and Improving Natural-Language Processing (in Machines) with Natural Language-Processing (in the Brain).” In *Advances in Neural Information Processing Systems* 32, edited by H. Wallach et al., 14954–64. .