Sample Research Summary

Samuel Debray

November 23, 2022

Abstract

The progress of AI in the last few years has allowed great progress to be made both in computer science and cognitive sciences. Capitalising on techniques used to compare AI and the brain on linguistic tasks, this project attempts to use AI to understand the cognition of advanced mathematics. We did so in three different ways. Firstly, we gathered a mathematical corpus from Wikipedia to generate a mathematical vocabulary and its semantic embeddings using GloVe, which computes distributional semantic representations. We then found that the first principal component of the GloVe embeddings of a global corpus made up of mathematical and non-mathematical pages enables to retrieve the fMRI activations in the brain's mathematical network. Finally, we used a pretrained model of the Transformer, a state-of-the-art neural network, to analyse advanced mathematical and general knowledge statements. We showed that the model makes a distinction between meaningful and meaningless statements, and that its behaviour correlates with that of non-mathematicians when presented with meaningless mathematical statements.

1 Introduction

In the last few years, the field of Artificial Intelligence (AI) has been the subject of great upheaval. The development of state-of-the-art neural networks like the Transformer (Vaswani et al., 2017) has made it possible to build bridges between algorithms and the brain (Caucheteux and King, 2022; Pereira et al., 2018; Schrimpf et al., 2021). Computer scientists have also realised that AI models are endowed with great reasoning abilities and are now able to automatically prove mathematical theorems (Polu et al., 2022; Polu and Sutskever, 2020) and to solve non-trivial computational problems (Charton, 2021; Charton et al., 2021; d'Ascoli et al., 2022; Kamienny et al., 2022; Lample and Charton, 2019).

A fair amount of research has been conducted on the links between AI and the brain's treatment of language on the one hand, and on AI's mathematical abilities on the other hand. Suprisingly, though, virtually nothing has been done on AI and the cognition of advanced mathematics. This project is a first step in this direction, capitalising on the recent discoveries on mathematical cognition made by Amalric and Dehaene, 2016, 2018. More precisely, there are two main questions we seek to answer:

(i) Do distributional semantic representations actually capture mathematical semantics? And can they be used to analyse fMRI data?

(ii) Do Transformer-class models, which proved to be gifted in mathematics, process maths in the same way as humans?

By looking in this direction, we aim at understanding better the way advanced mathematics are processed in the brain. Ultimately, this could lead to the emergence of artifical scientists, capable of performing scientific research on their own or, at the very least, to help scientists in the most tedious parts of their job; all by imitating the human brain. In short, the mutual benefits of AI and neuroscience are potentially immense.

2 Use of Distributed Semantic Representations

2.1 A Vocabulary of Mathematics

We created a $1,000$-word mathematical vocabulary by reviewing manually the vocabulary computed by the GloVe pipeline (Pennington et al., 2014) from the mathematical articles of French Wikipedia. For each word of the vocabulary, we obtained a 500-dimensional GloVe semantic vector, which is supposed to encode semantic information obtained from cooccurrences.

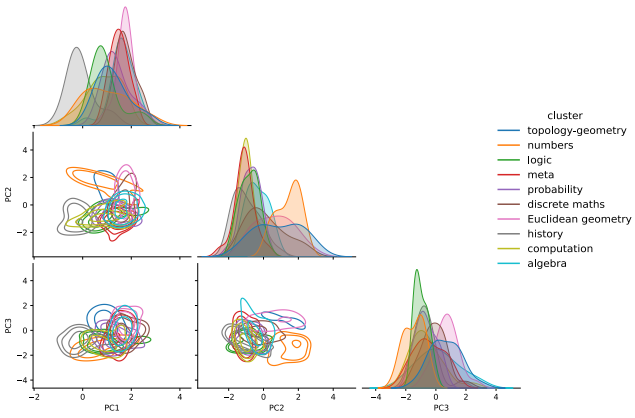

Fig. 1. Kernel density estimate plot of the twenty most frequent words of each cluster in the planes PC1-PC2, PC1-PC3 and PC2-PC3. Levels encompass 30%, 50% and 80% of the probability mass. PC1 accounts for 6.87% of the observed variance, PC2 for 6.20% and PC3 for 5.67%.

We performed a PCA (Wold et al., 1987) of the vectors and split them into ten clusters that we labelled by hand. We then represented 10 vectors by cluster in the planes PC1-PC2, PC1-PC3 and PC2-PC3 as depicted on figure 1.

No cluster seems to stand out, except maybe "hisory" (which contains the names of famous mathematicians) on PC1 and "numbers" on PC2. Overall, the features encoded by the first principal components are not easily identifiable.

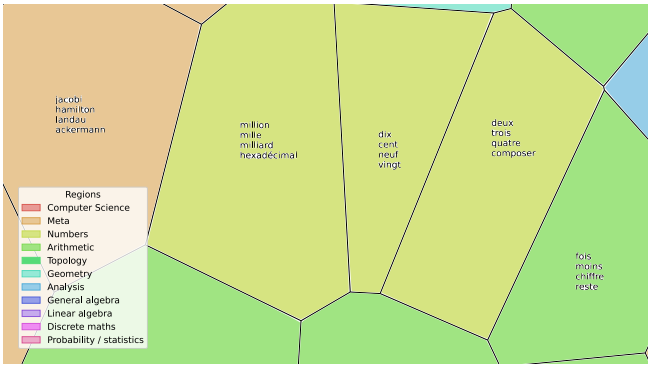

Fig. 2. View of a region of the clustered two-dimensional semantic map of mathematics.

We then obtained a 2D semantic map of mathematics by dividing the vectors into 100 clusters using spectral clustering (von Luxburg, 2007) and by projecting them in two dimensions with tSNE. A region of the map is presented on figure 2. The borders between clusters have been computed using Voronoi tessellation and the colours were assigned by hand to reflect the perceived mathematical tune of each cluster.

The spacial distribution of clusters does not allow to indentify larger clusters dedicated to geometry or algebra for instance. Interestingly, names of mathematicians are not grouped with their field of interest but are all clustered together, thus creating "history of mathematics" clusters. Moreover, some basic parts of mathematics, like arithmetics and numbers, are split into several clusters of increasing level of elaboration (e.g. "very small numerals" < "larger numerals" < "powers of ten" or "basic operations" < "modular arithmetic").

2.2 Modelling Brain Activations

We reanalysed the fMRI data from Amalric and Dehaene, 2016 using parametric GloVe embeddings instead of categorical regressors. The embeddings were computed from all French Wikipedia pages (not only the mathematical articles this time).

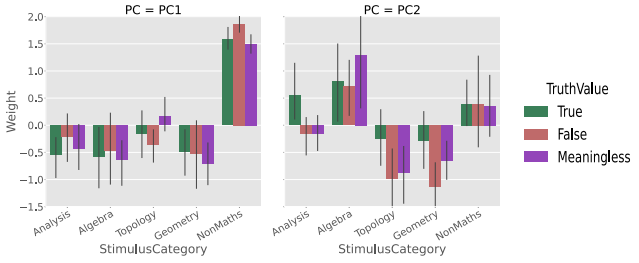

Fig. 3. Projection of the stimuli's embedding onto PC1, PC2 and PC3 as a function of their category and truth value (bars show 95% confidence intervals).

We performed a PCA on the embeddings of stimuli and found that the first PC encodes the mathematical or non-mathematical nature of stimuli. The PCs stop to be interpretable from PC2 however (see figure 3).

This approach allowed us to retrieve the mathematical network in mathematicians highlighted by Amalric and Dehaene, 2016 (see figure 4). The other PCs, however, had no significant effect.

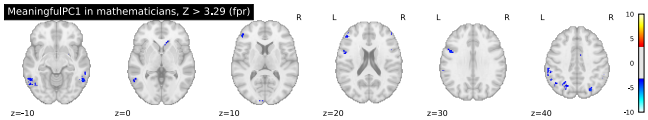

Fig. 4. Effect of PC1 in mathematicians, FPR correction $q < 10^{-3}$.

3 Use of Transformer-class models

We evaluated GPT-fr, a model of the Transformer trained on a large dataset in French, on stimuli from two mathematical experiments conducted at NeuroSpin (the one from Amalric and Dehaene, 2016 and one that is not published yet).

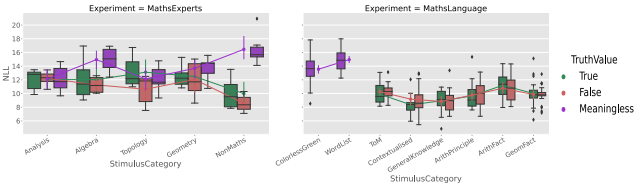

Fig. 5. GPT-fr's maximum negative log likelihood by stimulus as a function of stimuli's category and truth value.

As shown on figure 5, GPT-fr is not able to distinguish between true and false or even between mathematical and non-mathematical stimuli, but only between sensical and non-sensical ones.

Fig. 6. Percentage of subjects judging the stimulus is not true against GPT-fr's maximum negative log likelihood (one point represents one stimulus).

We then probed whether GPT-fr's prediction score correlated with subjects' estimation of the truthfulness of stimuli. As shown on figure 6, the correlation is very poor except for meaningless stimuli in non-mathematicians.

4 Discussion

4.1 General Discussion

Section 2.1 evaluates how well GloVe captures the mathematical semantics of mathematical words. The clustering shows that GloVe does capture a fair amount of mathematical semantics, as spectral clustering brings out an interesting classification of mathematics. However, the variance of the GloVe embedding of the vocabulary spans over many

directions, which results in not very explanatory principal components, that are poorly interpretable in addition.

Regarding the fMRI data, it seems that the GloVe embeddings do not provide more information than a mere categorical regressor. However, we showed that the first principal component of the embeddings makes a clear distinction between mathematical and non-mathematical stimuli, and enables to retrieve the mathematical network from Amalric and Dehaene, 2016.

Finally, in section 3, we showed that GPT-fr is able to make the distinction between meaningful and meaningless statements, and that its behaviour correlates with that of non-mathematicians when presented with meaningless statements. However, it seems that the mathematical abilities of the default GPT-fr model (i.e. with no additional training) of the Transformer do not go beyond this distinction and are thus very limited.

4.2 Future Work

First, the two-dimensional clustered map of mathematics we obtained informs us about how GloVe represents mathematical semantics, and especially how it connects concepts together. Here, connections are represented by the clustering, but they can also be quantified directly from the vectors by cosine similarity:

$$\mathrm{cosine}(\mathbf{u}, \mathbf{v}) := \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\|_2 \cdot \|\mathbf{v}\|_2}.$$

It would be interesting to compare this representation with that of humans, and especially to see (i) if humans predict the same similarities as GloVe and (ii) if they produce the same clustering. We would exepct the answer to these two questions to depend on the mathematical training of subjects.

Second, we found that the principal components of the GloVe embeddings of the global vocabulary do not enable a more fine-grained analysis of fMRI data than a mere categorical model. This might be due to the dimension reduction performed by the PCA, as the principal components are no longer interpretable from the second component. This problem might be solved by using other methods to lower the dimensionality of the GloVe embeddings (e.g. tSNE or maybe even asking for three-dimensional GloVe vectors directly), or by performing PCA across voxels like Huth et al., 2016.

Finally, training a Transformer's model (from scratch or by fine-tuning) instead of using a pretrained model could help improve performance both at classifying statements and at predicting subjects' behaviour. If so, the hidden states of the model should better predict brain activity on mathematical tasks. In the spirit of Pereira et al., 2018 the analysis of fMRI data thanks to these hidden states could also enable to train a decoder able to predict subjects' behaviour (i.e. their evaluation of a statement as true, false or meaningless) from brain activation.

References

Amalric, M., & Dehaene, S. (2016). Origins of the brain networks for advanced mathematics in expert mathematicians. *Proceedings of the National Academy of Sciences*, *113*(18), 4909–4917.

Amalric, M., & Dehaene, S. (2018). Cortical circuits for mathematical knowledge: Evidence for a major subdivision within the brain's semantic networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1740).

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1).

Charton, F. (2021). Linear algebra with transformers.

Charton, F., Hayat, A., & Lample, G. (2021). Learning advanced mathematical computations from examples.

d'Ascoli, S., Kamienny, P.-A., Lample, G., & Charton, F. (2022). Deep Symbolic Regression for Recurrent Sequences [arXiv:2201.04600 [cs]].

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.

Kamienny, P.-A., d'Ascoli, S., Lample, G., & Charton, F. (2022). End-to-end symbolic regression with transformers [arXiv:2204.10532 [cs]].

Lample, G., & Charton, F. (2019). Deep Learning for Symbolic Mathematics.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, *9*(1), 963.

Polu, S., Han, J. M., Zheng, K., Baksys, M., Babuschkin, I., & Sutskever, I. (2022). Formal Mathematics Statement Curriculum Learning.

Polu, S., & Sutskever, I. (2020). Generative Language Modeling for Automated Theorem Proving. *arXiv:2009.03393 [cs, stat]*.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *arXiv:1706.03762 [cs]*.

von Luxburg, U. (2007). A Tutorial on Spectral Clustering. *arXiv:0711.0189 [cs]*.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1), 37–52.