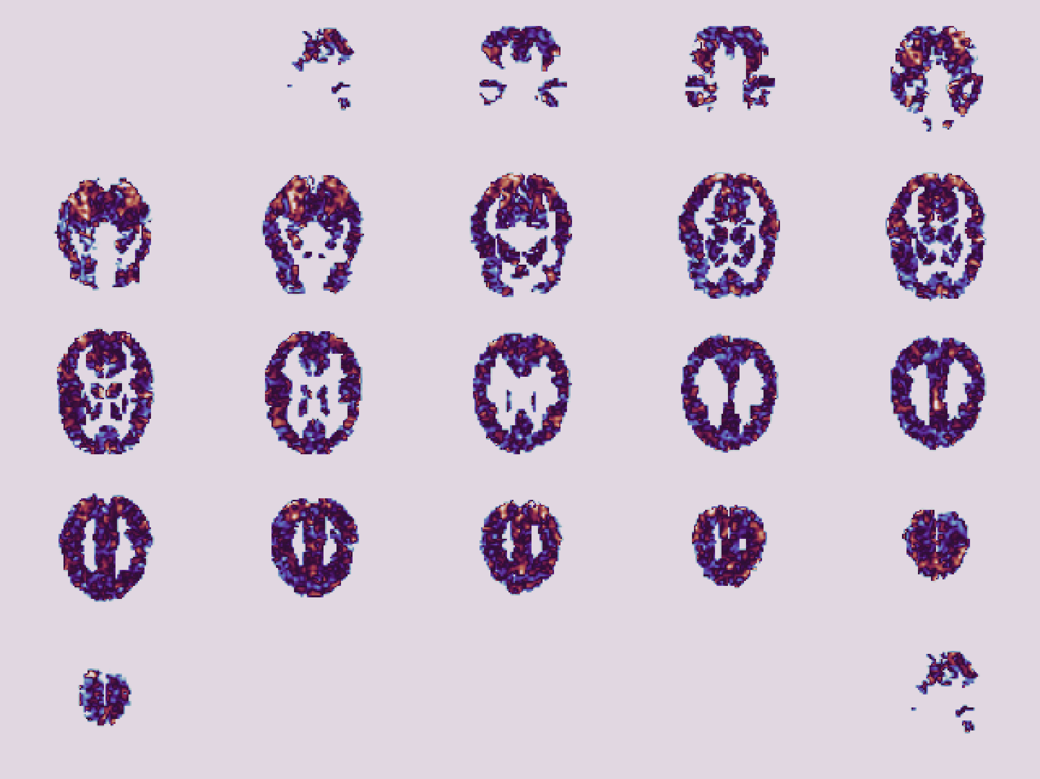
Introduction to Machine Learning | CSD361

Semantic Interpretation of Distributed Neural Activations

Interpreting nouns from fMRI scans



Aman Kanchi Pendyala | 2010110074

Surya Raghunath | 2010111065

Samyak Jain | 2010110763

Introduction

Mind and brain act together, even if the question of equivalence is contentious in philosophy and science. Since this relation stands to reason, one may attempt to interpret the mind's contents by observing the brain alone.

Using insights from computational linguistics and neuroimaging, Mitchell et al. 2008 (from now, "Paper") demonstrated that to the extent that the words for semantic concepts couccur in natural language, their mental representations too bear similarity. They showed that if a concept can be well represented, one can predict with good accuracy what a human brain may look like while beholding it as seen through the MRI machine.

The question is whether the reverse can also be established as strongly; that is whether semantic meaning can be interpreted from neural activity.

We show that by taking a similar approach to word representations, one can establish a strong relation (one at least as good as that of the Paper's) between neural activation patterns and perceived meaning.

Reproducing the Paper

To develop an understanding of the nature of the data and how one could go about “reversing” the work of the Paper, we began by first reproducing it.

The data shared by the authors of the Paper, being as it is fairly old, is not readily receptive to modern machine learning treatment. Some work went into standardizing the data.

The fMRI brain scans are represented as [51, 61, 23]-shaped tensors of real values roughly within (-10, 10). However, the distributions of ‘voxel’ activations were noticed to not only be skewed, but also with entirely different parameters in every scan.

We began by coercing the data to a more Gaussian distribution using the Yeo-Johnson transformation, translating the values to have a mean of 0 and scaling them to have a variance of 1. If this choice seems not strongly motivated it must be noted that when attempting the Paper’s reverse, we implement batched training, and there is a mountain of literature establishing that “normalized” data is conducive to much more stable and successful training in deep networks trained with Stochastic Gradient Descent. Additionally, our results match what the Paper reports, which is to be expected given the linear approach to modelling.

To summarize the Paper’s work,

(with some handwaving for brevity’s sake)

- Each of the 60 nouns in the data is represented as a 25-vector of sliding-window co-occurrence values with 25 pre-selected verbs in a large natural language corpus, scaled to unit L_2 norm.
- The model is a basis set of 25 scan-shaped tensors, each meant to represent/learn one of the verbs.
- When a word is input to the model (as its 25-vector), the predicted brain scan is the componentwise sum of the basis tensors, each scaled by the corresponding word-vector component.
- Thus training such a model (the Paper uses MLE, we resort to GD out of convenience) to match the true brain scan corresponding to the word results in the model learning

“intermediate” representations of each of the 25 verbs as they might be thought of as appearing in scans.

- Each time, the model is trained on only 58 of the 60 available words, and the remaining 2 are put aside to measure “accuracy” with. With T1 and T2 as the true scans and P1 and P2 being the corresponding predictions, the correctness criterion for one run of this ${}^{60}C_2$ operation is: S(T1, P1) + S(T2, P2) > S(T1, P2) + S(T2, P1) where S is the Cosine-similarity of two vectors (higher S = more similar). This is essentially a test for “confusion.”
- The Paper reports (and we reproduce) that with this setup, the model is correct ~77% of the time, with $p < 10^{-11}$ per their estimate.

Our reproduction of the Paper deviates only to the extent that while they take the S over 500 pre-chosen voxels found to have a high correlation to the meaning of the word, we take it over the entire scan. Of course, given that not the entirety of the scan correlates with each prediction, this relaxation should not *increase* the measured accuracy, at least significantly.

Convolutional approach

Our original approach was to treat this problem as the classic image classification problem, with convolutions, softmax output with one-hot class encodings, and carrying over well-established image classification methods and heuristics.

We state at the outset that results were mixed and did not lead to any insights except maybe that one must exercise caution going at all nails with the same deep-learning-shaped hammer. We express more thoughts under the section Discussion.

Data Augmentation

It is common in image classification to rotate, scale, translate, flip available training images to the effect of emulating a much larger dataset than is actually available. Along these lines, given the tiny size of the scan dataset, we adopted the random-rotation augmentation, keeping the range of rotation small so as not to disorient the scans too much.

1. 3D Convolutions

With the assumption that spatial features are important to inference of meaning from scans, we used 3D convolution layers, followed by a Dense classifier with a Softmax output. The data augmented as explained above was batched 64-wide and trained against the Categorical Cross-entropy as is standard in softmax classification.

2. Slice-wise 2D Convolutions

3D filters, though fantastic for their low parameter count, are relatively expensive in that a great many filters are required to make up for this very low count, making them compute- and memory-intensive.

An alternate approach we attempted was to train conventional 2D convolutional filters over the slices of the brain scan, without sharing parameters across slices. This was to ensure that each set of filters learned features relevant to its specific slice of the scan. Unfortunately, such custom forward passes are inconvenient to make efficient and require some fiddling with complex Einsums, and we felt that it was not worth the effort given the mixed results we had already witnessed with the 3D convolution approach. We did however go forward with the more inefficient route and note that the results were no better.

In both approaches, the failures boiled down to one the following

- Complete lack of convergence - constant loss throughout training
- Overfitting - decreasing loss over the train set but a constant test-set loss, despite the many regularization attempts – data augmentation, batch-normalization, dropout,

across many different hyperparameter settings.

Category Classification

The idea of treating this problem as one of noun classification is inherently flawed. Even if in principle a person's scans can be classified with great accuracy to represent one of the 60 nouns, this information is by and large useless, given the sheer number of nouns (noun concepts) in any human language; in the tens of thousands at least.

There is however some value in being able to classify scans into broad-sweeping categories, and we attempted these deep learning approaches to instead classify the scans into the very same categories as laid out by the Paper:

Artificial object, building, building part, tool, furniture, kitchen, vehicle, insect, vegetable, bodypart, clothing, animal.

These are somewhat encompassing categories, and being able to classify a scan into these categories can be argued is somewhat valuable.

However, as above, we found it difficult to train a model to effectively classify scans into the categories, since given the small size of the data, the models were quite susceptible to overfitting or not converging, despite extreme regularization.

We posit that the failure boils down to the fact that brain scans, though spatial, are fundamentally not visual in the sense that patterns of activation across regions do not resemble one-another at all geometrically, and convolutions are antithetical to settings where this is true. Instead, any effective model must instead learn region-specific patterns that contribute to the final answer.

Distributed Representations

Numerical representation of words in natural language has seen a lot of progress since the publication of the Paper. A popular method of representing a word in deep-NLP is to represent it as a real-valued vector in a large-dimensional space. The vector for a word, at least before the advent of the Transformer, was inferred from enormous natural language corpora, to great effect.

These learned vectors possess many interesting properties, such as the popular demonstration that in these vector spaces, granted that they are effectively “trained,”

Queen - King ~= Woman - Man

To a great extent, these vectors seem to find very meaningful arrangements in their space, with words of similar meaning/category huddling together. In these spaces one finds the months of the year huddled together for example, not very far from where the days of the week are huddled.

Many algorithms for forming these word-vectors exist, and many pre-trained dictionaries are open-source for public use. We decided to use the GloVe dictionary.

Linearity

The Paper makes the assumption that intermediate, latent aspects for concepts are composed linearly in neural activation, stating that such an assumption is fairly common in brain imaging literature.

We began by investigating the extent to which this somewhat lax assumption carries over into word-vectors, at least in these very specific settings.

Fixing our basis as the GloVe-50 representations of the same 25 verbs from the Paper, we found that an identical treatment to a linear sum of the basis scaled by the couccurrence coefficients results in a prediction with a non-confusion (the S function) accuracy of a remarkable 90% on unseen words.

Additionally, a reenactment of the paper with a learnable basis as opposed to a fixed basis showed that the accuracy was only marginally better at ~92%, justifying that just as with the intermediate representations for brain scans, a verb basis is remarkably good at composing into (strictly, *close to;* see Discussion) nouns.

For the rest of our work we decided to proceed with the fixed basis as opposed to the learned one seeing as how there is only a marginal difference in representational power.

Concept

With this toolkit of GloVe-50 vectors, we defined the goal of learning a model as such:

Given a brain scan as input, predict the coefficients with which the basis sum results in a prediction close to the actual word-vector of the noun corresponding to the input scan.

This approach theoretically extends the scope of prediction to neural activations corresponding to any semantic concept at all, as opposed to being tied down to only 60 nouns or 12 categories. This formulation can be thought of as a continuous map from brain-scan space to word-vector space which is arguably immensely valuable.

Approach

Our model is similar to that of the Paper with the crucial reversal that the basis is fixed and it is the coefficients that must be inferred. We took two fairly different approaches to this general idea for modelling the data, but both involve predicting the coefficients with which to make a linear combination of the fixed GloVe verb basis.

Importantly, the inputs are now, instead of 3D tensors, long vectors acquired by flattening the tensors. No data augmentation was implemented. This we maintain is completely acceptable since the Paper's authors took great care in making the scans precisely similar geometrically so a voxel at index [i, j, k] always represents the same brain region across scans.

Sparse Training

A la Paper, in each run, the model is trained on 58 of the 60 nouns, and a confusion test (S function from above) is taken against the remaining 2 to establish correctness. Naturally, since the task is of predicting the coefficients over a fairly small dataset, the model overfits eventually and we had to empirically find the sweet spot for early-stopping that results in good accuracy.

Dense training

Our distinction of “sparse” and “dense” has to do with the treatment of the data.

Whereas the sparse model is trained exclusively on the 58-split, this as mentioned above is conducive to overfitting especially given the linear (therefore simple) nature of the model and the small size of the dataset. With additional work of recombining the data and making the model semi-linear, a much more powerful model can be learned.

Data recombination

- Once the 58-split is made over brain scans, the targets are taken to be their corresponding GloVe-50 vectors.
- Two new batch-wide arrays, B1 and B2 are generated by randomly sampling from the 58 scan-target pairs, with repetition.
- B1 consists of X1 and Y1, the array of scans and the array of targets, similarly for B2.
- A batch-wide array R is generated, whose elements are scalars in [0, 1], to serve as combination ratios.
- The final Batch is B = (X, Y), where
    - X = X1 * R + X2 * (1 - R)
    - Y = Y1 * R + Y2 * (1 - R)
    - The operation * is the pairwise scaling a list of tensors by a list of scalars

This results in entire regions in the input and output spaces being represented and receptive to training, as opposed to individual (and very innumerous) points to be mapped pairwise. This can be compared to the motivation (and success) of the Variational Autoencoder (VAE). Although the latter is by design probabilistic, the general idea of using only finitely many data points to effectively map out the entire data distribution, is identical.

Here however, a lot of weight is put on the linearity assumption of word-vectors by not only translating them with one another but also scaling them, since the target to comb(scan1, scan2) now is comb(target1, target2). But since these word vectors are well-learned, a charitable interpretation allows that word-vector interpolation may indeed correspond, even if vaguely, to semantic interpolation.

Model

The model here, to account for the increased data complexity, was made semi-linear. That is, while the prediction is still the linear combination of the fixed basis vectors scaled by the predicted coefficients, the prediction of the coefficients itself is no longer linear, but a result of two ReLU layers and a Sigmoid output layer, whose output is rescaled to sum to 1.

It is to be noted here that despite this inclusion, the model ultimately has *fewer* parameters than the Paper and our Convolutional models.

Results

One distinction here is that instead of Cosine-similarity, the S function here is the negative of the Euclidean distance between prediction and output. In word-vector space, it isn't enough for two vectors to only point in the same direction to be deemed similar. Closeness is also required.

With the original sparsely-trained model, accuracy on unseen words along the same S function as above was recorded in the 61-65% range over many runs, with 500 58-splits each run.

With the densely-trained semi-linear model, accuracy was much higher, in the 78-80% range over many runs of 500 58-splits.

We could not measure the accuracy for each of the ${}^{60}C_2$ 58-splits due to limited compute. The Dense model needed magnitudes more training time to converge given the complexity of the spaces being mapped.

Significance of results

Holding that an arbitrary model would be correct 50% of the time at the limit, we calculate the improbability of a model's being correct in at least 80.2% of 500 experiments, the highest accuracy recorded in our runs. This is a simple binomial setting, and the value turns out to be $p < 2.5 \times 10^{-44}$. Our p-value estimation method differs from that of the Paper where they establish it empirically. Nevertheless, an accuracy of ~80% is extremely significant given the theoretical limit of ~90% achieved by complete-information predictions using the cooccurrence coefficients. We consider our work a great success.

Discussion

Critique of classification as the right approach

As we have stated above, there is limited if any value in classifying scans into noun classes. There are very many nouns in any human language and at least as many noun-like concepts a mind may conceive of therefore activation patterns a human brain may exhibit. Treating them orthogonally is dismissing the shared, overlapping information implicit in the activations.

If the Paper had similarly treated the data as independent X-Y maps, there would be no basis with which to discover latent similarities between neural activations and how these similarities correspond to semantic similarities.

There is some value in rather classifying the scans into categories, although again a human concept never really exclusively belongs to one category or the other. An insect is an animal, clothing is artificial, and a dining table is furniture while also having to do with the kitchen.

Besides, the fewer the categories into which to classify, the more vacuous the distinction, and the more numerous, the more "overlap" information is lost or rather, discarded by design.

Limits

The measure for "accuracy," both in the Paper and identically in our work, is somewhat lax. It is not a very strong statement to make, that a predicted output is closer to the true output than to an arbitrary (statistically independent) other vector from a similar, but strictly different distribution. We defend this limitation in the section Future Avenues.

Additionally, as mentioned above, we could not measure across all possible 58-splits due to limited compute. However, 500 is a big enough fraction of 1770 as to be representative.

Future Avenues

As mentioned in Limits under Discussion, the measure of accuracy of prediction is somewhat weak. It can be interpreted, with some leeway, as “the prediction is closer to the true vector than most other vectors are.” However, the volume swept by the same Euclidean distance grows polynomially with the dimensionality of a space, meaning a greater potential for other vectors to be within this hypersphere, therefore closer. Since word-vectors often occupy very large-dimensional spaces, the statement starts to gain significance: If despite this great potential for false competitors the prediction is closer than most, the prediction is valuable.

We propose a hypothetical method with which, given time-series scans of a person’s train of thought, a significant amount of semantic inference can be made. This comes with some assumptions which we will later discuss.

Proposed Method

Let the time-series scans be made at discrete time-steps with corresponding scans s_1, … s_n.

With our model as *m*, the brain-scan space as *B*, and the word-vector space as *W*, $m: B \rightarrow W$.

At each time-step i, a prediction p_i is made as $m(s_i)$.

Thus having mapped the series of scans to a series of vectors in *W*, we acquire a predicted time-series of word representations of the scan series.

However, the model is fallible at any one time-step since it is not perfect. At any one time-step, there is a large subset of *W* that could be a potential match to the corresponding word-prediction p_i.

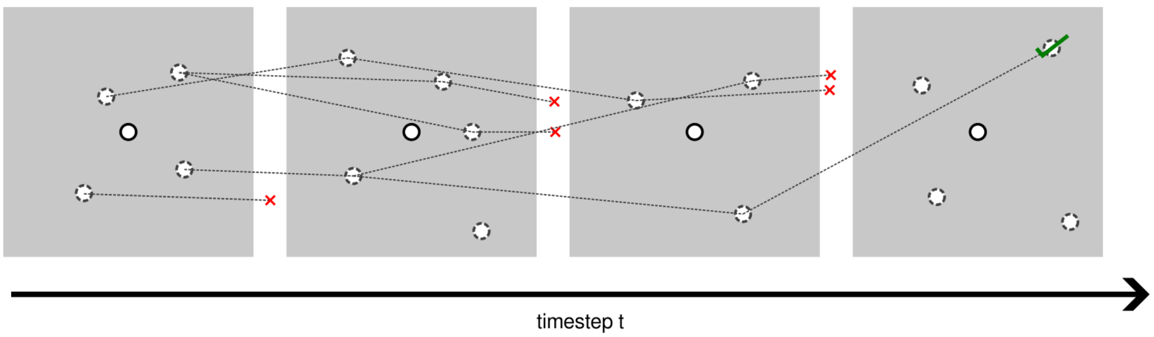
With time-series of any reasonable length, this problem boils down to a multiple-constraint problem with only one solution in the vast majority of cases – the natural-language interpretation/sentence. In this setting, the more constraints available, the easier it is to zero in on good solutions.

However the measure of the real-ness, that is the plausibility, of an arbitrarily constructed sentence is quite difficult to define.

A solution is to train a new sequence-to-sequence model that inputs these predicted time-series, as the Generator in a Generator-Adversarial Network (GAN), where the Discriminator's objective is to distinguish the Generator's output sentences from sentences sampled from natural language corpora. It is common in GAN settings to train against multiple losses. The second loss here besides the adversarial loss could be the distance from nearest candidates at each timestep, so that the generator's outputs are not just plausible sentences (mode collapse is common in GAN training) but in fact relevant to the inputs.

The Generator in this way could learn to map these close-but-not-quite time-series to perfectly plausible natural language sentences.

An illustration of a hypothetical inference process over word-vectors in R2. As the time-series progresses, the many branches towards plausible sentences over candidate word-vectors dwindle rapidly, and the partial inaccuracies are corrected for by this reduction in possible outcomes. The task of determining the plausibility of a sentence can be delegated to a GAN distinguishing real sentences from these generated sentences.


○ Model's prediction
◌ Candidate real word-vectors
✕ No continuation of sentence
✓ End of a plausible sentence
┈ Possible continuation
timestep t

Limitations

Of course this operates under a lot of assumptions and we put it forth more as a thought experiment in the direction that this work is useful than as a practical proposition necessarily.

The weakest assumption is that the internal abstractions for a complex thought process are identically linear to the word-sequence that is vocalized. Chomsky, whenever prompted to speak about language models, talks about how there is nothing at all linear about how language is perceived, as opposed to spoken.

Another is that complex trains of thought, seen through an MRI machine, can be neatly distinguished into discrete timesteps. The obvious expected behaviour of the brain is that semantically distinguishable activation patterns overlap heavily across time, making delineating boundaries an entire new dimension of difficulty in itself. However one could postulate that these brain-states may interpolate continuously from distinguishable pattern to distinguishable pattern, allowing the inference itself to be just as continuous, like the semantic interpolation imagined under Data recombination.

There is also the assumption of the existence of or even the possibility of collecting such data. Functional imaging has come a long way since the Paper was published, but acquiring such data would require either asking the participants to entertain certain predetermined thoughts during the imaging, or asking them to later recollect what they were thinking about when the imaging was underway, both of which would raise strong questions about fidelity to real, organic trains of thought, or correctness of recollection.

Additionally, training only on nouns, we can not make the claim that the latent features learned also extend to all other aspects of language, nor that a verb-basis is just as apt for them, without a strong foothold in neurolinguistic literature.

However, this last point can be ameliorated by collecting more fMRI data a la Paper for adjectives, adverbs, and so on, and training a larger model. This is an explosion in the complexity of data required but not a conceptual failing in itself.

Closing

We have established that it is indeed possible to learn a map from brain scans to word-vectors, with great accuracy.

However, the data is quite old, and its acquisition and nature highly idealised. It can not be claimed outright that being able to predict nouns from brains that were expressly instructed to think about the nouns, leads to being able to infer any meaning from brains acting in the real world, where they are constrained by and concerned with countless simultaneous considerations. It is entirely possible that the same brain may exhibit a given noun entirely differently in different environments, as opposed to a white laboratory vacuum. To this extent, the significance of the results is left to the charitability of the reader, but little more is asked of the reader than by Paper.

Our discovery that the word-vector basis behaves much like the latent semantic activation basis points in the general direction that there are indeed grounds on which to stand, for any attempts to try and decipher the brain from the NLP perspective.

Work of this nature raises concerns about the possibility of privacy violation of the most fundamental kind, and enforceability of thought-crime, but there is also huge medical value in trying to map intention to action for those debilitated by any of the countless neurophysiological afflictions a human body may suffer. One must pick and choose the merits of science.