Exploring Neurons in Multi-Modal Language Models

Sam Acquaviva **Lowell Hensgen**

Abstract

We apply a procedure that explains the compositional role of individual neurons in deep neural networks to multiple new domains. First, we apply the procedure to a multi-modal neural network. We analyze whether the trends found in single modality neural networks are the same as the trends in the dual modality neural network on the Natural Language Inference task. We modify the procedure to take into account the semantic similarity of words when encoding concepts, and analyze how this difference in encoding changes how well we can explain the role of individual neurons.

1 Introduction

Despite the large successes of deep neural networks (DNNs) in computer vision and natural language processing (NLP), they still largely operate as black-boxes. By analyzing the role of individual neurons in these models, we can better understand DNNs, increase trust in DNNs' decisions, and assist DNNs in making ethical decisions (Lipton, 2016). There is extensive work to explain the role of specific neurons in language models (Dalvi et al., 2018; Mu and Andreas, 2020) and in image classification/generation tasks (Zeiler and Fergus, 2013; Zhou et al., 2017, 2015). However, to our knowledge, there has not been any work to explore the role of individual neurons in multi-modal networks that connect text and images. We investigate neurons in the Contrastive Language–Image Pre-training model (CLIP) (Radford et al., 2021), a model that can encode and compare both text and images. We answer whether the model's unique learning objective to link images and text causes its neuronal representations to largely differ from other language models. More specifically, we extend the techniques of (Mu and Andreas, 2020) to CLIP by fine-tuning CLIP's text embedding model on the Natural Language Inference (NLI) task and comparing individual neurons' alignments with concepts to the alignments found in the original paper. We also find that encoding the semantic meaning of words in that word's concept changes the relationship between neuron explainability and firing accuracy.

2 Related Work

2.1 Explaining Role of Neurons

A lot of preexisting work explains how individual neurons represent linguistic concepts in language models (Dalvi et al., 2018; Durrani et al., 2020; Karpathy et al., 2015; Mu and Andreas, 2020). In (Dalvi et al., 2018), the authors present two techniques for analyzing the role of individual neurons: *Linguistic Correlation Analysis* which is a supervised method for finding the most important neuron for a given extrinsic task, and *Cross-model Correlation Analysis* which is an unsupervised method for finding the most important neurons for the model itself. However, the usefulness of these techniques is limited because the techniques rely on manual inspection of the outputs of the analyses.

One difficulty in automating the explanation of concepts represented by neurons is determining what concepts to search for. With a simple list of coarse concepts (e.g. verb, water), many concepts will be distributed widely across neurons (Fong and Vedaldi, 2018). Perhaps more fine-grained, complex concepts are needed to explain individual neurons' behavior. (Mu and Andreas, 2020) addresses this difficulty by starting with a list of individual concepts (e.g. verb, water), then using a compositional search combining different concepts with logical operators to build more complex concepts. The procedure uses NOT, AND, and OR logical operators to produce concepts in logical forms. It performs a beam search with $B = 10$, building upon the 10 most descriptive neurons for a given formula

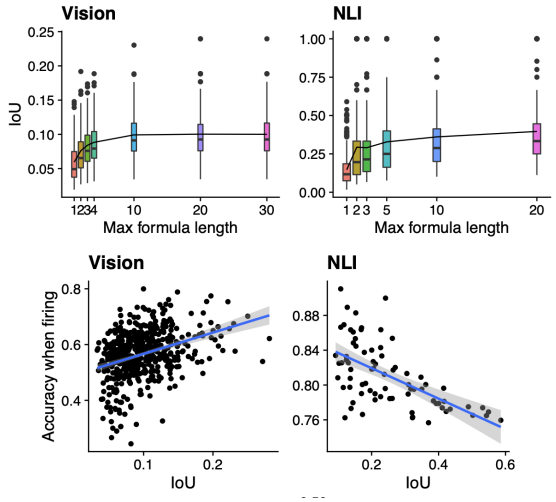

Figure 1: From (Mu and Andreas, 2020), **top:** as the complexity of the concept increases, so does its match with neuron activations. **bottom:** in the vision model, accuracy increases when neurons that we can explain fire. In the language model, accuracy decreases when neurons that we can explain fire.

length (using the IOU metric described in 3 and narrowing down to the ten best formulas of the next length. In the original paper, they analyze a vision model for image classification and an NLP model for Natural Language Inference (NLI) tasks. NLI is the task of determining whether, given a *premise*, a *hypothesis* is true (entailment), false (contradiction), or undetermined (neutral). For example, the premise, "The man plays soccer" contradicts the hypothesis "The man is sleeping".

We are interested in two of the questions that (Mu and Andreas, 2020) answers:

1. Do neurons learn compositional concepts?
2. Do interpretable neurons contribute to model accuracy?

They find that, yes, neurons learn compositional concepts (see Figure 1, top). They also find that, in the vision model, the neurons that we are better able to explain are more accurate, and in the language model, the neurons that we are better able to explain are less accurate (see Figure 1, bottom). The meaning of the metrics in Figure 1 are explained in 3.

2.2 CLIP

CLIP is pre-trained on text-image pairs, and learns both a text encoder and image encoder as illustrated in Figure 2. CLIP's training objective is to maximize the similarity between its embedding of an

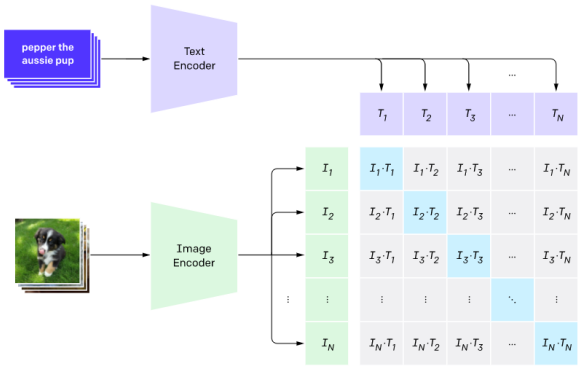

Figure 2: From (Radford et al., 2021), this diagram shows the pre-training architecture of CLIP.

image and its embedding of the text paired with that image. This form of supervision aligns the text and image representations, and makes CLIP especially useful for zero-shot image classification, Across a suite of 27 datasets measuring tasks such as fine-grained object classification, optical character recognition (OCR), activity recognition in videos, and geo-localization. CLIP's zero-shot prediction outperformed the state-of-the-art Noisy Student EfficientNet-L2 on average across 27 datasets (Radford et al., 2021). However, to our knowledge, CLIP's textual embeddings have only been used to help encode the task, and have not been fine-tuned for typical NLP tasks such as NLI.

3 Evaluation Metrics

Our two main evaluation objectives are those in Figure 1, which answer the two questions asked by (Mu and Andreas, 2020) in 2.

The match between a concept and a neuron can be represented by the Intersection over Union (IoU) between the mask of the neuron and the mask of the concept, where $IoU(A, B) = \frac{|A \cap B|}{|A \cup B|}$ (see Figure 3). For neurons in the final layer of the network, we calculate the IoU of each neuron and the concept it best aligns with. We repeat this for many formula lengths to see if the positive correlation between IoU and concept complexity found in (Mu and Andreas, 2020) holds in CLIP.

To compute how the accuracy of a model correlates with the explainability of its neurons, we calculate the IoU between a neuron and its best concept as described in Figure 3. Then, we calculate the performance of the model on the task when each neuron fires past some threshold.

4 Methods

We answer whether the two metrics mentioned – correlation between neuron and concept complexity, correlation between neuron accuracy and explainability – hold in the following scenarios:

1. When we use a different model (CLIP) instead of Resnet-18 and BiLSTM for image and NLP tasks, respectively.

2. When we change the encoding of concepts to be more complex.

4.1 Using CLIP Instead of BiLSTM and Resnet

We directly replicate the techniques of (Mu and Andreas, 2020) for NLI. (Mu and Andreas, 2020) used the BiLSTM architecture from (Bowman et al., 2016) to encode the premise and hypothesis. Then, these representations are concatenated along with their element-wise product and difference, and fed to a Multi-layer Perceptron (MLP) with a softmax layer. We use the same architecture, but use the CLIP text encoder instead of a BiLSTM and only train the weights of the MLP. Using these new models in both domains, we see if the findings of (Mu and Andreas, 2020) mentioned in 2 and shown in Figure 1 hold.

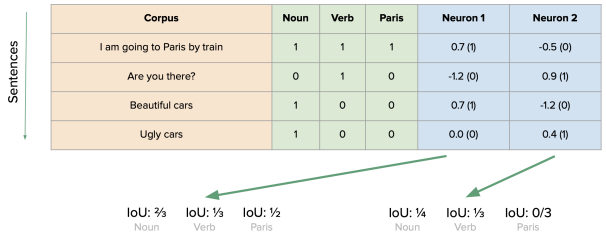

Figure 3: An example of masks for linguistic concepts (green) and neurons (blue), and the Intersection over Union (IoU) between these masks. The 1s and 0s represent the masks. For a concept, 1 indicates the concept is in the sentence; for a neuron, (1) indicates that the neuron activation to the sentence passed some threshold (0 in this example). Looking specifically at the IoU between Neuron 1 and the Noun concept, you can see that the intersection of the masks is 2 because the concept exists *and* the neuron activates passed the threshold for the sentences "I am going to Paris by train" and "Beautiful cars". Similarly, the union is 3 because the concept exists *or* the neuron is activated enough for the same two sentences and the sentence "Ugly cars" (Noun exists in the sentence, but Neuron 1 does not activate). So, the IoU between Neuron 1 and Noun is $\frac{2}{3}$.

4.2 Alternative Encoding of Concepts

We also explore how different encodings of concepts in neurons affect the findings in Figure 1. In (Mu and Andreas, 2020), the encoding scheme of individual concepts is relatively simple. We try an alternative encoding scheme. The original encoding scheme only encodes logical compositions of given word concepts (e.g. "Paris AND France") as in a sentence if that exact word appears in the sentence (see Figure 3). Instead, we use WordNet (Miller, 1995), a large database containing encodings of semantic relations between words, to add the synonyms of a word to a concept mask before computing IOU. This encoding scheme is motivated by the fact that the original encoding scheme does not capture any of the semantic similarity between words. For example, under the encoding scheme from (Mu and Andreas, 2020), the concept "water" will not be encoded in the sentence "I swam in a lake", although it seems that the two are related. With the new encoding, if lake and water are synonyms the concept water will now be related to the original sentence. We encode part-of-speech in the same way as the original encoding scheme.

5 Results

5.1 Re-implementation using CLIP

We used CLIP's text encoder to create a latent representation of the premise and hypothesis, then we trained an MLP with a final softmax layer on the Stanford NLI dataset. For each pre-trained weight in the final layer of the MLP, we compute accuracy when that neuron fires past some threshold.

First, we analyze the results using the IOU concept encoding as described by (Mu and Andreas, 2020). As seen in Figure 4, when using a formula length of one and CLIP as the encoder, the negative accuracy-when-firing and IOU correlation was statistically significant. When using our re-implementation of the BiLSTM, it is not. Using a formula length of two, as seen in 5, the correlation between accuracy when firing and IoU becomes more negative in both CLIP and the baseline BiLSTM. The joint text-image training objective increased the accuracy for the NLI task, but didn't change the trend in neuron explainability much.

We also validate that neuron-concept IoU increases as the max formula length increases. Using CLIP as the encoder, the average IoU with a formula length of 1 was 0.19, and the average formula

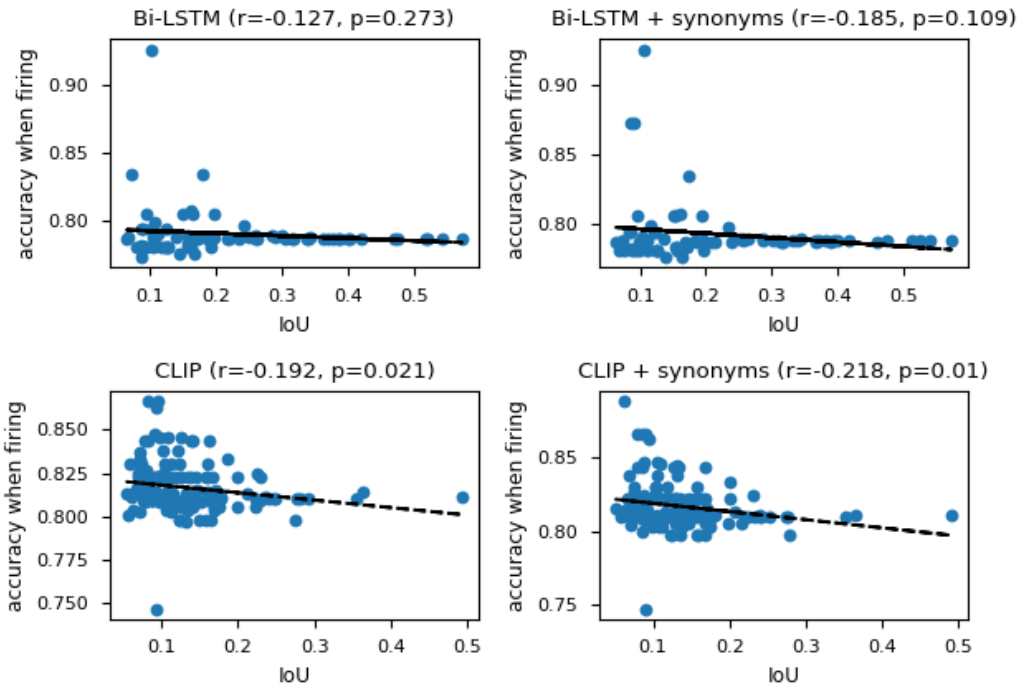

Figure 4: Formula Length One Results. **top:** the relationship between neuron explainability and model accuracy when using the BiLSTM with formula length one. **bottom:** the relationship between neuron explainability and model accuracy when using CLIP. **left:** Concept encoding scheme from (Mu and Andreas, 2020). **right:** Our concept encoding scheme that includes synonyms of concepts.

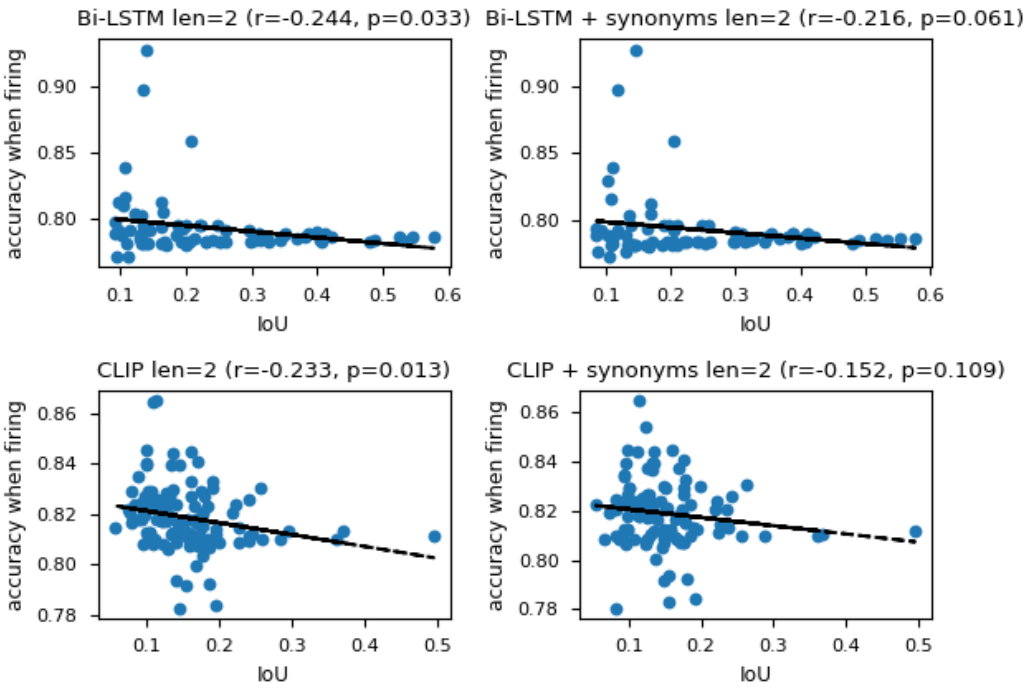

Figure 5: Formula Length Two Results. **top:** the relationship between max formula length and neuron IoU when using the concept encoding scheme from (Mu and Andreas, 2020) for classification on different datasets. **bottom:** the relationship between max formula length and neuron IoU when using the Word2Vec concept encoding scheme for classification on different datasets.

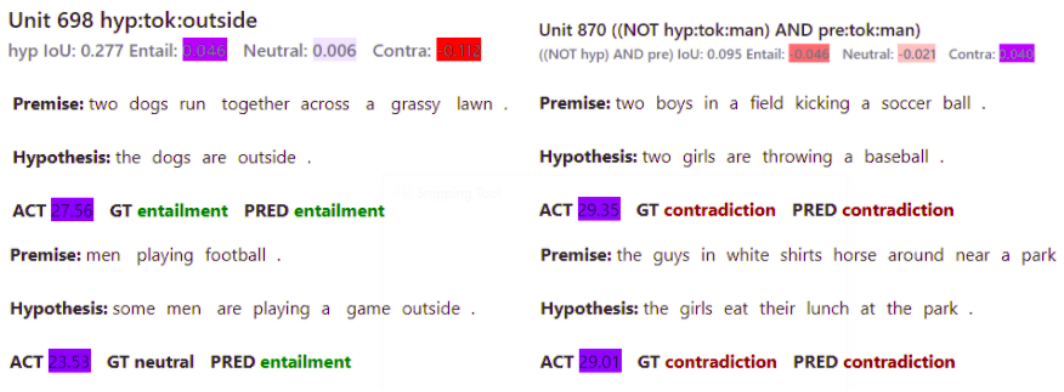

Figure 6: **left:** Example high IOU neuron with formula length one, best concept found is hyp:tok:outside meaning outside is in the hypothesis. **right:** Example neuron with formula length two, with the best length two formula found indicating NOT man in the hypothesis AND man in the premise. Included with each neuron are two test examples where the neuron has high activation along with the ground truth (GT) and model prediction (PRED).

length with a formula length of 2 was 0.21. This is in line with 1.

5.2 New Concept Encoding

We analyzed the WordNet-aware concept encoding described in 4 for both the Bi-LSTM encoder and CLIP encoder with a formula length of 1 and 2. With a formula length of 1, the trend between accuracy-when-firing and IoU decreases when using our alternative concept encoding, as shown in 4. With a formula length of 2, the opposite happens: the trend becomes less negative, as shown in 5.

However, the average IoU of a neuron's top concept did not change much. The mean difference between a neuron's IoU with its top concept when using or not using the alternative encoding is 0.0014. This similarity suggests that embedding the semantic information of a word in the sentence does not make a large difference in explaining that neuron's firing behavior.

5.3 Individual Neuron Qualitative Analysis

In 6, we show examples of two neurons from our CLIP models with normal IOU concept encoding. In neuron 698, we see evidence for why more complex neurons are necessary to solve the NLI task. If the neuron really did only encode the concept "outside", it would always highly weight the word outside in the hypothesis to being entailment. The word "outside" should have no relation to the entailment of the hypothesis, so the model would have much lower accuracy.

In the right neuron of 6 we see Unit 870: a neuron that has its highest IoU with an intuitive formula of length 2. Having a masculine noun (man, boys) in the premise and not having one in the hypothesis could be indicative of subjects being changed in the sentences and a contradiction. But this heuristic isn't in the spirit of the task – the neuron doesn't seem to be actually matching up subjects, just checking for the presence of men and their subsequent absence. Thus, it learns a simple heuristic that can effectively predict contradictions in short, simple sentences used by the dataset. Probing this neuron's highest-IoU points to reasonable adversarial examples: the premise-hypothesis pair "the boys said hello to the girls" and "the girls said hello" would likely be weighted towards contradiction by Unit 870.

Perhaps this also gives some insight into why more explainable neurons are less accurate. If a neuron can be easily explained by a formula, maybe it learns some heuristic of the training dataset which is less general than neurons we can't explain with a formula.

6 Conclusion

We extended past research in the field of understanding individual neurons in Natural Language Processing to multi-modal models. Specifically, we showed that

1. Interpretability of neurons in CLIP's text encoder are negatively correlated with accuracy, across many tasks.

2. Using a more sophisticated encoding of concepts has little effect of interpretability of neurons.

This suggests that using the unique joint learning objective between images and text maintains

trends in neuron interpretability. This also suggests that results of (Mu and Andreas, 2020) might be general in NLI – explainability of neurons in NLP tasks anti-correlated with accuracy of the model, regardless of the specific encoder. The robustness of results when using a word and its synonyms versus a single word suggests that encoding the semantic information of a word in the concept is not necessary to explain a neuron's firing activity. Neuron interpretability could be independent of the concept calculation, as long as the concept is encoded reasonably.

Future research could further explore encoding concepts in natural language sentences to improve neuron explainability. Additionally, we, like (Mu and Andreas, 2020), only analyzed the neurons at the output layer. Future research could analyze how neurons in earlier layers of the MLP encode compositional concepts. Another natural extension of our work is to explore how other pre-trained language models represent concepts in neurons.

References

Samuel R. Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D. Manning, and Christopher Potts. 2016. A fast unified model for parsing and sentence understanding. *CoRR*, abs/1603.06021.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James R. Glass. 2018. What is one grain of sand in the desert? analyzing individual neurons in deep NLP models. *CoRR*, abs/1812.09355.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. *CoRR*, abs/2010.02695.

Ruth Fong and Andrea Vedaldi. 2018. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. *CoRR*, abs/1801.03454.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *CoRR*, abs/1506.02078.

Zachary Chase Lipton. 2016. The mythos of model interpretability. *CoRR*, abs/1606.03490.

George Miller. 1995. Wordnet: a lexical database for english. *Princeton University*, 38.

Jesse Mu and Jacob Andreas. 2020. Compositional explanations of neurons. *CoRR*, abs/2006.14032.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Matthew D. Zeiler and Rob Fergus. 2013. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901.

Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. 2017. Interpreting deep visual representations via network dissection. *CoRR*, abs/1711.05611.

Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. 2015. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150.