Interpreting Neural Networks Using Minimal Training Data

Josue Torres-Fonseca[1], Jacob Andreas[2], Evan Hernandez[2]
Boise State University[1], Massachusetts Institute of Technology[2]

Abstract

Today's machine learning models are seen as black boxes taking in input and producing output without much knowledge of its internal workings. This has led to problems such as bias and overly complex models. The Language and Intelligence Group at MIT has proposed a model (called MILAN, for mutual-information-guided linguistic annotation of neurons) that, given a neuron, generates a description by searching for a natural language string that best describes image regions in which the neuron is active. Even though this model works well it requires too much training data to be useful for other applications. My research aims to fix this problem by creating a new implementation of MILAN which generates sentences from the WordNet corpus and picks the best sentence using CLIP (Contrastive Language–Image Pre-training). CLIP does not require any training data to complete this task instead only requiring a list of possible words used to describe the images being analyzed. Results show that with less training data this new implementation of MILAN can accurately annotate many of the neurons however neurons looking at more abstract concepts are unsuccessfully annotated. Possibly suggesting that MILAN needs access to both the image regions and the entire image.

Introduction

We know how neural networks learn but not what they are learning

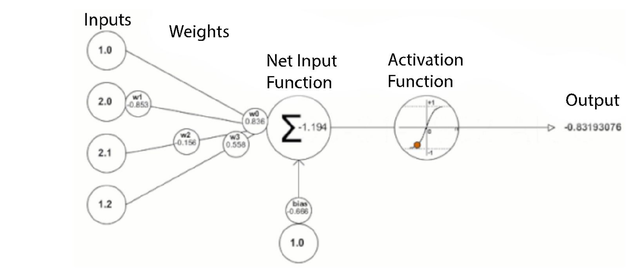


Figure 1: Neuron's inner workings

MILAN was built to describe the roles of individual neurons in a neural network meant to analyze images

For each neuron MILAN can find image regions which cause the neuron to activate the most and produce fine grained descriptions for these image regions

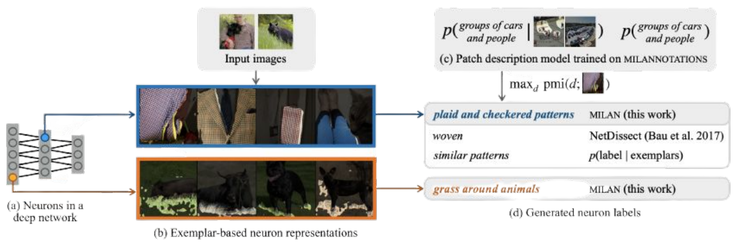


Figure 2: Milan pipeline

This was done using a machine learning network requiring 20k human generated image captions across 7 models with different network architectures, datasets, and tasks

Goal: Find a way to captions these images without using Milan Annotations Dataset

Motivation

Many machine learning models today are trained on data from the Internet

Unfortunately, much of the data from the Internet is biased leading to problems where machine learning models are inadvertently racist, sexist, etc.

MILAN provides a way to find bias in machine learning models

Methods

To reduce the need of training data we produce descriptions from the WordNet corpus. To provide ground truth for the model we use CLIP

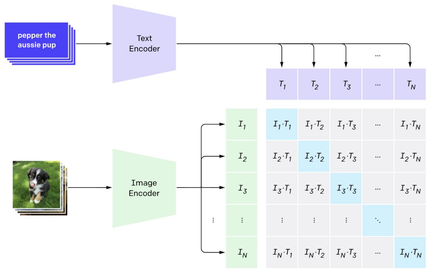
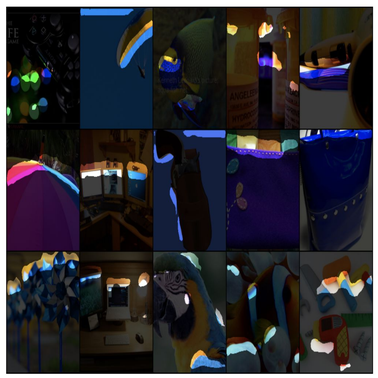


Figure 3: Clip encoding a description and an image and comparing them

CLIP originally looks at the whole image, however we modified CLIP so that it can generate scores for descriptions only given the regions of an image

Two-word descriptions were generated by first passing into CLIP all words from WordNet that exist in a database of concreteness scores. Then the top 32 nouns were found using CLIP and then commonly used prepositions were prepended to these nouns and the resulting prepositional phrases were ranked



Human Annotations:
- Blue edges
- Blue and orange areas

Predicted Caption: At Lifestyle

Human Annotations:
- Jack-o-lanterns
- The brightest parts of the images

Predicted Caption: After September

Results

The new pipeline performs worse than NetDissect and MILAN with a precision of 0.191, recall of 0.104, and f-score of 0.134. While NetDissect has a f-score of 0.24 and MILAN achieved a f-score of 0.38. Although arguably some of the predicted captions describe the pattern better than the human annotations.

However, with AlexNet which contains images that require spatial reasoning it struggles on ideas such as "above text or horizontal lines on top of building"

Model	CE	ND	$p(d \mid E)$	$\mathrm{pmi}(d; E)$
AlexNet-ImageNet	.01	.24	.34	**.38**
AlexNet-Places	.02	.21	.31	**.37**
ResNet-ImageNet	.01	.25	.27	**.35**
ResNet-Places	.03	.22	.30	**.31**

Figure 5: F-Scores from Bert by model

Conclusion

Human Annotations:
- Stripes
- Images with Stripes

Predicted Caption: Off Wearing

The new machine learning pipeline can create two-word descriptions for the neurons using less data by using CLIP but struggles to create descriptions for abstract ideas/descriptions (e.g. spatial reasoning)

Future work involves looking deeper into these cases and more thoroughly testing what CLIP is capable of

References & Acknowledgements

- David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In Computer Vision and Pattern Recognition (CVPR), 2017.
- Jesse Mu and Jacob Andreas. Compositional explanations of neurons. In Advances in Neural Information Processing Systems, 2020.
- Radford, A., Kim, J., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I.. (2021). Learning Transferable Visual Models From Natural Language Supervision.
- Hernandez, E., Schwettmann, S., Bau, D., Bagashvili, T., Torralba, A., & Andreas, J. (2022). Natural Language Descriptions of Deep Visual Features. In the Eleventh International Conference on Learning Representations (ICLR 2022).

Thank you to Jacob Andreas' lab for helping me with my research here at the Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory. Also thank you so much to the MSRP staff for giving me the opportunity and funding necessary to be at MIT. Most of all thank you to the MSRP cohort. You are all an amazing group of people and I hope the best for all of you.