Deep Dream for LLM’s

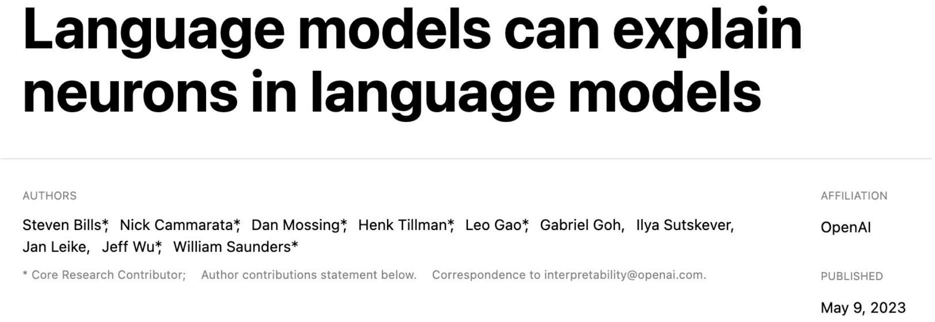
By Peter Chatain and Scott Viteri

Deep Dream for LLM’s

Deep Dream ->

Motivation:

1. We want sentences that describe a neuron’s behavior
2. Naive deep dream produces gibberish
   a. True for **images**, and for **text**
3. Find a regularization technique that constrains to recognizable sentences.
4. GPT-N uses recognizable sentences to produce explanations



Language models can explain neurons in language models

AUTHORS

Steven Bills*, Nick Cammarata*, Dan Mossing*, Henk Tillman*, Leo Gao*, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu*, William Saunders*

* Core Research Contributor; Author contributions statement below. Correspondence to interpretability@openai.com.

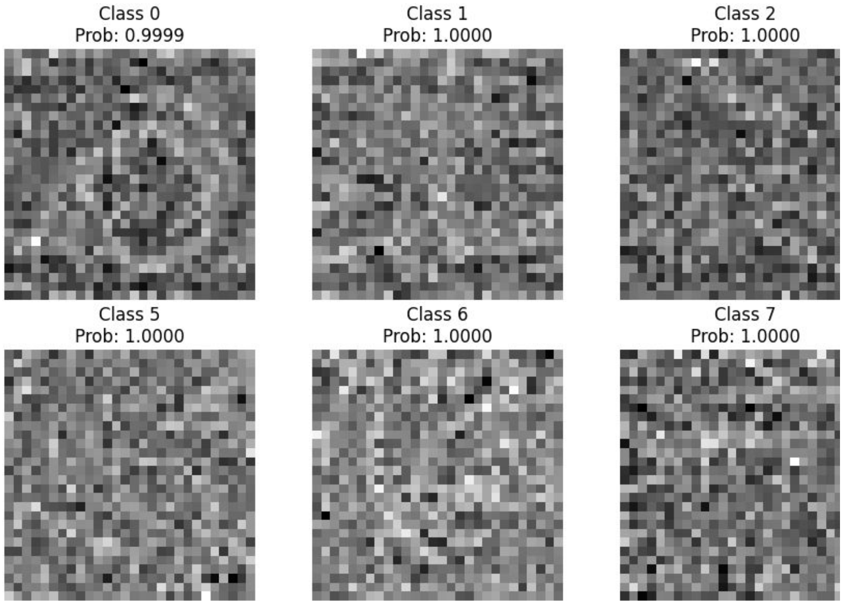
AFFILIATION

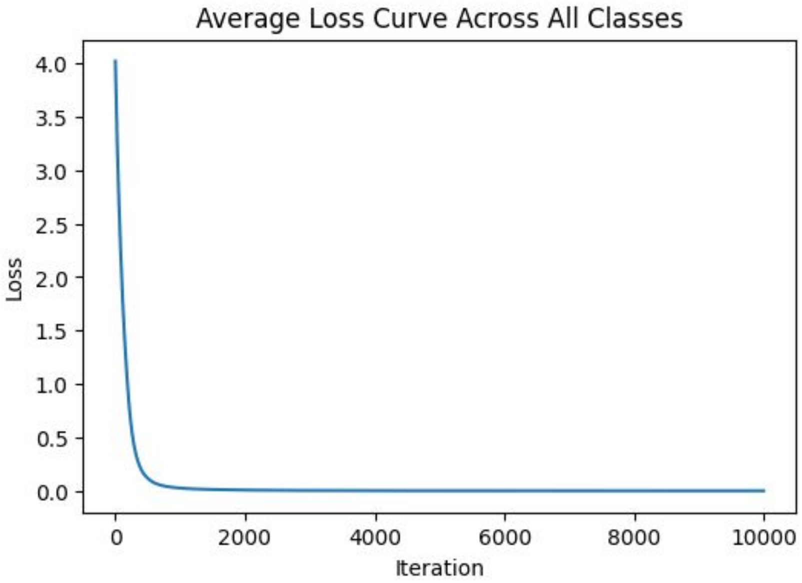
OpenAI

PUBLISHED

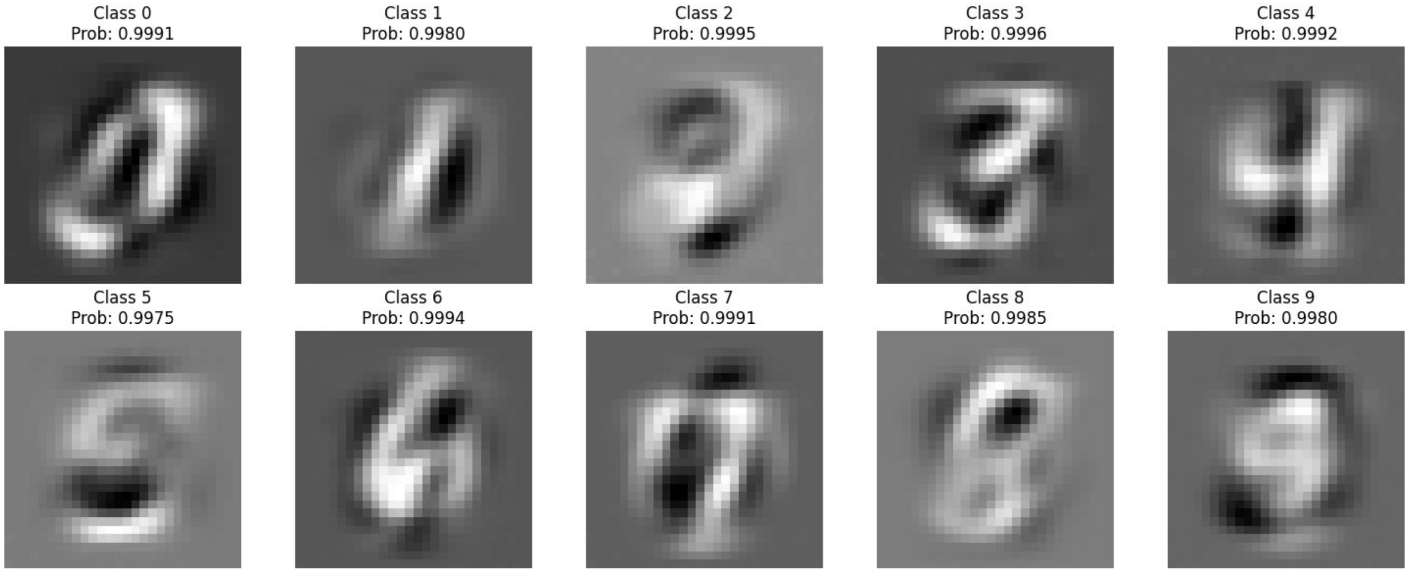
May 9, 2023

Problem

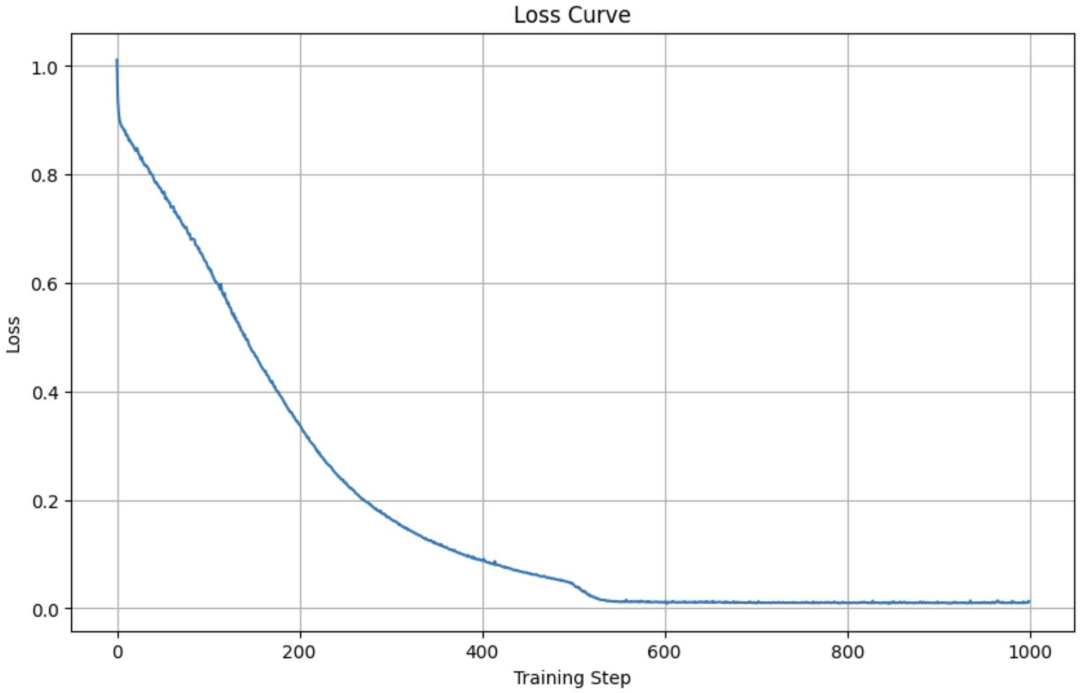

Class 0
Prob: 0.9999
Class 1
Prob: 1.0000
Class 2
Prob: 1.0000
Class 5
Prob: 1.0000
Class 6
Prob: 1.0000
Class 7
Prob: 1.0000


Average Loss Curve Across All Classes
Loss
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
2000
4000
6000
8000
10000
Iteration

Solution


Class 0
Prob: 0.9991
Class 1
Prob: 0.9980
Class 2
Prob: 0.9995
Class 3
Prob: 0.9996
Class 4
Prob: 0.9992
Class 5
Prob: 0.9975
Class 6
Prob: 0.9994
Class 7
Prob: 0.9991
Class 8
Prob: 0.9985
Class 9
Prob: 0.9980

Auto-encoder loss



Future Work

1. Training better autoencoders
2. Placing the auto-encoder in intermediate layers, not just after the first embedding layer.
3. Instead of searching with gradient descent, search over tokens restricted to only the most plausible tokens as judged by the language model itself