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

Introduction

- language model의 각각의 neuron들이 어떠한 feature를 represent 하는지 파악하고 싶으나, neuron의 수가 엄청나게 많은 만큼 모두 분석하는 것이 어렵다.

- 해당 논문은 language model로 각각의 neuron들이 어떠한 역할을 하는지 human-friendly하게 설명하는 방법을 제시한다.

- 또한 language model이 생성한 설명을 정량적으로 평가할 수 있게 하여, 설명이 얼마나 정확한지 파악할 수 있는 방법 또한 제시한다.

Overview

- 다음과 같은 세팅에서 진행함:
  - **Subject Model**: neuron의 behavior를 해석 하고 싶은 모델이다. 해당 논문에서는 GPT-2 XL을 사용하였고, MLP layer의 neuron만 해석한다.
  - **Explainer Model**: 활성화된 neuron에 대해 해석을 하고 human-friendly한 설명을 생성하는 모델이다. 해당 논문에서는 GPT-4를 사용하였다.
  - **Simulator Model**: Explainer Model이 생성한 설명을 바탕으로 어떠한 neuron이 활성화 되어있는지 예측하는 모델이다. 해당 논문에서는 GPT-4를 사용하였다.

- 다음과 같은 알고리즘으로 진행함:
  - **Step 1**: Explainer Model로 neuron의 behavior에 대한 설명 생성
  - **Step 2**: 생성한 설명을 바탕으로 Simulator Model이 어떠한 neuron이 활성화 되는지 예측함
  - **Step 3**: 예측된 결과와 실제 활성화된 neuron을 비교하여 점수를 산출함

Methodology

Step 1

- Explainer Model에 입력되는 프롬프트에 <neuron number>와 {<token>, <activation value>} pair들을 주어 해당 neuron의 behavior를 설명하도록 함

<system prompt>

```
We're studying neurons in a neural network. Each neuron looks for some particular thing in a short document. Look at the parts of the document the neuron activates for and summarize in a single sentence what the neuron is looking for. Don't list examples of words. The activation format is token<tab>activation. Activation values range from 0 to 10. A neuron finding what it's looking for is represented by a non-zero activation value. The higher the activation value, the stronger the match.
```

+

<demonstrations>

```
Neuron 1
Activations:
<start>
the 0
sense 0
of 0
together 3
ness 7
in 0
our 0
town 1
is 0
strong 0
. 0
<end>
<start>
[prompt truncated …]
<end>

Explanation of neuron 1 behavior: the main thing this neuron does is find phrases related to community
```

+

<user prompt>

```
Neuron 4
Activations:
<start>
Esc 0
aping 9
the 4
studio 0
, 0
Pic 0
col 0
i 0
is 0
warmly 0
affecting 3
<end>

Explanation of neuron 4 behavior: the main thing this neuron does is find
```

Methodology

Step 2

- Simulator Model에게 Explainer Model이 생성한 설명과 <neuron number>와 <token>들을 주어 <token>들의 activation 활성화 값을 예측하도록 함

<system prompt>

```
We're studying neurons in a neural network. Each neuron looks for some particular thing in a short document. Look at an explanation of what the neuron does, and try to predict how it will fire on each token. The activation format is token<tab>activation, activations go from 0 to 10, "unknown" indicates an unknown activation. Most activations will be 0.
```

+

<demonstrations>

```
Neuron 1
Explanation of neuron 1 behavior: the main thing this neuron does is find phrases related to community
Activations:
<start>
the  unknown
sense unknown
of 0
together 3
ness 7
in 0
our 0
town 1
is 0
strong 0
. 0
<end>
<start>
[prompt truncated …]
```

+

<user prompt>

```
Neuron 4
Explanation of neuron 4 behavior: the main thing this neuron does is find present tense verbs ending in 'ing'
Activations:
<start>
Star  unknown
ting  unknown
from  unknown
a     unknown
position unknown
of unknown
strength unknown
<end>
```

Methodology

Step 3

- Simulator Model이 생성한 activation value를 실제 activation value와 비교하여 얼마나 잘 예측했는지 점수로 표현함

- Simulator Model 생성한 activation value의 scale (1~10)이 실제 activation value의 scale과는 다르므로 단순히 이 둘 사이에 pearson 상관계수인 ρ를 사용함.

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

x_i: 예측한 i번째 activation value

$\bar{x}$: 예측한 activation value의 평균

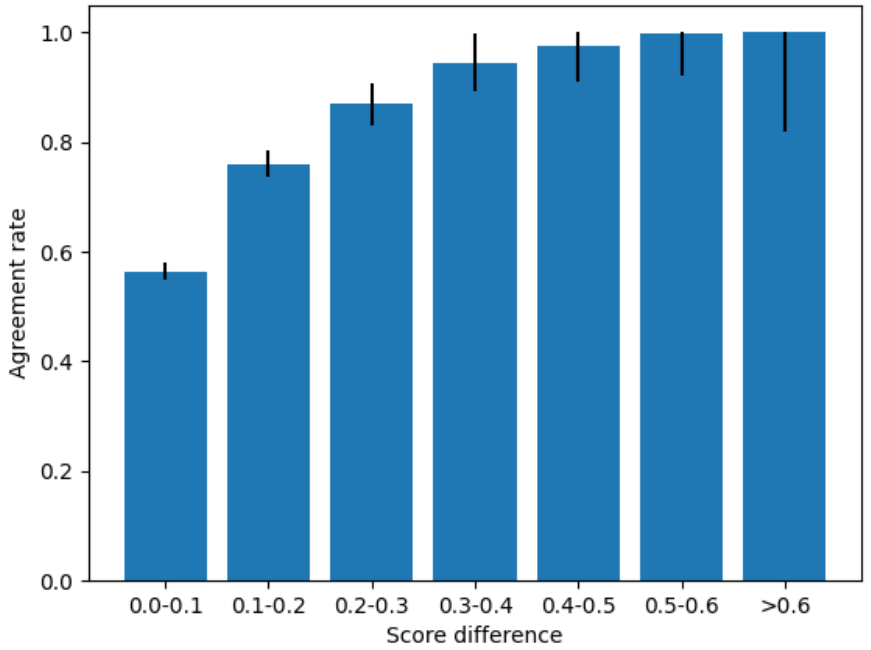
y_i: 실제 i번째 activation value

$\bar{y}$: 실제 activation value의 평균

Methodology

Validation of scoring

- ρ 가 둘 사이의 제대로 된 비교를 하지 못할 수 있으므로 점수에 대한 검증 과정이 필요함.
- Simulator가 예측한 activation value와 실제 activation value가 얼마나 비슷한지 사람에게 1~5의 scale로 평가하게 하도록 함. ρ가 높을수록 사람이 평가한 점수도 높게 나와, ρ가 점수로 사용하기 적합하다는 결론을 내림.

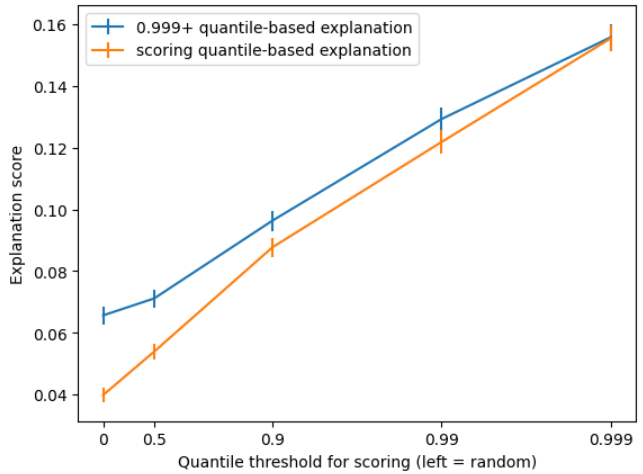


Methodology

Algorithm parameters and details

- Explainer Model의 prompt에 어떠한 {<token>, <activation value>} pair가 들어가는 것이 좋은지 탐구함.
- 다음과 같은 ablation study를 통해 "top-activating" 텍스트 (activation value를 특정 quantile로 분류했을 때 특정 quantile 이상의 token이 최소 하나가 있는 텍스트)를 사용하기로 함.
  1. 한 텍스트에 16-64개의 token이 들어있는 경우에 ρ가 가장 높았다.
  2. 10개 이상의 텍스트를 사용하는 것은 ρ값을 크게 변화시키지 못했다.
  3. 랜덤 텍스트나 특정 quantile 보다 낮은 토큰을 가진 텍스트를 사용했을 때 "top-activation" 텍스트를 사용했을 때보다 성능이 감소하였다.

<quantile threshold의 변화에 따른 ρ의 변화>



Algorithm parameters and details

- Simulator Model의 prompt에 어떠한 {<token>, <activation value>} pair가 들어가는 것이 좋은지 탐구함
- 다음과 같은 방법을 고려함

방법	내용	단점
random-only	랜덤 텍스트를 사용함	ρ의 표준편차가 지나치게 크게 나타남
top-and-random	5개의 top-activating과 5개의 랜덤 텍스트를 모두 사용함	지나치게 broad한 설명에 대해서만 높은 ρ가 나타났다.

- 두 방법이 가지고 있는 장단점 때문에 두 방법 모두 사용하기로 함

Methodology

Revising Explanations

- Explainer Model이 일부 neuron에 대해 설명을 지나치게 broad하게 생성하는 경우가 있음
  - 예시) 13 layer의 1352번째 neuron이 활성화되는 토큰은 "all" 이여서 설명은 "the term 'all' along with related contextual phrases." 라고 생성하지만 텍스트를 분석해보니 모두 "all" 토큰 전에 "not" 토큰이 있다. 실제 neuron이 "all" 에 대해서만 활성화 되어서 ρ도 높게 나타났지만 결코 좋은 설명이라고 할 수 없다.

:실제 neuron이 활성화된 token

<"all" 토큰 전에 "not" 토큰이 나타난다는 것을 알 수 있다>

Sorry, none of our films matched your search criteria.

Please check to make sure you checked off the correct filters at the top of the page.

Try a more general search.

Not all films from our collection are available online.

It's possible that the film you're searching for is

is updated every hour.

Please note: not all stray animal pictures are posted on this site. Please visit the shelter regularly to look for your lost pet.

We will do our best to help with your search, but as the owner, you are ultimately responsible to look for and identify your pet.

your destination. In the unfortunate event of shipping damage, complimentary wedges cannot be replaced

for free if they are damaged in transit.

NOTE ABOUT COMPLIMENTARY WEDGES

All AA grade handles are shipped with complimentary wedges - if wedges are required.

Not all handles

pop and minimalist designs and finely inked giclée prints; then The Neue Modern Press is the shop for you!I also can create custom orders, please contact me through Etsy conversations if you would like me to create some works not currently featured in my shop. Please note that not all custom orders can/

Methodology

Revising Explanations

- 이러한 문제를 해결하기 위해 Explainer Model이 자기가 설명한 설명을 검토하도록 함

<초기 설명>

the term ‘all' along with related contextual phrases

<explainer model>

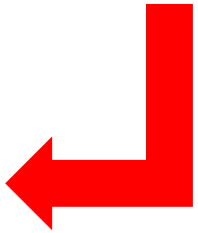
<설명에 대한 램덤 문장 10개 생성>

All films from our collection are available online.
All stray animal pictures are posted on this site.
...
Please note that all custom orders can /

<subject model>

<생성한 문장에 대한 activation value 값 확보>

Token	Activation
All	0
Flims	0
...	...
Online	0

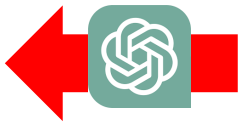


<확보한 activation value를 통해 prompt 구성 >

Revise following explanation based on evidence. Activating words highlighted with **.
<Explanation>
Not **All** films from our collection are available online.

<Evidence>
All films from our collection are available online.

<explainer model>



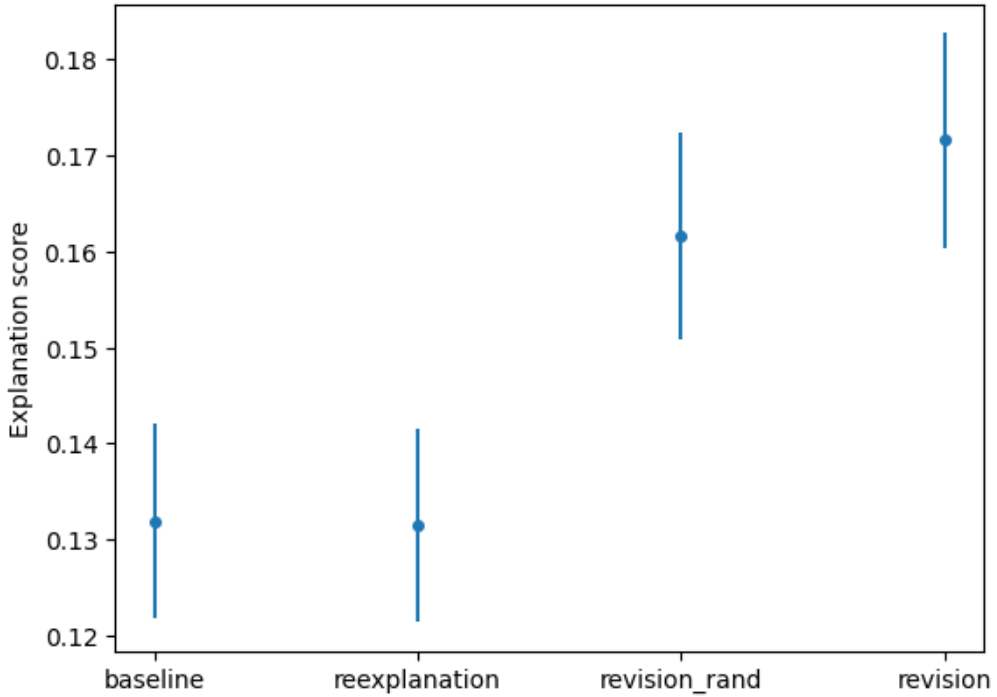
<수정된 설명>

the term ‘not all' along with related contextual phrases

Methodology

Revising Explanations

- 그러나 검토된 설명(reexplanation)을 주었을 때가 검토하지 않았을 때(baseline)보다 ρ가 감소함.
- 초기 설명 + 검토된 설명(Revision)을 주자 baseline 보다 ρ가 증가함.
- 초기 설명 + 0이 아닌 activation value를 가진 랜덤 문장에 대한 설명(Revision_rand)를 주자 revision과 비슷한 ρ를 보임

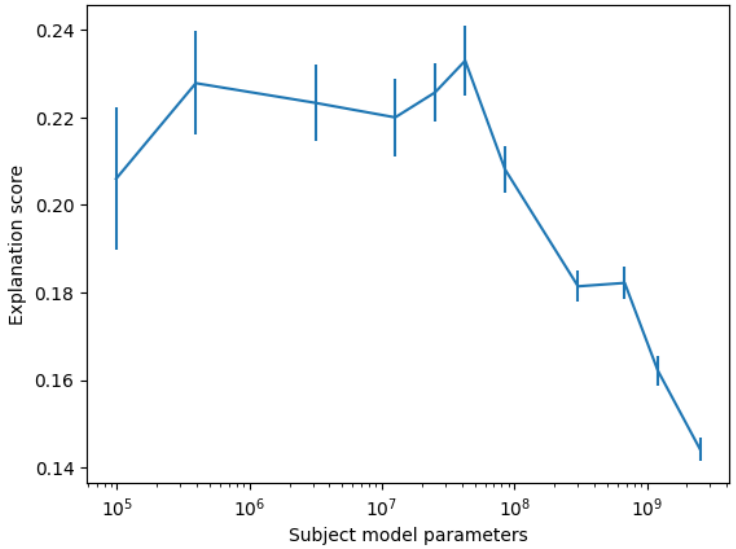


Experiment

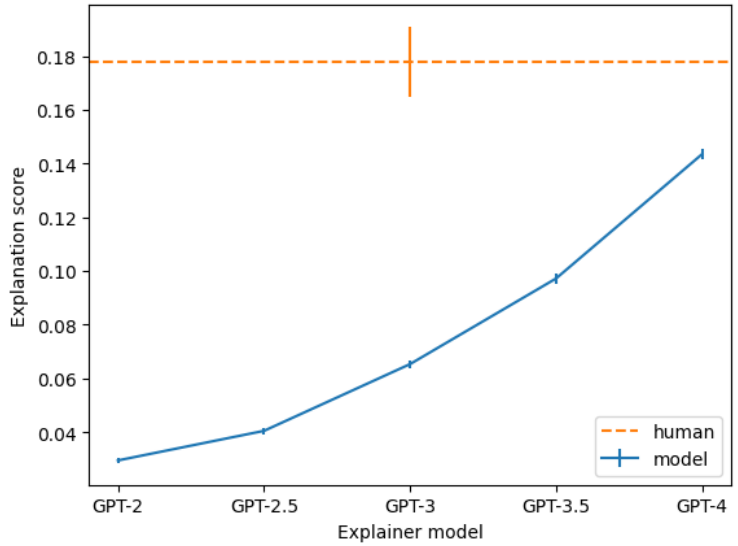
Model scaling trends

- Subject Model, Explainer Model, Simulator Model를 scaling 했을 때 ρ의 변화를 확인함
- Subject Model은 GPT-3 series 모델을 사용하였는데, 대체적으로 특정 parameter까지는 증가하다가 이후로는 감소하는 양상을 보임
- Explainer Model, Simulator Model 모두 모델이 좋아질수록 ρ가 증가함

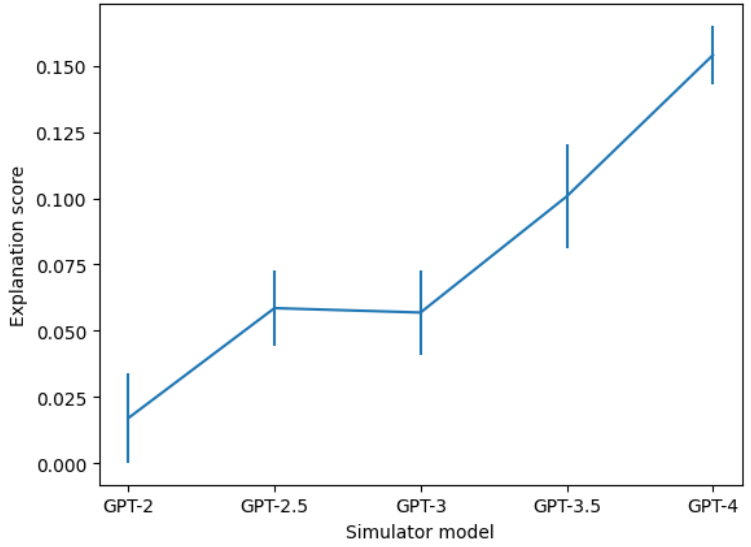
<Subject Model trend>



<Explainer Model trend>



<Simulator Model trend>



Experiment

Qualitative results – 높은 ρ를 중심으로

<생성한 설명>

<Neuron이 집중하는 토큰>

:실제 neuron이 활성화된 token

references to Canadian people, places, and entities.

ambaud told The Canadian Press on Thursday.

The people of the Dene Tha' First Nation tribal community live, farm fish, trap and hunt near where the spill happened. According to a statement released by Dene Tha' Chief James Ahnassay, the spill "seriously affected harvesting areas

Public Inc. announced the launch of Sit Kicker. Sit Kicker is a nationwide initiative focused on encouraging Canadians who work in office settings to reduce sedentary behaviour and "kick the sit" out of their work habits by shifting workplace culture towards more stand-friendly physical work environments. The Public Health Agency of Canada is providing

numbers that are part of a fraction or series.

Brexit, lawyers warn Tobacco and alcohol companies could win more easily in court cases such as the recent battle over plain cigarette packaging if the EU Charter of Fundamental Rights is abandoned, a barrister and public health professor have said. Getty 25/43 'Thousands dying' due to fear over non-existent statin side-effects

soup." –NME compendium Getty Images 5/14 On Beyonce's lack of talent "[If the word "artistry"] applies to Beyoncé then f**k me." –National Post, Feb. 2015 Getty Images 6/14 On the losers who go to rehab "Didn't go into rehab like

Experiment

Qualitative results – 낮은 ρ를 중심으로

<생성한 설명>

<Neuron이 집중하는 토큰>

:실제 neuron이 활성화된 token

numbers and numeric expressions related to sports scores and statistics.

and turned the ball over on offense. It seemed that every shot the Hornets were taking besides a three by Kemba Walker was coming up short, and the Nuggets took an early 12-5 lead with some nice play inside and some threes by the rookies, Emmanuel Mudiay and Nikola Jokic. Turnovers

Pop will actually strike, but luckily this is the first game of the night so we should know at least an hour before tip-off. Here you can find value plays in Danny Green, Patty Mills, and maybe David West. Can always run Kawhi or Aldridge if and hope they can reach their value before being

words or phrases related to actions or events.

<|endoftext|>Recent results:

Sep 2, 2017 6:30:00 PM - Sep 22, 2017 6:30:03 AM

Sep 22, 2017 6:30:03 AM - Mar 22, 2018 10:01:50 PM

Mar 22, 2018 10:01:50 PM - May

Clean the Bedding

Wash all the clothes like the hammock or sleep sack etc... on Weekly Basis.

Basis. Only use hot water for washing the clothes.

Don't use detergents because the smell of the detergent may irritate ferret.

4

Limitation

- 일부 Neuron이 하나의 특징이 아닌 여러 개의 특징을 나타낼 수 있으므로 설명 가능하지 않을 수 있다.
- 단순히 neuron의 activation이 아닌 neuron의 causal effect를 나타내는 circuit을 통해서 neuron의 behavior를 설명하는 것이 더 정확할 수 있다.
- Explainer Model과 Simulator Model 사이에서 steganography 현상이 발생할 수 있다.
  - 예시) Subject Model은 feature X를 가지고 있는데 Explainer Model이 잘못 Y라고 말할 수 있고 Simulator Model은 X feature에 해당하는 토큰에 높은 activation value 값을 부여할 수 있다.

Conclusion

- LLM을 통해 각각의 neuron들이 어떠한 behavior를 보이는지 human-friendly 하게 설명하는 방법을 제시함

- LLM이 생성한 답변을 정량적으로 평가하는 방법을 제시하여, 설명이 얼마나 정확한지 파악할 수 있도록 함

SPARSE AUTOENCODERS FIND HIGHLY INTERPRETABLE FEATURES IN LANGUAGE MODELS

Hoagy Cunningham[*12], **Aidan Ewart**[*13], **Logan Riggs**[*1], **Robert Huben**, **Lee Sharkey**[4]
[1]EleutherAI, [2]MATS, [3]Bristol AI Safety Centre, [4]Apollo Research
{hoagycunningham, aidanprattewart, logansmith5}@gmail.com

Introduction

- Neuron은 feature들을 polysemantic하게 나타낸다.
- Polysemantic?
  - 한 개의 feature -> 한 개의 neuron에 대응하지 않고
  - 여러 개의 feature -> 한 개의 neuron에 대응함

- 이렇게 polysemantic한 feature들은 적은 수의 neuron으로 많은 feature들을 나타낼 수 있다는 장점이 있지만 interpretable하지 않다는 단점이 있음

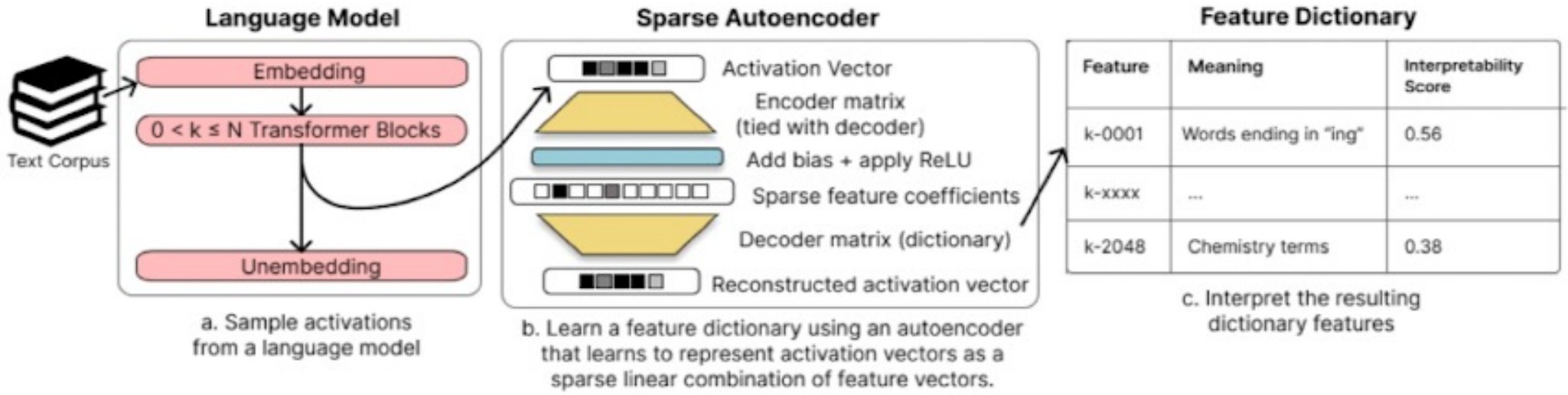
Introduction

- 따라서 polysemantic한 feature를 interpretable 하게 만들기 위해 monosemantic한 feature들로 재구성해야 함
  →이를 'dictionary learning' 이라고 함

- 해당 논문에서는 sparse autoencoder를 dictionary로 활용하는 방법을 제시함

Methodology

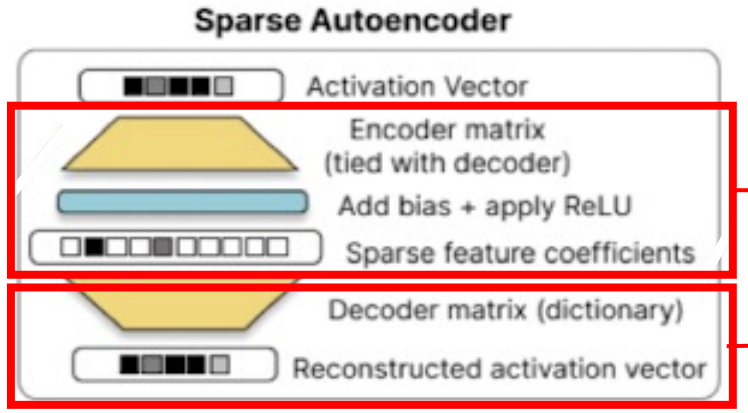
Overview



1. Transformer의 residual stream, MLP layer, attention의 head등에서 internal activation을 추출함
2. Internal activation을 sparse autoencoder를 통해 더 작은 feature들로 나타냄
3. 각 feature들이 가지는 의미를 GPT-4로 해석하여 interpretability score 산출

Methodology

Sparse autoencoder



$$\mathbf{c} = \mathrm{ReLU}(M\mathbf{x} + \mathbf{b})$$

$$\hat{\mathbf{x}} = M^T\mathbf{c} = \sum_{i=0}^{d_{\text{hid}}-1} c_i \mathbf{f}_i$$

$$\mathcal{L}(\mathbf{x}) = \underbrace{||\mathbf{x} - \hat{\mathbf{x}}||_2^2}_{\text{Reconstruction loss}} + \underbrace{\alpha||\mathbf{c}||_1}_{\text{Sparsity loss}}$$

$\mathrm{x} \in \mathbb{R}^{d_{in} \times n_{token}}$

$M \in \mathbb{R}^{d_{hid} \times d_{in}}$

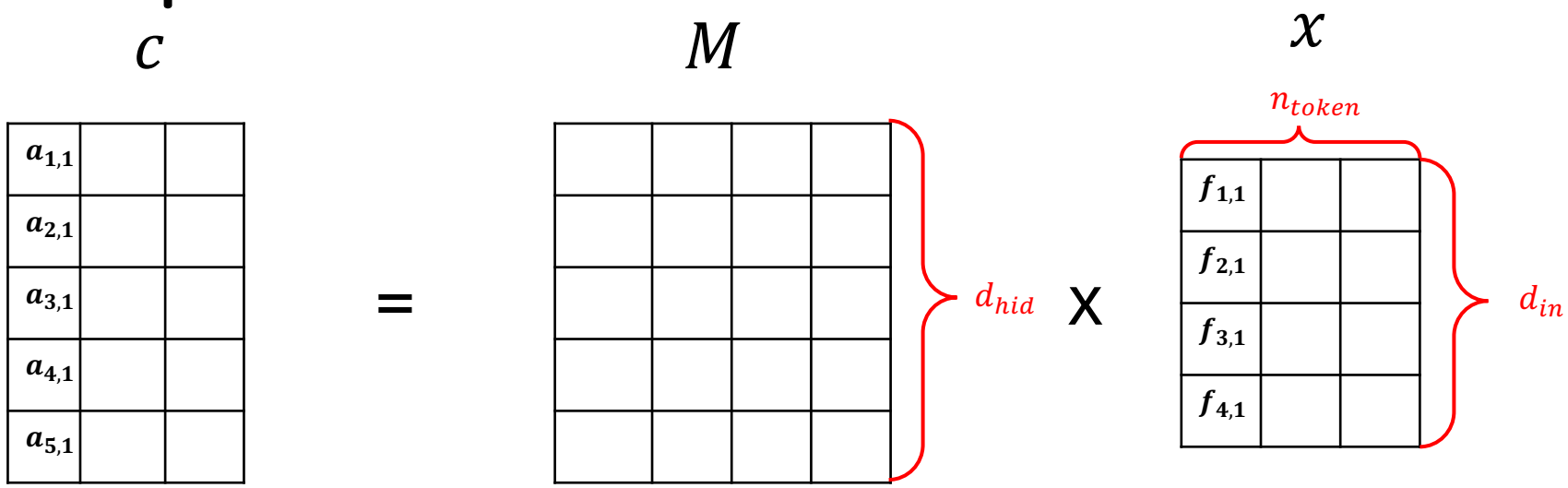
$b \in \mathbb{R}^{d_{in}}$

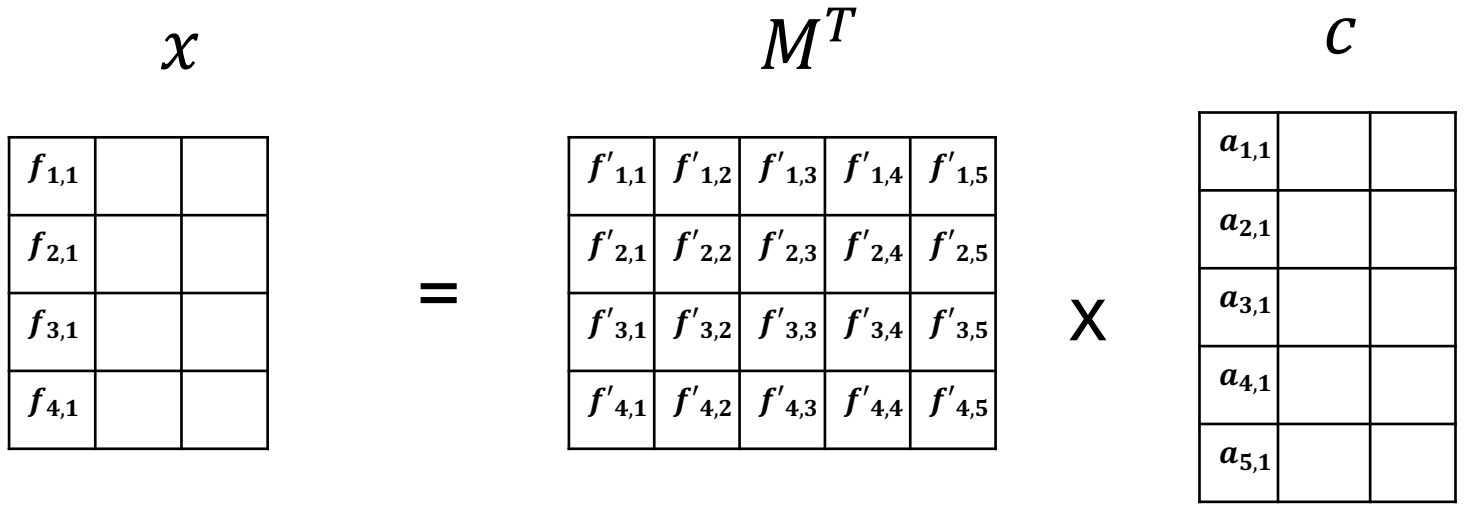
$c \in \mathbb{R}^{d_{hid} \times n_{token}}$

1. Decoder matrix는 encoder matrix의 transpose를 사용함
2. (x와 $\hat{x}$의 reconstruction L2 loss) + (c의 L1 loss) 를 통해 M을 학습함

Methodology

Simple example

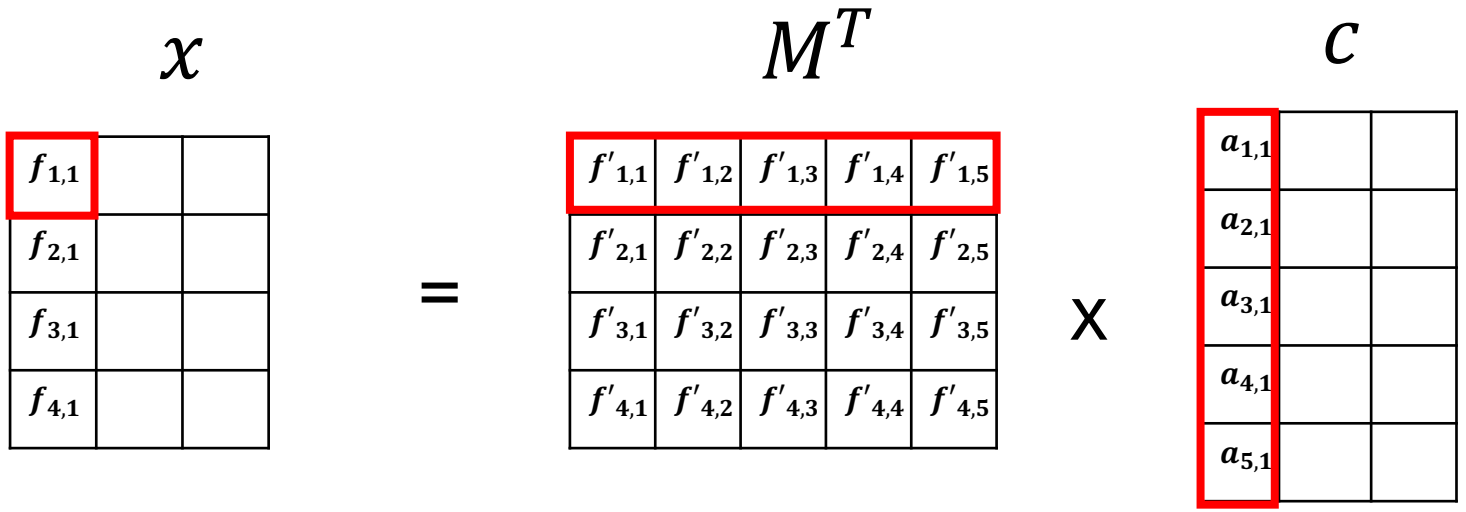




*한개의 neuron이 한개의 feature에 대응한다고 가정함

Methodology

Simple example



$$f_{1,1} = a_{1,1} f'_{1,1} + a_{2,1} f'_{1,2} + a_{3,1} f'_{1,3} + a_{4,1} f'_{1,4} + a_{5,1} f'_{1,5} + a_{6,1} f'_{1,6}$$

→ 즉, 한 개의 feature를 5개 feature들의 linear combination으로 나타낼 수 있다!

Experiment

Interpreting dictionary features

1. 각 dictionary feature들이 어떤 것을 나타내는지 파악하기 위해 interpretability score를 계산함

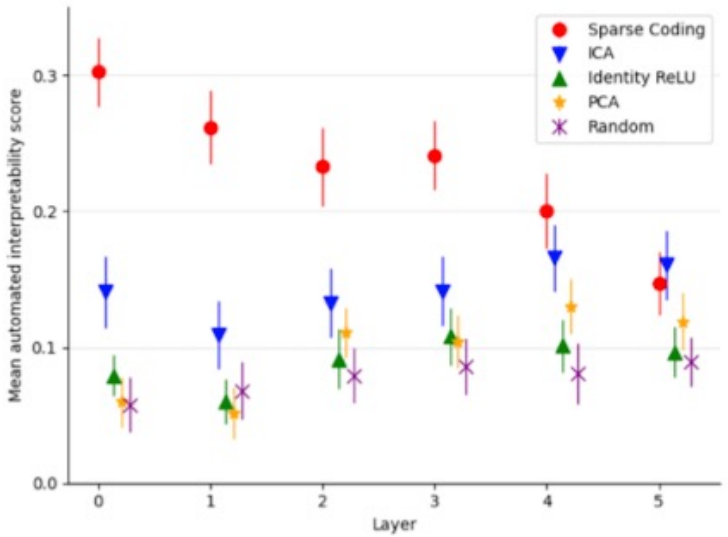
2. 다음의 dictionary feature를 추출하는 방법들과 비교함
    1. Default basis
    2. Random directions
    3. PCA (Princple Component Analysis)
    4. ICA (Independent Component Analysis)

3. baseline 모델로 Pythia-70M와 Pythia-410M을 사용함

4. Autoencoder를 모델의 residual stream에만 적용함

Experiment

Interpreting dictionary features



1. Layer가 낮을 수록 spase encoding 방법이 다른 방법보다 더 interpretable한 feature들을 추출함
2. 그러나 layer가 높아질수록 다른 방법들과의 차이가 줄어듬

   → 이에 대해 해당 논문은 “Layer가 높아질수록 더 복잡한 feature들이 나타나게 되는데, LLM은 이러한 feature의 pattern을 파악하지 못하기 때문에 interpretability가 감소한다” 라고 언급하고 있다.

Experiment

Identifying causally-important dictionary

1. 모델이 다른 방법들과 비교하여 얼마나 feature를 잘 나타내는지 확인하기 위해 IOI(Indirect Object Identification) task를 수행함

→ IOI task: 서로 독립적인 2개의 clause가 포함된 문장이 있을 때 첫번째 clause에 indirect object (IO)와 subject(S)가 주어지면 두번째 clause에 나타나는 빈칸을 IO로 맞추는 task

→ ex) "Then, Alice and Bob went to the store. Alice gave a snack to __ "

S IO S → Bob

2. 다음과 같은 순서로 task를 수행함
   1. 문장의 IO를 다른 단어로 변경함 (e.g, Bob → Vanessa)
   2. Feature Subset F에 속한 feature f들에 대한 변형된 문장의 feature coefficient $\bar{c}_{i,j}$ 를 구함
   3. Feature Subset F에 속한 feature f들에 대한 original 문장의 feature coefficient를 $c_{i,j}$ 를 구함
   4. 다음과 같은 식으로 모든 토큰의 input을 변경함

$$\mathbf{x}'_i = \mathbf{x}_i + \sum_{j \in F} (\bar{\mathbf{c}}_{i,j} - \mathbf{c}_{i,j}) \mathbf{f}_j$$

x_i: i번째 토큰
$\bar{c}_{i,j}$: 변형된 문장의 i번째 토큰의 j번째 coefficient
$c_{i,j}$: original 문장의 i번째 토큰의 j번째 coefficient

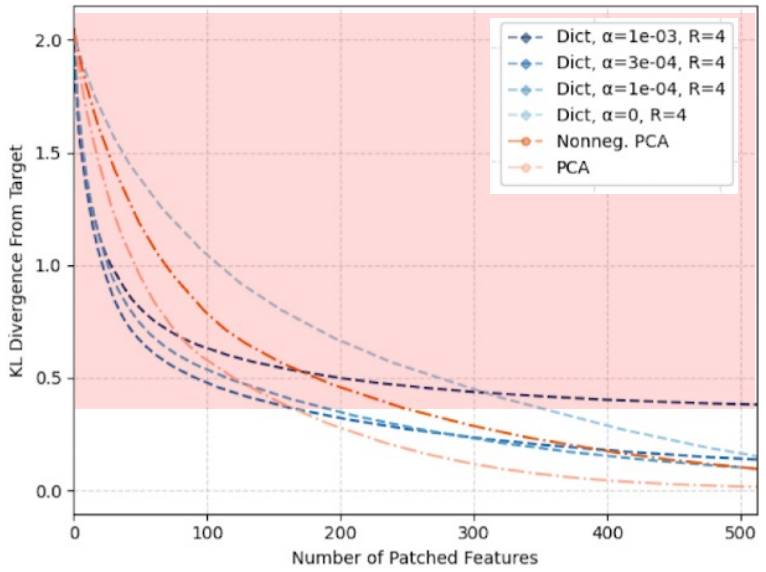
5. x_i를 모델에 통과시켰을 때 나오는 logit y 와 x'_i를 모델에 통과시켰을 때 나오는 logit z 사이의 KL divergence를 계산함

$$D_{KL}(\mathbf{z} || \mathbf{y})$$

Experiment

Identifying causally-important dictionary

<11번째 layer에서 변경하는 feature의 개수에 따른 KLD의 변화>

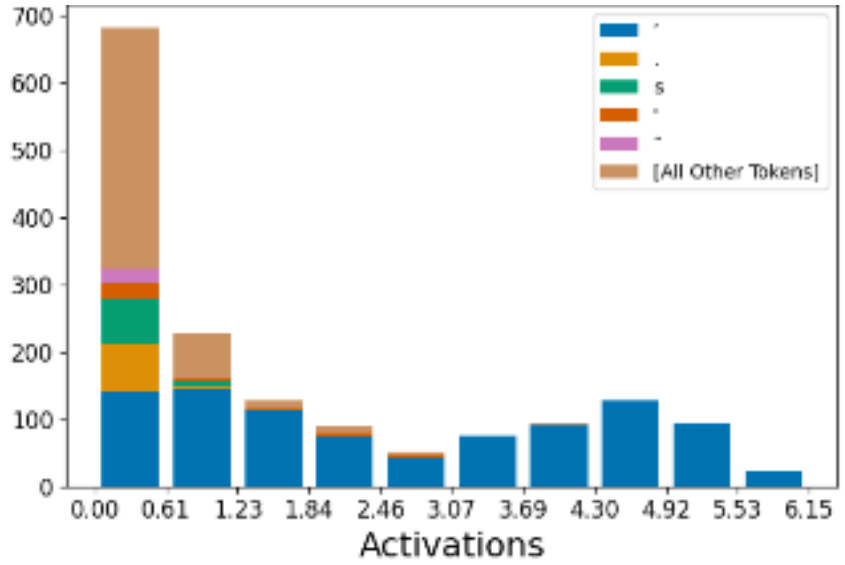


1. 여기서 R=h_{hid}/h_{in}, α는 sparsity loss를 조절하는 계수
2. Patch하는 feature의 수가 많을수록 KL divergence 값이 감소함
3. Sparse encoding($\alpha < 0$)이 PCA 보다 같은 수치의 KL divergence를 달성하기 위해 변형해야 하는 feature의 개수가 적음. 즉, sparse encoding 방법이 PCA보다 핵심적인 feature들을 더 잘 표현한다는 것을 의미함

Experiment

Case Studies

<556번째 dictionary feature에 나타나는 activation의 histogram>

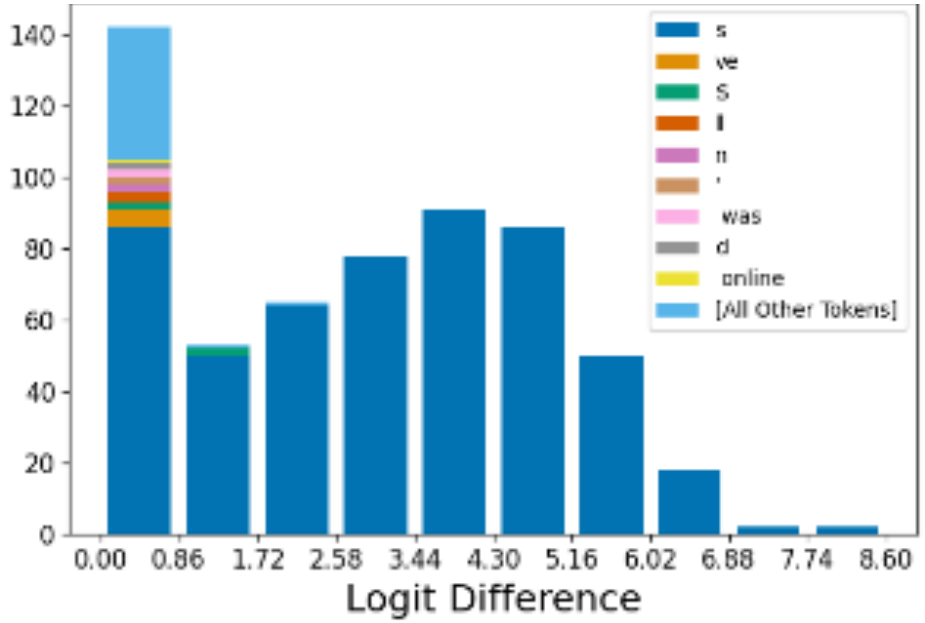


1. 각각의 dictionary feature들이 monosemantic한 feature를 나타내는지 확인하기 위해 토큰에 대한 activation의 histogram 확인함
2. 위 그래프는 556번째 dictionary feature에 나타나는 activation의 histogram을 표현하고 있는데 (단, 몇 번째 layer인지는 나와 있지 않다), apostrophe가 가장 많이 나타난다는 것을 통해 해당 dictionary feature가 apostrophe를 나타낸다는 것을 알 수 있다.

Experiment

Case Studies

<Output logit difference>



1. apostrophe에 해당하는 dictionary feature가 최종 output logit에 끼치는 영향을 파악하기 위해 apostrophe에 해당하는 dictionary feature를 residual stream에서 제거했을 때 나타나는 최종 output logit 값의 분포와 제거하지 않았을 때 output logit 값의 차이를 계산함
2. apostrophe 다음에 가장 자주 나오는 s가 logit difference에서 가장 많이 나타난다는 것을 알 수 있음

Conclusion

1. Sparse autoencoder를 통해 polysemantic한 특징을 가지는 feature들을 monosemantic한 feature들로 disentangle 하는 방법을 제시함

2. 그러나 다음과 같은 한계점을 가지고 있음
    1. Dictionary feature들을 학습하기 위해 사용한 Loss가 0으로 수렴하지 않았다. 이것은 dictionary feature가 모델이 표현하는 모든 feature를 표현하지 못한다는 의미이다.
    2. Autoencoder를 residual stream에만 적용하였다. MLP에도 적용해봤지만, 활성화되지 않은 feature들이 많아서 별 효과가 없었다고 한다.
    3. IOI task를 통해서만 dictionary feature를 분석하였기 때문에 neuron들에 대한 완벽한 해석이 이루어지지 않았다.

감사합니다