- “An Observation on Generalization”
  Talk given by Ilya Sutskever at Simons Institute

- https://www.youtube.com/watch?v=AKMuA_TVz3A&ab_channel=SimonsInstitute

Slides created by Yegon Kim
2024/10/15

Supervised Learning

- Supervised learning is conceptually simple
- It is easy to create theories of it, e.g. uniform convergence bound
- We know that we are **doing our best** to a certain extent

$$\mathbf{Pr}_{S\sim D^{|S|}}\left[\text{Test}_D(f) - \text{Train}_S(f) \leq \sqrt{\frac{\log|\mathcal{F}| + \log 1/\delta}{|S|}} \quad \textit{for all } f \in \mathcal{F}\right] > 1-\delta$$

Uniform convergence bound

Unsupervised Learning

- Unsupervised learning: optimize one objective, for a different objective
- What if the current unsupervised methods are **very bad? (↑regret)**
- This can give some people nightmares

Unsupervised Learning

- How could we possibly create a theory of unsupervised learning?

Guiding examples

- Upstream dataset X, Downstream dataset Y
- e.g. Dataset X of English sentences, dataset Y of French sentences
  Using X can help Y
- e.g. Dataset X of bits sampled from uniform distribution
  Using X won't help with any Y

Compression

- Prediction = Compression
- A good model M can be turned into a good compressor C, and vice versa

Chunk	Compressor	Raw Compression Rate (%) enwik9	ImageNet	LibriSpeech	Random
∞	gzip	32.3	70.7	36.4	100.0
	LZMA2	23.0	57.9	29.9	100.0
	PNG	42.9	58.5	32.2	100.0
	FLAC	89.5	61.9	30.9	107.8
2048	gzip	48.1	68.6	38.5	100.1
	LZMA2	50.0	62.4	38.2	100.0
	PNG	80.6	61.7	37.6	103.2
	FLAC	88.9	60.9	30.3	107.2
	Transformer 200K	30.9	194.0	146.6	195.5
	Transformer 800K	21.7	185.1	131.1	200.1
	Transformer 3.2M	17.0	215.8	228.2	224.0
	Llama 2 (7B)	8.9	53.4	23.1	103.2
	Chinchilla 1B	11.3	62.2	24.9	108.8
	Chinchilla 7B	10.2	54.7	23.6	101.6
	Chinchilla 70B	**8.3**	**48.0**	**21.0**	100.8

"Language Modeling Is Compression" [Deletang et al, 2024]

We can 'solve' MNIST up to ~78% accuracy with the following code-golfed obscurity:

```
c = lambda z: len(gzip.compress(z.tobytes()))

def ncd(x, y):
    return (c(x + y) - min(c(x), c(y))) / max(c(x), c(y))

cls = [(x, c(x), l) for x, l in training_set]

correct_predictions = sum([np.array_equal(Counter(
    [l for _, _, l in sorted([(ncd(x1, x), x, l) for x, _, l in c
     key=lambda t: t[0])[:5]]).most_common(1)[0][0], label)
     for x1, label in test_set])
```

"78% MNIST accuracy using GZIP in under 10 lines of code" [Serlier, 2023]

Compression

- Kolmogorov complexity K(X)
  Length of the shortest program that outputs X

- For any computable compressor C,

$$K(X) \leq |C(X)| + K(C) + O(1)$$

Length of message — Length of program

Prediction error — Model complexity

Compression

- A deep neural net M **implements/simulates** a program
- SGD is doing program search
- $K(M)$ might be much smaller than the neural net implementation

- Goldblum et al, 2024. “Position: The No Free Lunch Theorem, Kolmogorov Complexity, and the Role of Inductive Biases in Machine Learning”

Formalization

- We can now formalize unsupervised learning as follows:
  - An algorithm C compresses Y
  - It is allowed access to X
  - What's the best algorithm C?

Formalization

- Conditional Kolmogorov complexity $K(Y|X)$
  Length of shortest program that outputs Y, when **allowed to probe X**

- For all X,

$$K(Y|X) \leq |C(Y|X)| + K(C) + O(1)$$

Formalization

- C(Y): **all the weights of deep NN** + *prediction error*
- C(Y|X): training code for X + $\Delta\theta$ (finetuning) + *prediction error*

- But is this truly the best we can do?

Joint compression

- As of yet, we do not have deep neural nets that can be truly conditioned on large **datasets** X
- But this is not much of an inconvenience:

$$K(X, Y) = K(X) + K(Y|X) + O(\log(K(X, Y)))$$

- “Just compress everything”

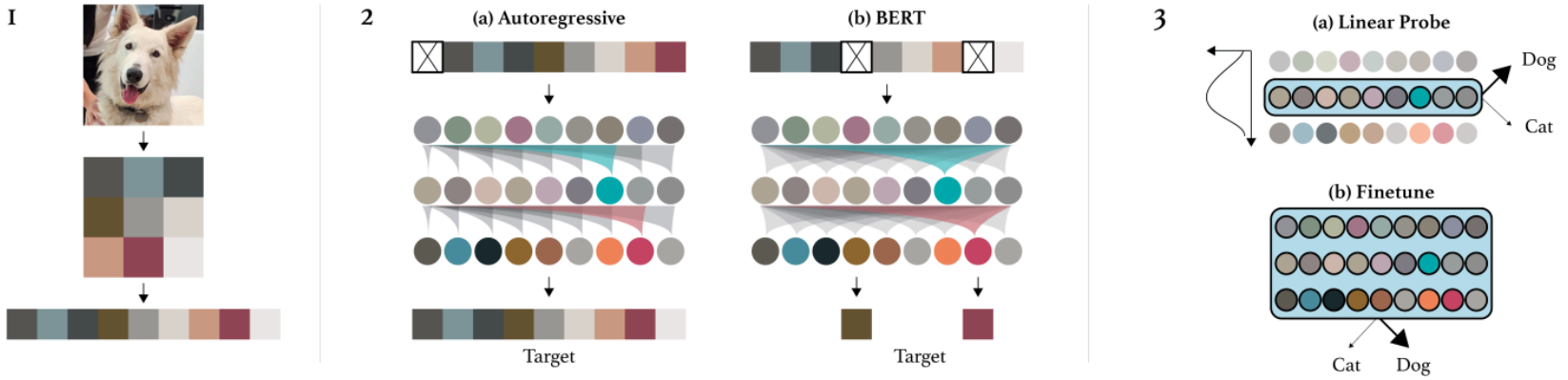
Experiments

- We don't really need this "theory" for explaining GPT, since the task is next token prediction on both X and Y.
  i.e. We are learning the same function on both datasets. It's already guaranteed that X will somewhat help Y.

- We therefore turn to image classification

Experiments

- “Generative Pretraining from Pixels” [Chen et al, 2020]



Experiments

Model	Acc	Unsup Transfer	Sup Transfer
CIFAR-10			
AutoAugment	98.5		
SimCLR	98.6	√	
GPipe	99.0		√
iGPT-L	99.0	√	
CIFAR-100			
iGPT-L	88.5	√	
SimCLR	89.0	√	
AutoAugment	89.3		
EfficientNet	91.7		√

To summarize

- Prediction on Y is compression of Y
- Prediction on Y with the help of X is compression of Y given X
- The best $C(X, Y)$ probably contains the best $C(Y \mid X)$