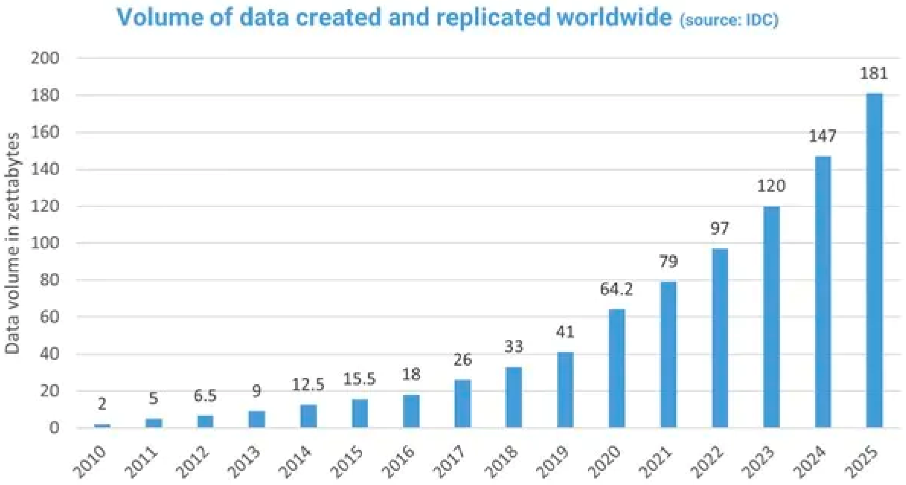
# Self Supervised Learning

Mehmet Saygin Seyfioglu
11/13/23

# What fuels the recent AI boom?

# What fuels the recent AI boom?

DATA



Taken from https://medium.com/@mwaliph/exponential-growth-of-data-2f53df89124

# What is the largest human-labeled dataset to date?

Ideas?

# What is the largest human-made dataset to date?

## How Large?

Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." *International journal of computer vision* 115 (2015): 211-252.

# What is the largest human-made dataset to date?

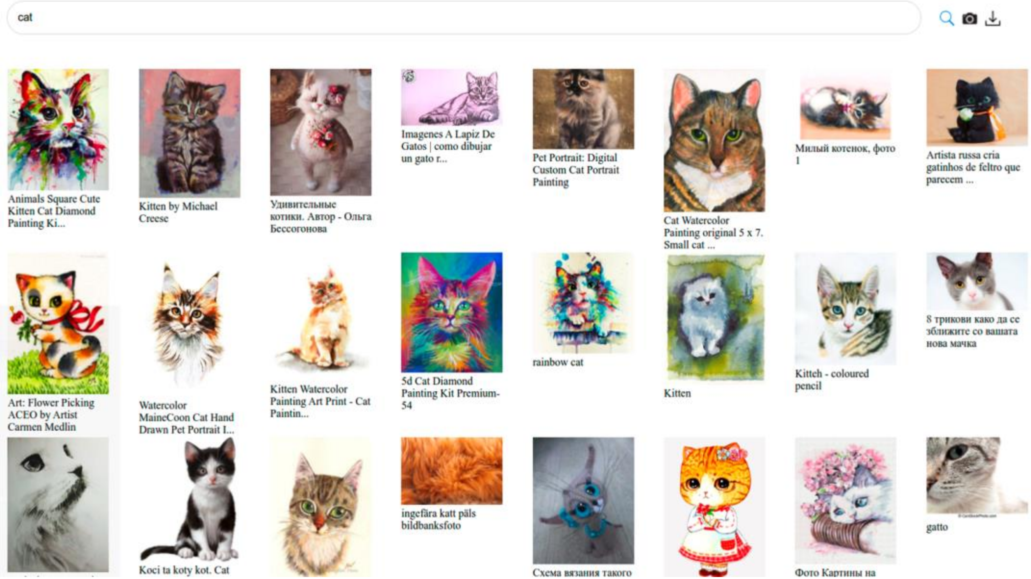
14M images and labels

How Large are the Datasets for training Foundational Models (Like CLIP, Stable Diffusion, GPT etc.)

Any ideas?

# How Large are the Datasets for training Foundational Models (Like CLIP, Stable Diffusion, GPT etc.)

## LAION 5B

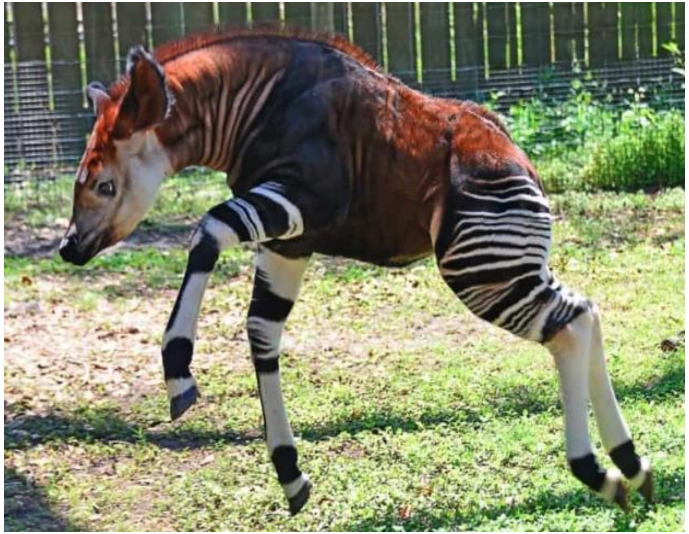

cat
Animals Square Cute Kitten Cat Diamond Painting Ki...
Kitten by Michael Creese
Удивительные котики. Автор - Ольга Бессоронова
Imagenes A Lapiz De Gatos | como dibujar un gato r...
Pet Portrait: Digital Custom Cat Portrait Painting
Cat Watercolor Painting original 5 x 7. Small cat ...
Милый котенок, фото 1
Artista russa cria gatinhos de feltro que parecem ...
Art: Flower Picking ACEO by Artist Carmen Medlin
Watercolor MaineCoon Cat Hand Drawn Pet Portrait I...
Kitten Watercolor Painting Art Print - Cat Paintin...
5d Cat Diamond Painting Kit Premium-54
rainbow cat
Kitten
Kitteh - coloured pencil
8 трикови како да се зближите со вашата нова мачка
Koci ta koty kot. Cat
ingefära katt päls bildbanksfoto
Схема вязания такого
Фото Картины на
gatto

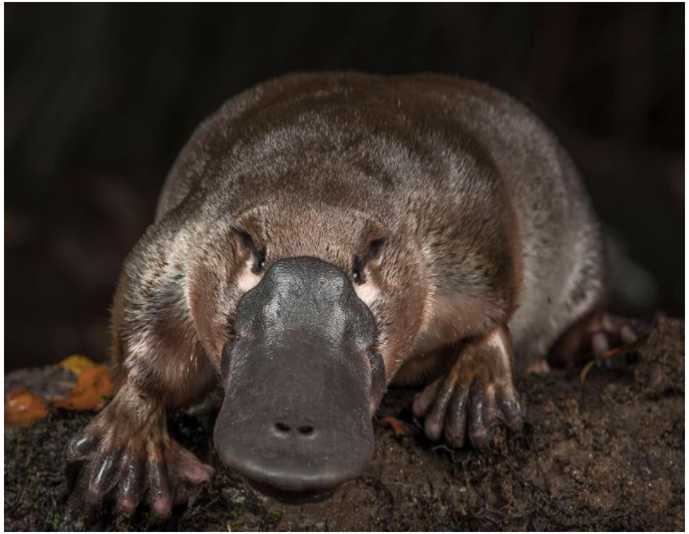
Impossible to rely on Human labels at this scale.

Schuhmann, Christoph, et al. "Laion-5b: An open large-scale dataset for training next generation image-text models." *Advances in Neural Information Processing Systems* 35 (2022): 25278-25294.

# Self Supervised Learning

What are these animals? Can you name them?





# Self Supervised Learning

- So SSL methods are generally encoding the similarity, or the difference of entities without specifically knowing what those entities are.

# Self Supervised Learning

- So SSL methods are generally encoding the similarity, or the difference of entities without specifically knowing what those entities are.
- Recent trend is to train a “Foundational Model” with a huge dataset in a self-supervised manner, which aims to learn some underlying generalizable facts then fine-tune this model for certain “downstream tasks”.

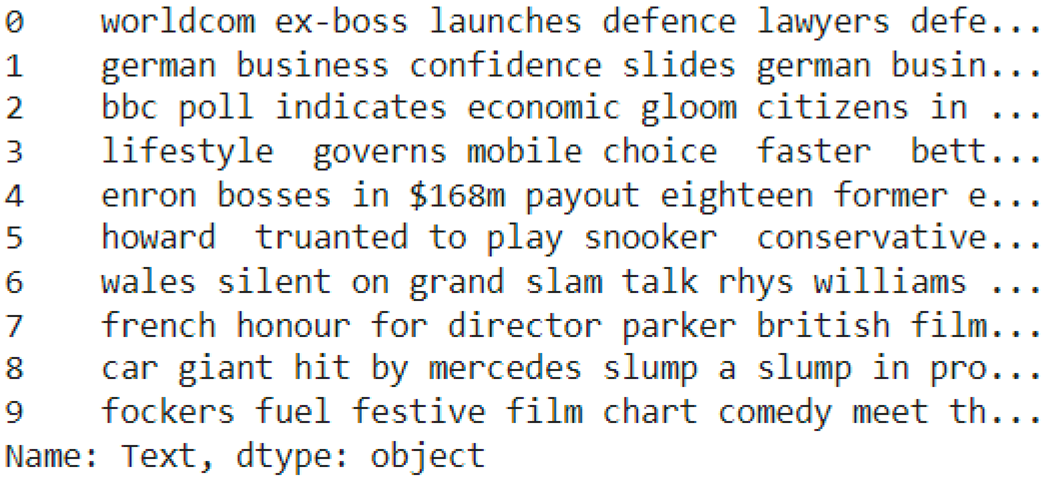
# Self Supervised Learning

- So SSL methods are generally encoding the similarity, or the difference of entities without specifically knowing what those entities are.
- Recent trend is to train a “Foundational Model” with a huge dataset in a self-supervised manner, which aims to learn some underlying generalizable facts then fine-tune this model for certain “downstream tasks”.

The trick is to create some pretext task, which generates some pseudo labels from unlabeled data. How?

# SSL for Natural Language Processing (NLP)

- How can we implement SSL techniques in Natural Language Processing?



```
0    worldcom ex-boss launches defence lawyers defe...
1    german business confidence slides german busin...
2    bbc poll indicates economic gloom citizens in ...
3    lifestyle  governs mobile choice  faster  bett...
4    enron bosses in $168m payout eighteen former e...
5    howard  truanted to play snooker  conservative...
6    wales silent on grand slam talk rhys williams ...
7    french honour for director parker british film...
8    car giant hit by mercedes slump a slump in pro...
9    fockers fuel festive film chart comedy meet th...
Name: Text, dtype: object
```

# SSL for Natural Language Processing (NLP)

- How can we implement SSL techniques in Natural Language Processing?
  - Context (align words that appear in similar context)
    - Any issues with this?

# SSL for Natural Language Processing (NLP)

- How can we implement SSL techniques in Natural Language Processing?
  - Context (align words that appear in similar context)
    - Any issues with this?
  - Masked Language Modeling (Predict the masked word using its context)
    - Predict Next word

# Word2Vec

Use a shallow neural network to learn word embeddings

$$\arg\max_{\theta} \prod_{(w,c)\in D} p(c|w;\theta)$$

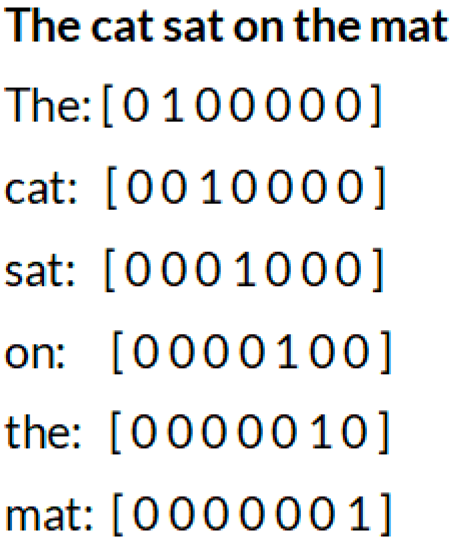
Let w denote the corpus of words and let c be the context of words for a given data set D. The word2vec model is trying to maximize the conditional probability p(c|w) by optimizing its parameters θ.

How do we solve this objective?

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems* 26 (2013).

# Word2Vec Skip-Gram

Assume we encoded each word in the corpus into a one hot vector.



**The cat sat on the mat**

The: [0 1 0 0 0 0 0]

cat: [0 0 1 0 0 0 0]

sat: [0 0 0 1 0 0 0]

on: [0 0 0 0 1 0 0]

the: [0 0 0 0 0 1 0]

mat: [0 0 0 0 0 0 1]

# Word2Vec Skip-Gram

Given one hot embedded word vectors, we can then try to do negative sampling to train a shallow network using softmax function:

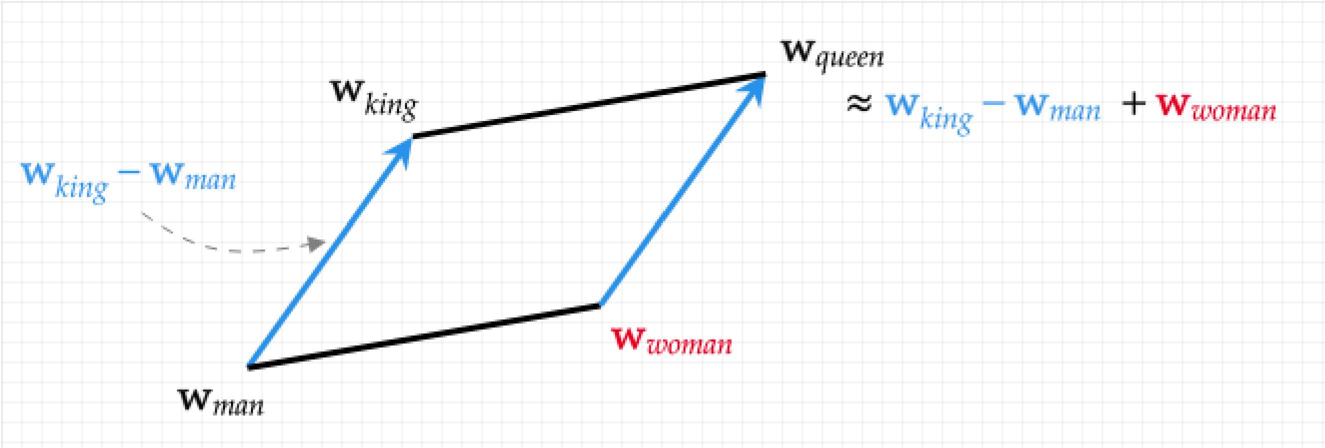
$$p(c|w;\theta) = \frac{e^{v_c . v_w}}{\sum_{(c' \in C)} e^{v_c{}' . v_w}}$$

The nominator here is a word w and its context c. On the denominator though, we randomly sample context from the dataset. This is called Negative Sampling.

So we didn’t label the dataset at all, but utilizing the context (word proximity)

# Word2Vec

Turns out, if you do this with a large enough dataset, you can learn vector representation of words which can be used in vector algebra!


$\mathbf{w}_{queen}$
$\approx \mathbf{w}_{king} - \mathbf{w}_{man} + \mathbf{w}_{woman}$
$\mathbf{w}_{king}$
$\mathbf{w}_{king} - \mathbf{w}_{man}$
$\mathbf{w}_{woman}$
$\mathbf{w}_{man}$

This was a wake up call for the field. Tom Mikolov came up with the word2vec algorithm at Google during 2013. What happened after then?

# What happened between 2013-2019?

If word2vec is great, why the research on SSL got stalled (!) in NLP?

# What happened between 2013-2019?

If word2vec is great, why the research on SSL got stalled (!) in NLP?

1. Folks mainly relied on LSTMs (recurrent nets) to learn from large datasets, and LSTMs does not have a capacity to encode long range dependencies.
2. We lacked data.

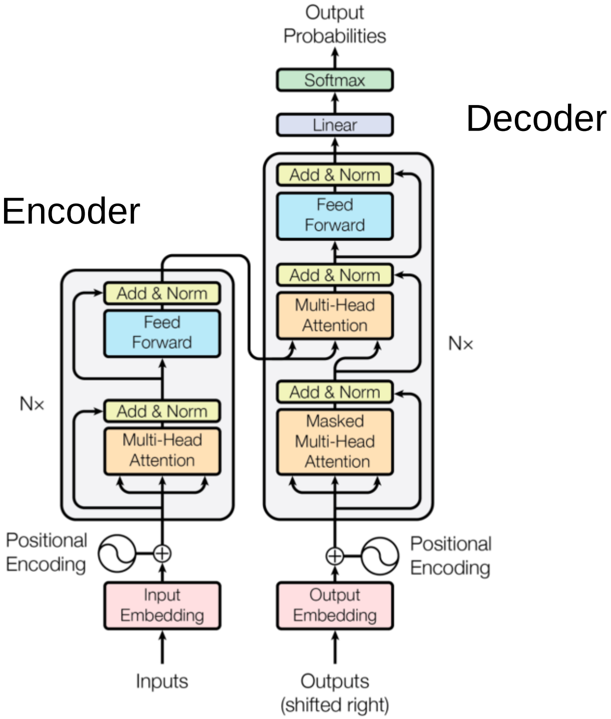
# What happened 2013-2019?

If word2vec is great, why the research on SSL got stalled (!) in NLP?

1. Folks mainly relied on LSTMs (recurrent nets) to learn from large datasets, and LSTMs does not have a capacity to encode long range dependencies.
2. We lacked data.

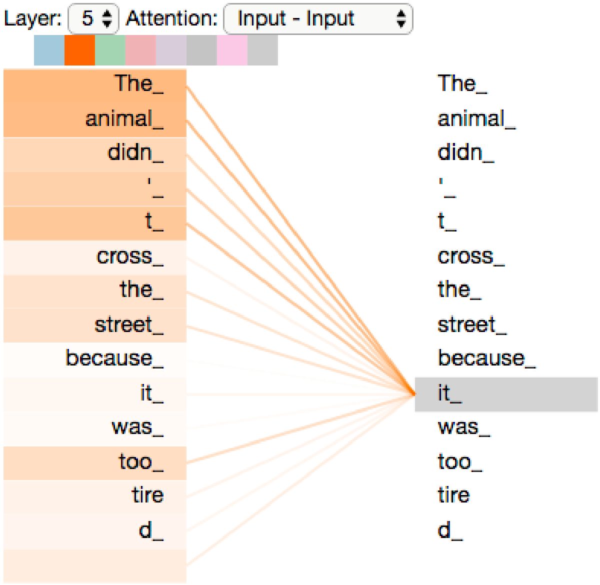
What architecture changed it?

# Transformer



Encoder

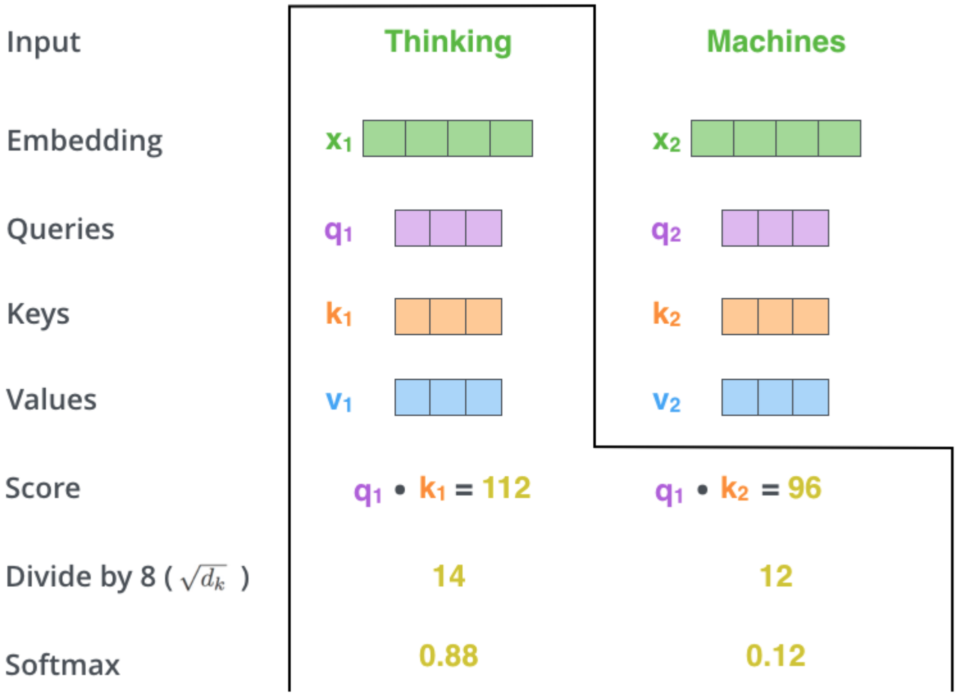
Decoder



Layer: 5 Attention: Input - Input

Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

# Transformer

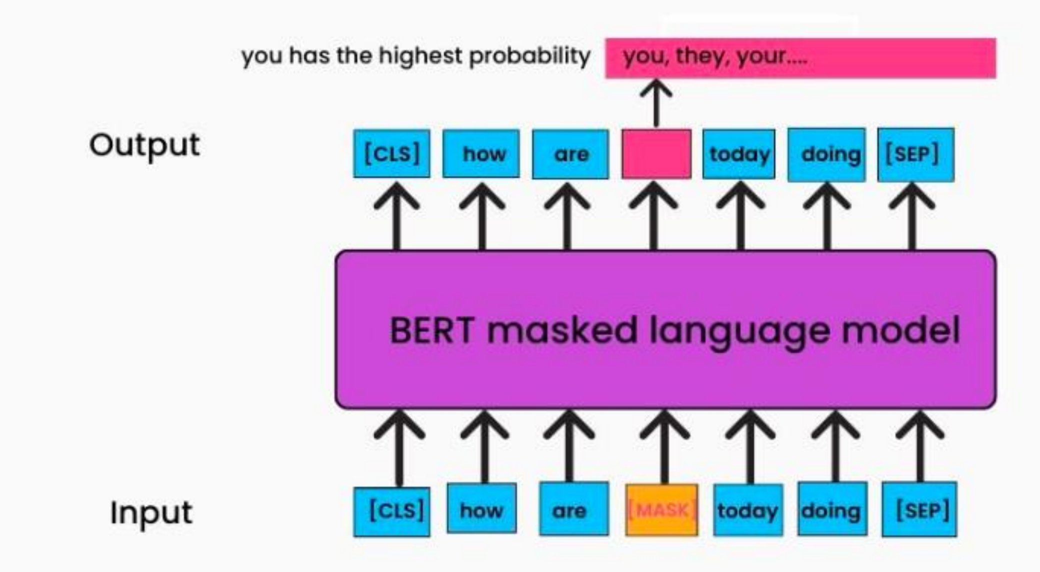


|  |  |  |
|---|---|---|
| Input | Thinking | Machines |
| Embedding | $x_1$ | $x_2$ |
| Queries | $q_1$ | $q_2$ |
| Keys | $k_1$ | $k_2$ |
| Values | $v_1$ | $v_2$ |
| Score | $q_1 \bullet k_1 = 112$ | $q_1 \bullet k_2 = 96$ |
| Divide by 8 ( $\sqrt{d_k}$ ) | 14 | 12 |
| Softmax | 0.88 | 0.12 |

Taken from https://jalammar.github.io/illustrated-transformer/

# BERT

Encoder-only architecture. 340 million parameters.



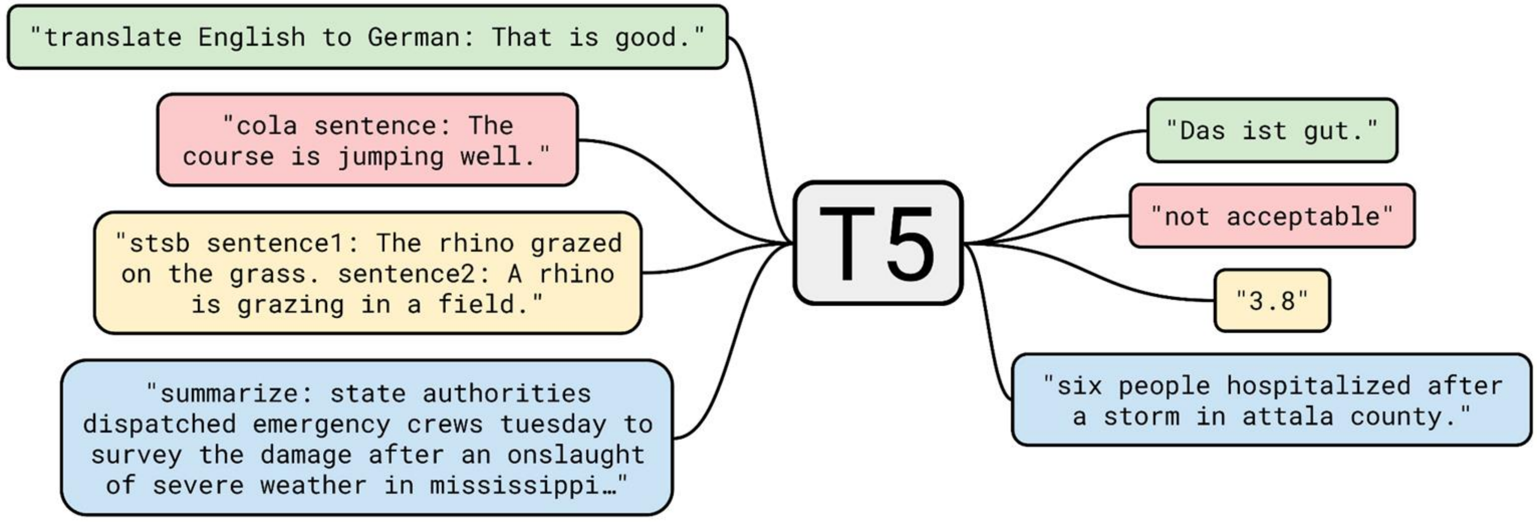
Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018)

Taken from https://www.turing.com/kb/how-bert-nlp-optimization-model-works

# T5

Encoder-Decoder model

Shuffling, Masked Language Modeling. 3b parameters



"translate English to German: That is good."

"cola sentence: The course is jumping well."

"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."

"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"

T5

"Das ist gut."

"not acceptable"

"3.8"

"six people hospitalized after a storm in attala county."

Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

# GPT

Decoder only architecture

Trained for the next word prediction (Turns out, this scales well with more data and model size, thus we have the modern GPTs now)

GPT2 - 1.75b parameters

…

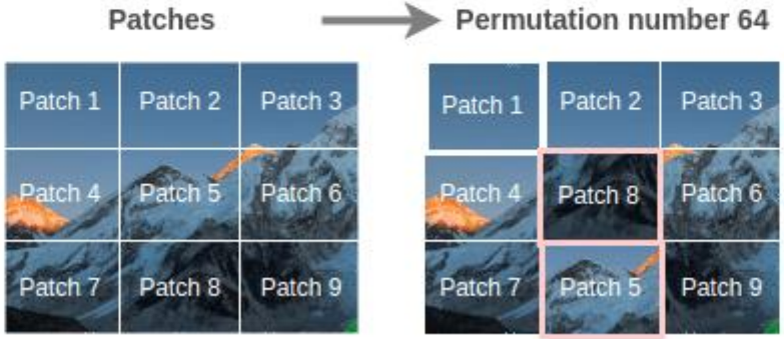
GPT4 - 275b parameters (estimation)

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

# SSL for Vision

How can we implement SSL techniques in Vision?

# SSL for Vision



Data Generation for Geometric Transformation Recognition

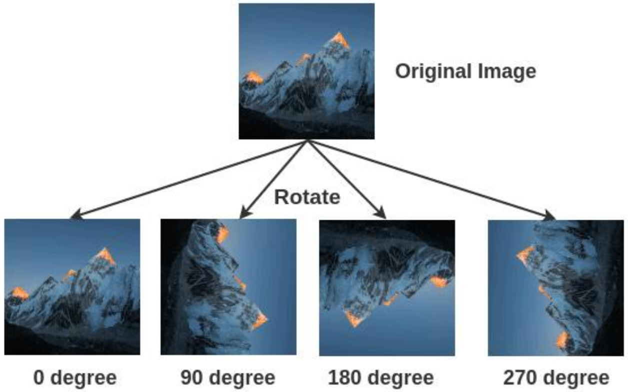
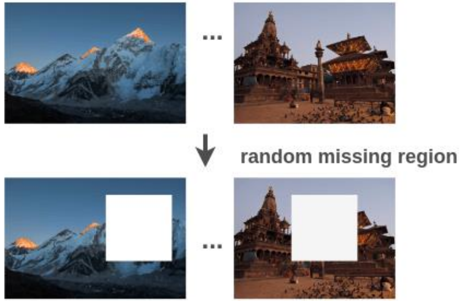


Image Inpainting Data Generation



# SimCLR

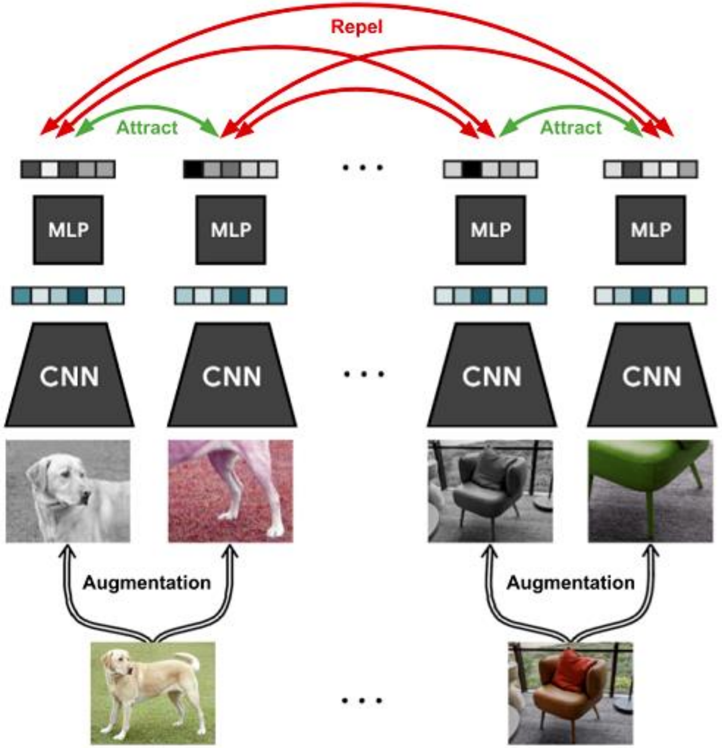


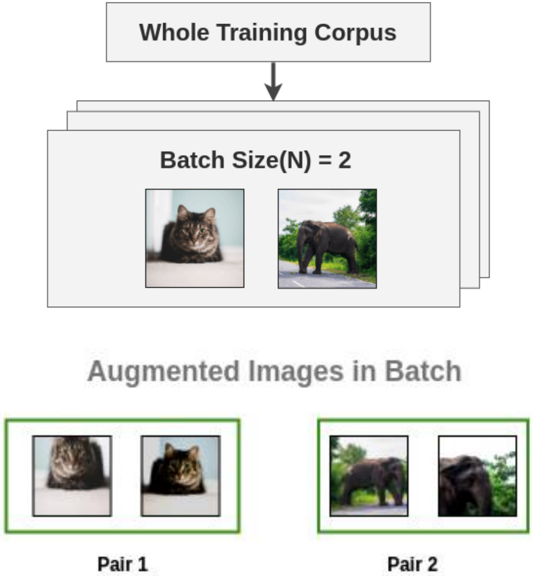
Figure taken from https://sh-tsang.medium.com/review-simclr-a-simple-framework-for-contrastive-learning-of-visual-representations-5de42ba0bc66

Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PMLR, 2020.

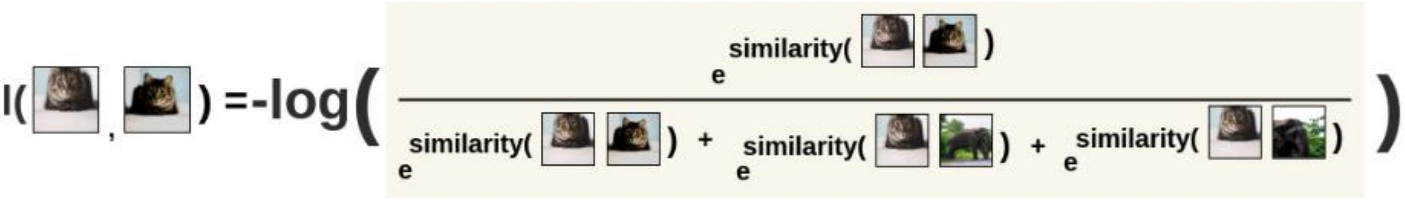
# SimCLR (cont’d)

$$L_{NCE} = -\sum_{i=1}^{n} log \frac{e^{\theta(t_1^i, t_2^i)}}{\frac{1}{b}\sum_{j=1}^{b} e^{\theta(t_1^i, t_2^j)}}$$

# SimCLR (cont’d)

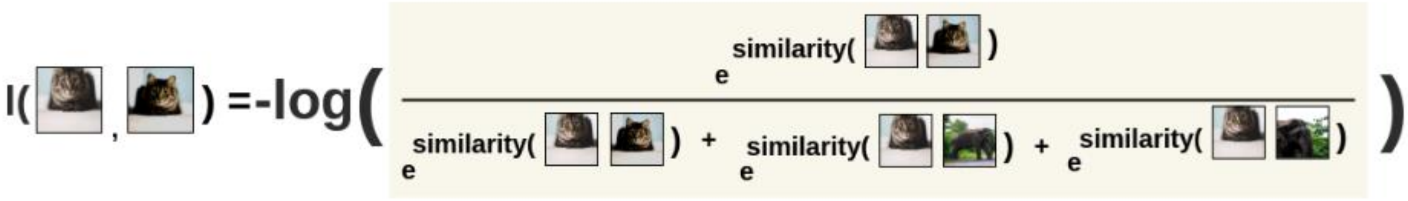


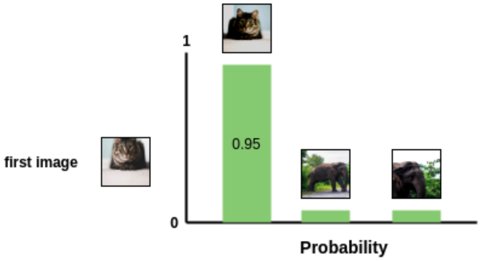
$$L_{NCE} = -\sum_{i=1}^{n} log \frac{e^{\theta(t_1^i, t_2^i)}}{\frac{1}{b}\sum_{j=1}^{b} e^{\theta(t_1^i, t_2^j)}}$$



# SimCLR (cont’d)

$$L_{NCE} = -\sum_{i=1}^{n} log \frac{e^{\theta(t_1^i,t_2^i)}}{\frac{1}{b}\sum_{j=1}^{b} e^{\theta(t_1^i,t_2^j)}}$$





Figures taken from https://amitness.com/2020/03/illustrated-simclr/

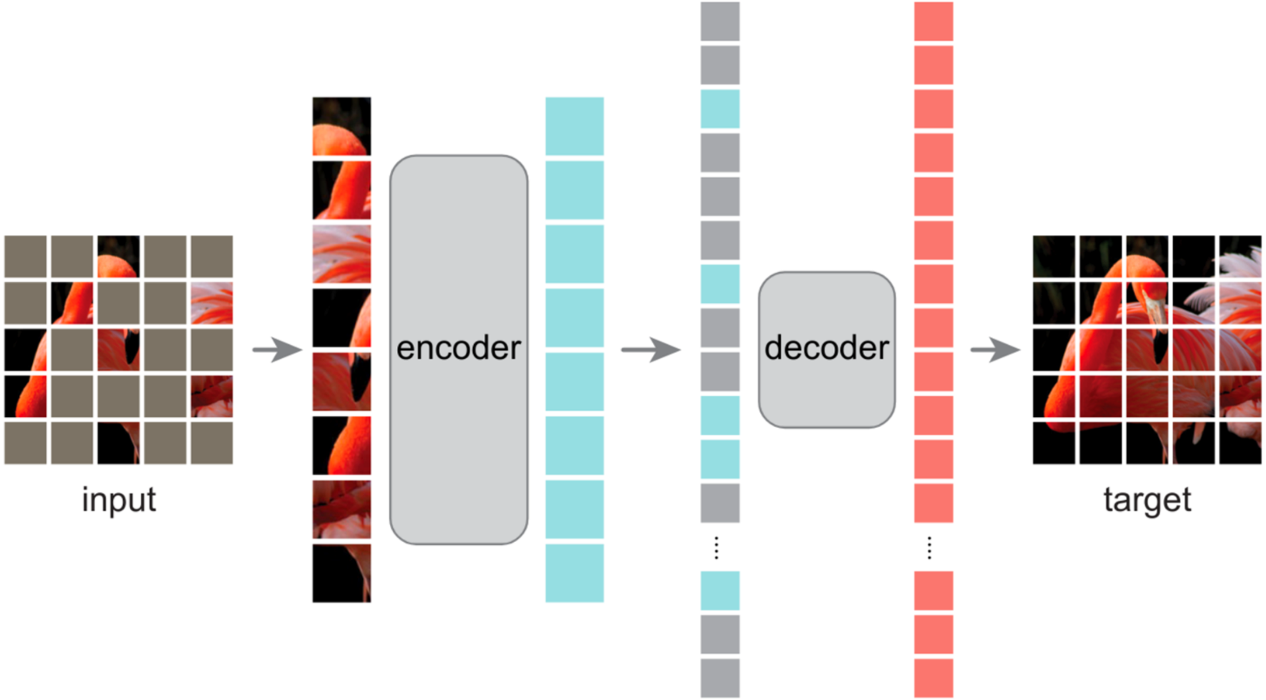
# The importance of Augmentations in SimCLR

What augmentations to use depends on the downstream task, and the dataset itself.

Let’s say you want to classify a data of apples of different colors.

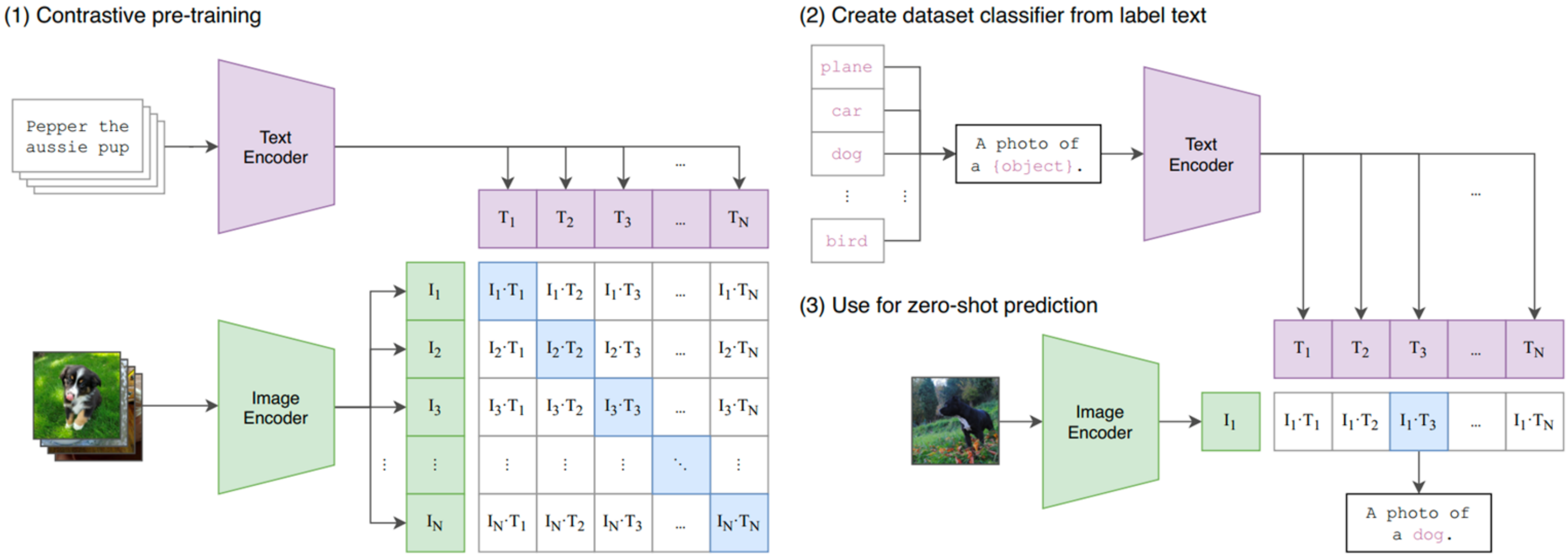
Would you use color augmentation in this task?

# Masked Autoencoders



He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

# CLIP


(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
$T_1$ $T_2$ $T_3$ ... $T_N$
Image Encoder
$I_1$ $I_2$ $I_3$ ⋮ $I_N$
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
⋮ ⋮ ⋮ ⋱ ⋮
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
$T_1$ $T_2$ $T_3$ ... $T_N$
(3) Use for zero-shot prediction
Image Encoder
$I_1$
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.