자연어 처리 및 응용 Day 2,3

The 13th KIAS CAC Summer School on
the Parallel Computing and Artificial Intelligence

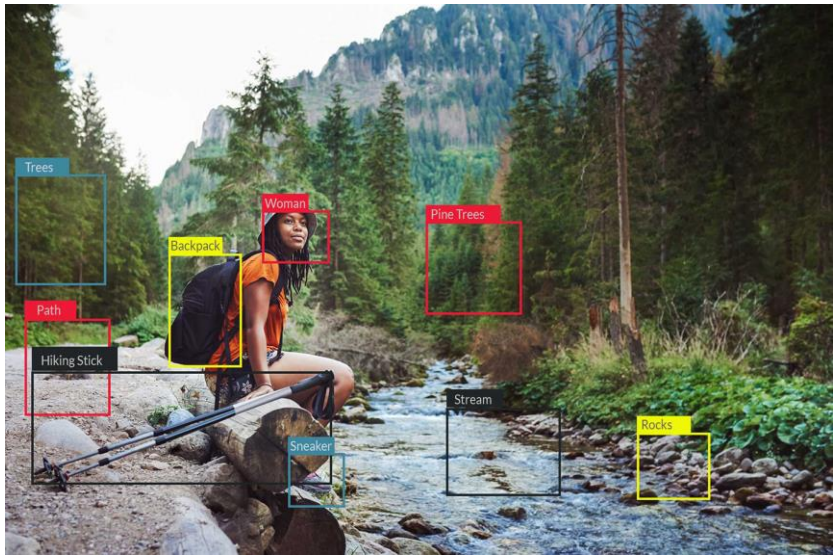
Eun-Sol Kim
한양대학교 컴퓨터소프트웨어학부

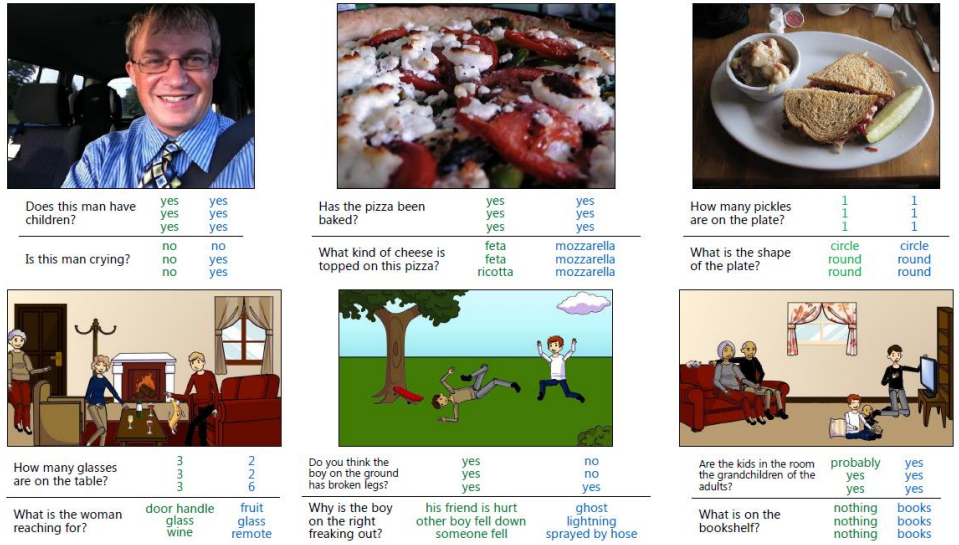
Traditional Approaches

Input: image

Output: Assign image to one of a fixed set of categories

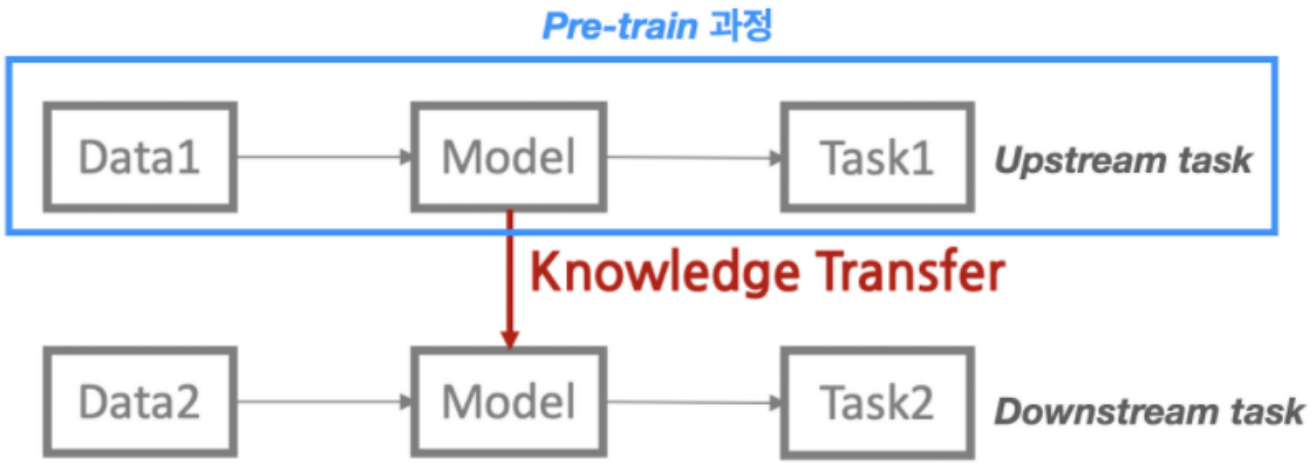
cat
bird
deer
dog
truck



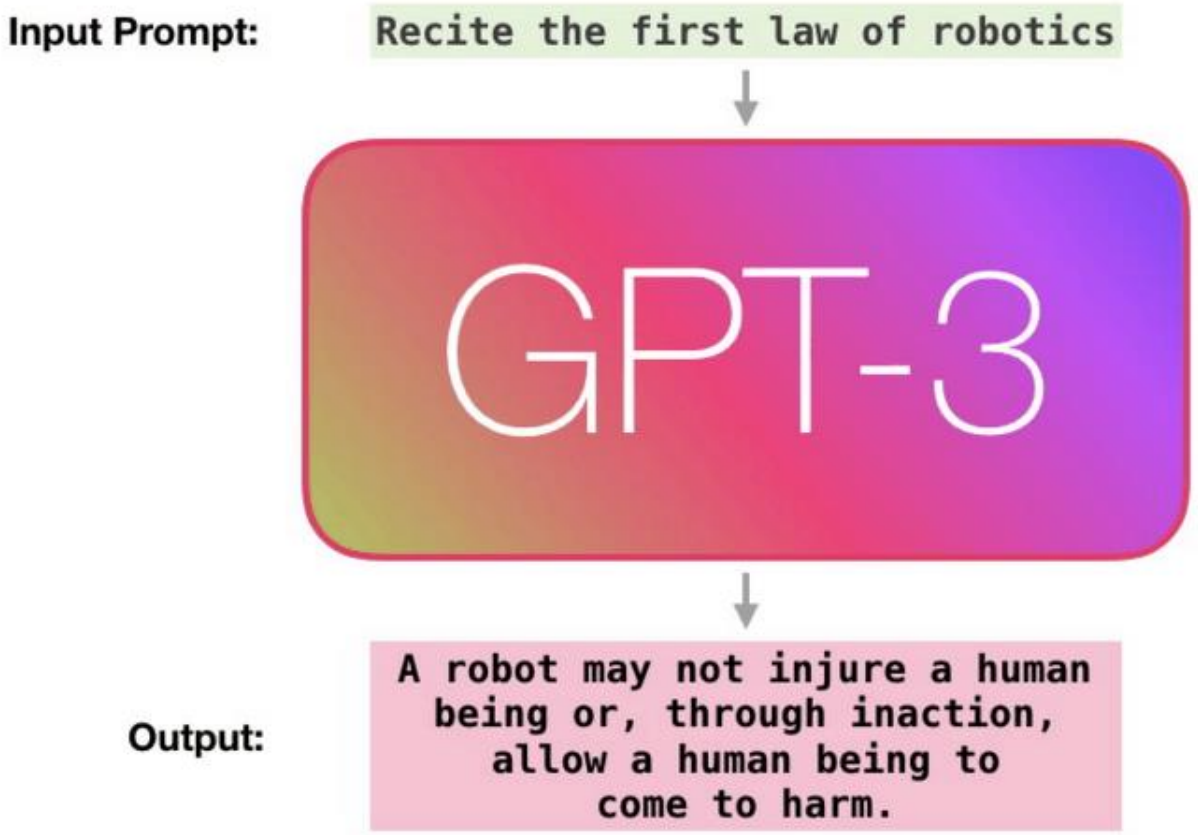


Pre-training, Fine-tuning, Zero-shot Learning

- Fine-tuning
- zero-shot
- one-shot
- Few-shot



GPT-3


Input Prompt:
Recite the first law of robotics
GPT-3
Output:
A robot may not injure a human being or, through inaction, allow a human being to come to harm.

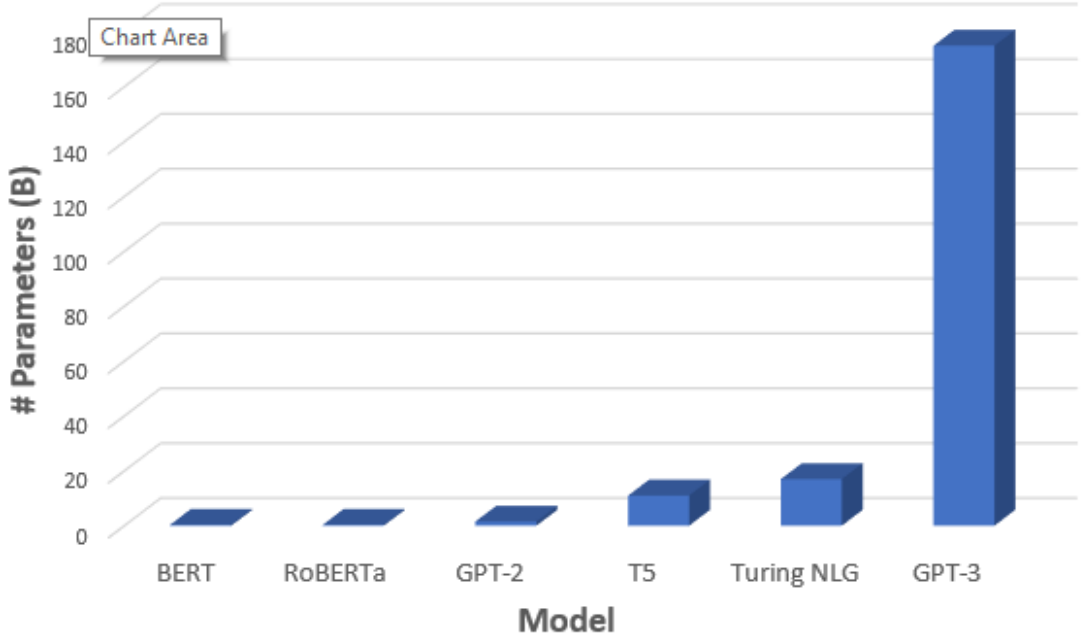
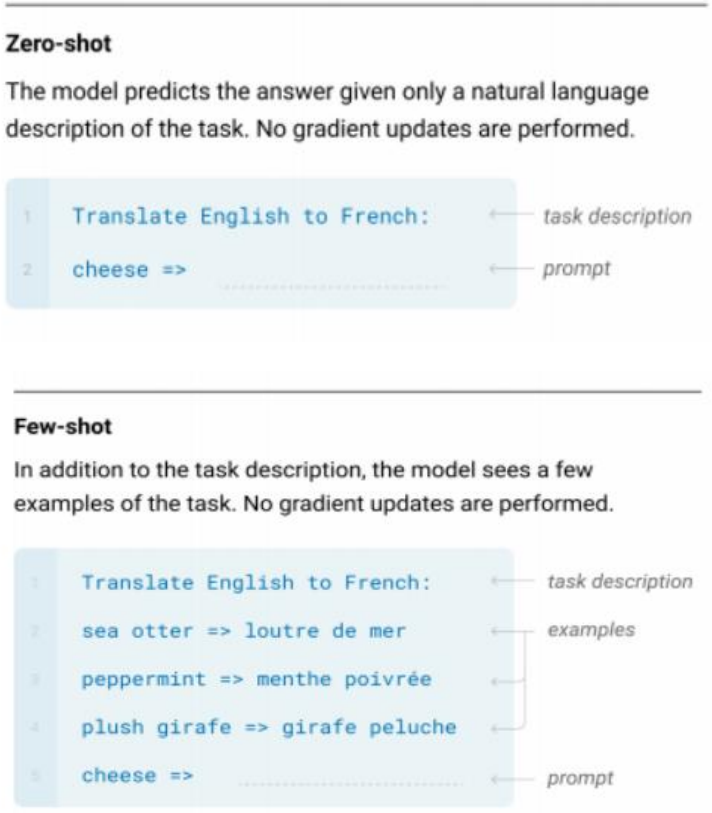
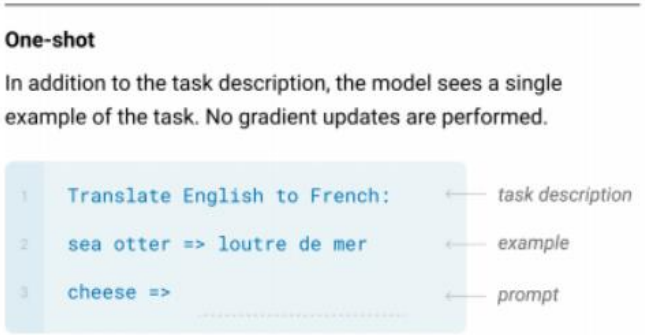


Chart Area
Parameters (B)
180
160
140
120
100
80
60
40
20
0
BERT
RoBERTa
GPT-2
T5
Turing NLG
GPT-3
Model

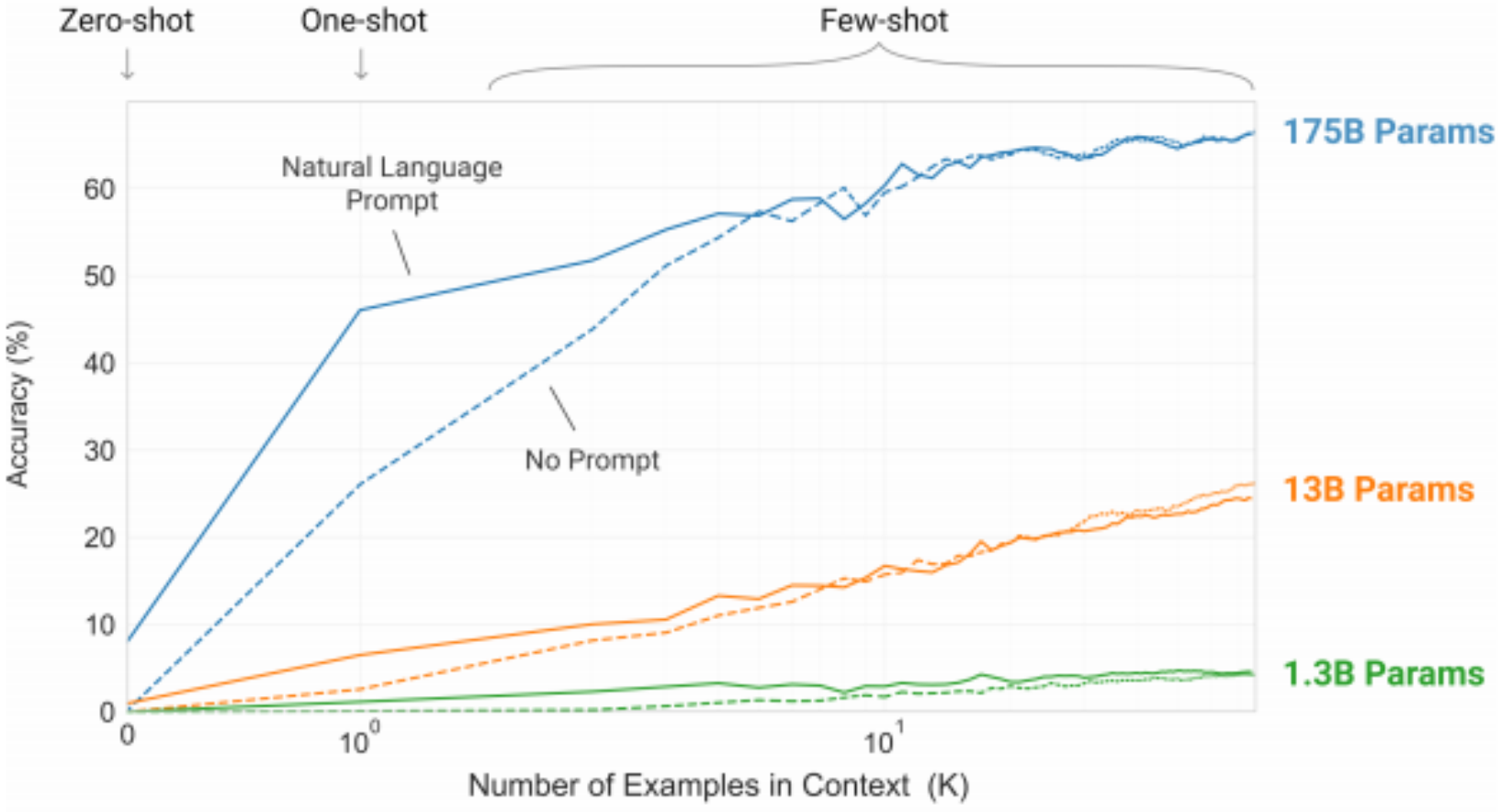
GPT-3



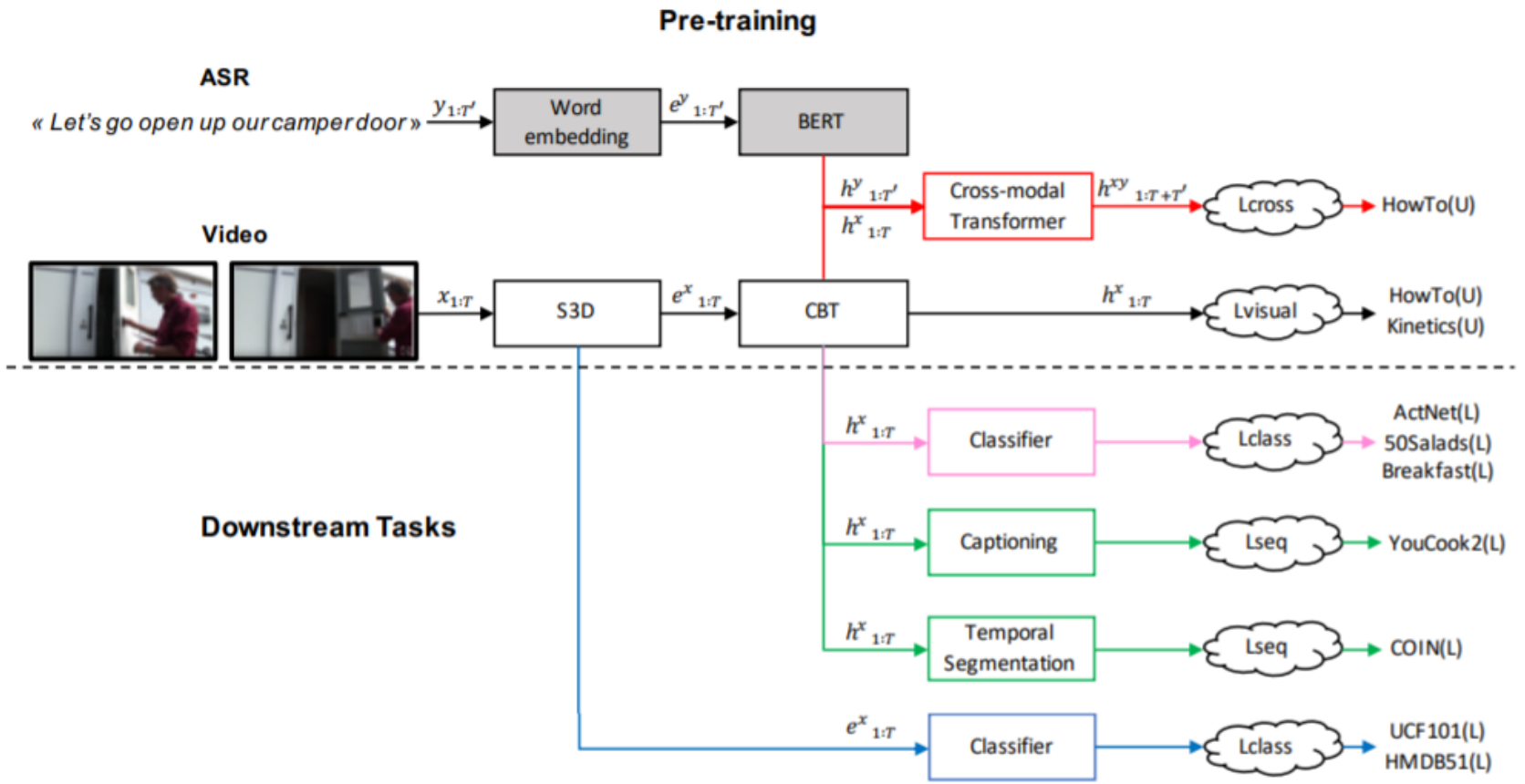


OpenAI, GPT3

GPT-3 성능

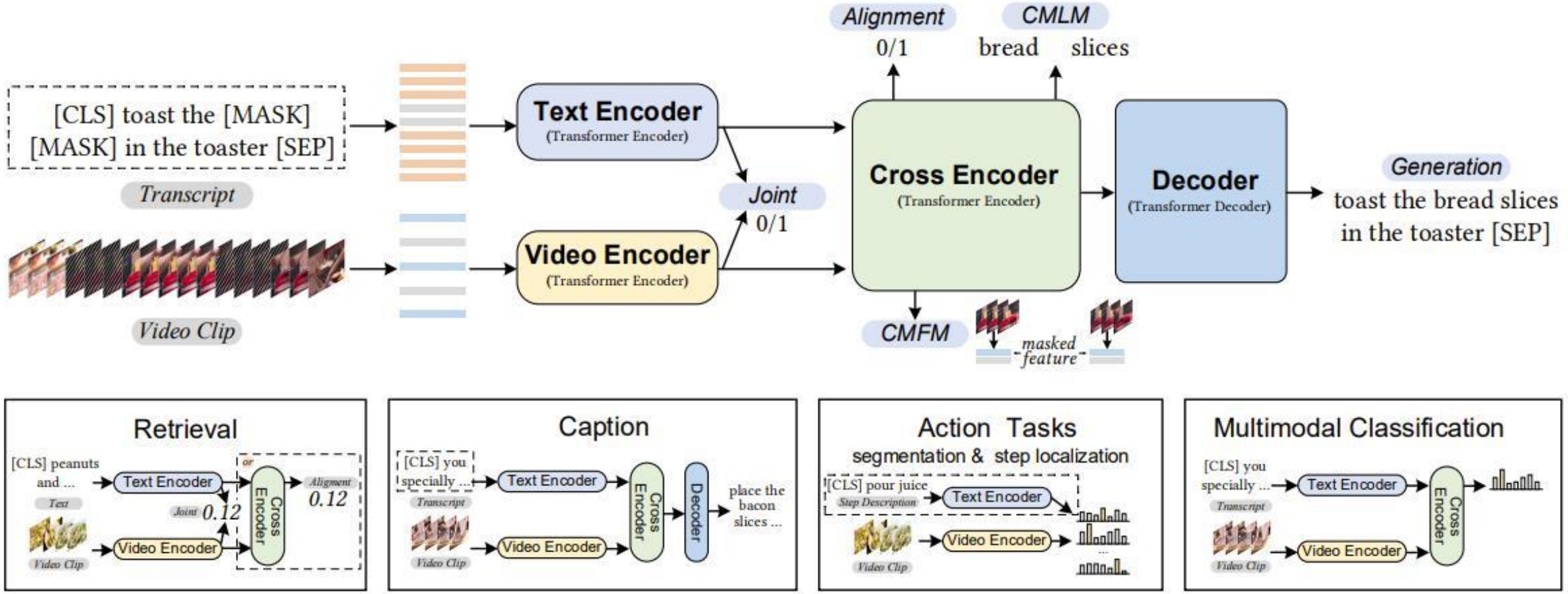

Zero-shot
One-shot
Few-shot
Natural Language Prompt
No Prompt
175B Params
13B Params
1.3B Params
Accuracy (%)
0
10
20
30
40
50
60
0
10^0
10^1
Number of Examples in Context (K)

Pre-training, Fine-tuning, Zero-shot Learning



https://arxiv.org/abs/1906.05743

UniVL


Alignment
0/1
CMLM
bread slices
[CLS] toast the [MASK] [MASK] in the toaster [SEP]
Transcript
Text Encoder
(Transformer Encoder)
Joint
0/1
Video Encoder
(Transformer Encoder)
Cross Encoder
(Transformer Encoder)
Decoder
(Transformer Decoder)
Generation
toast the bread slices in the toaster [SEP]
Video Clip
CMFM
masked feature
Retrieval
[CLS] peanuts and ...
Text
Text Encoder
Video Encoder
Video Clip
Joint 0.12
or
Cross Encoder
Alignment 0.12
Caption
[CLS] you specially ...
Transcript
Text Encoder
Video Encoder
Video Clip
Cross Encoder
Decoder
place the bacon slices ...
Action Tasks
segmentation & step localization
[CLS] pour juice
Step Description
Text Encoder
Video Encoder
Video Clip
Multimodal Classification
[CLS] you specially ...
Transcript
Text Encoder
Video Encoder
Video Clip
Cross Encoder

Self-attention

Self-supervised Learning

Self-attention

Attention is all you need

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Attention is **all you need**

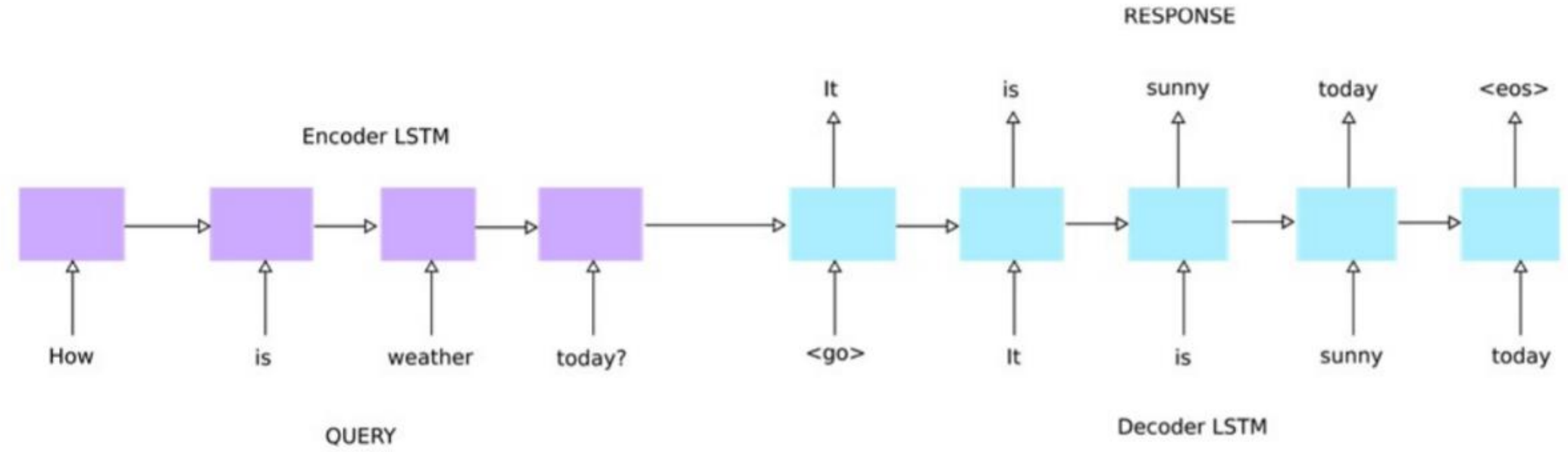
A Vaswani, N Shazeer, N Parmar… - Advances in neural …, 2017 - proceedings.neurips.cc

… the number of **attention** heads and the **attention** key and value dimensions, keeping the amount of computation constant, as described in Section 3.2.2. While single-head **attention** is 0.9 …

☆ Save Cite Cited by 45648 Related articles All 46 versions »

Background - RNN

- Sequence Modeling
- LSTM, GRU



Attention

Published as a conference paper at ICLR 2015

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho **Yoshua Bengio***
Université de Montréal

BackGround - Attention

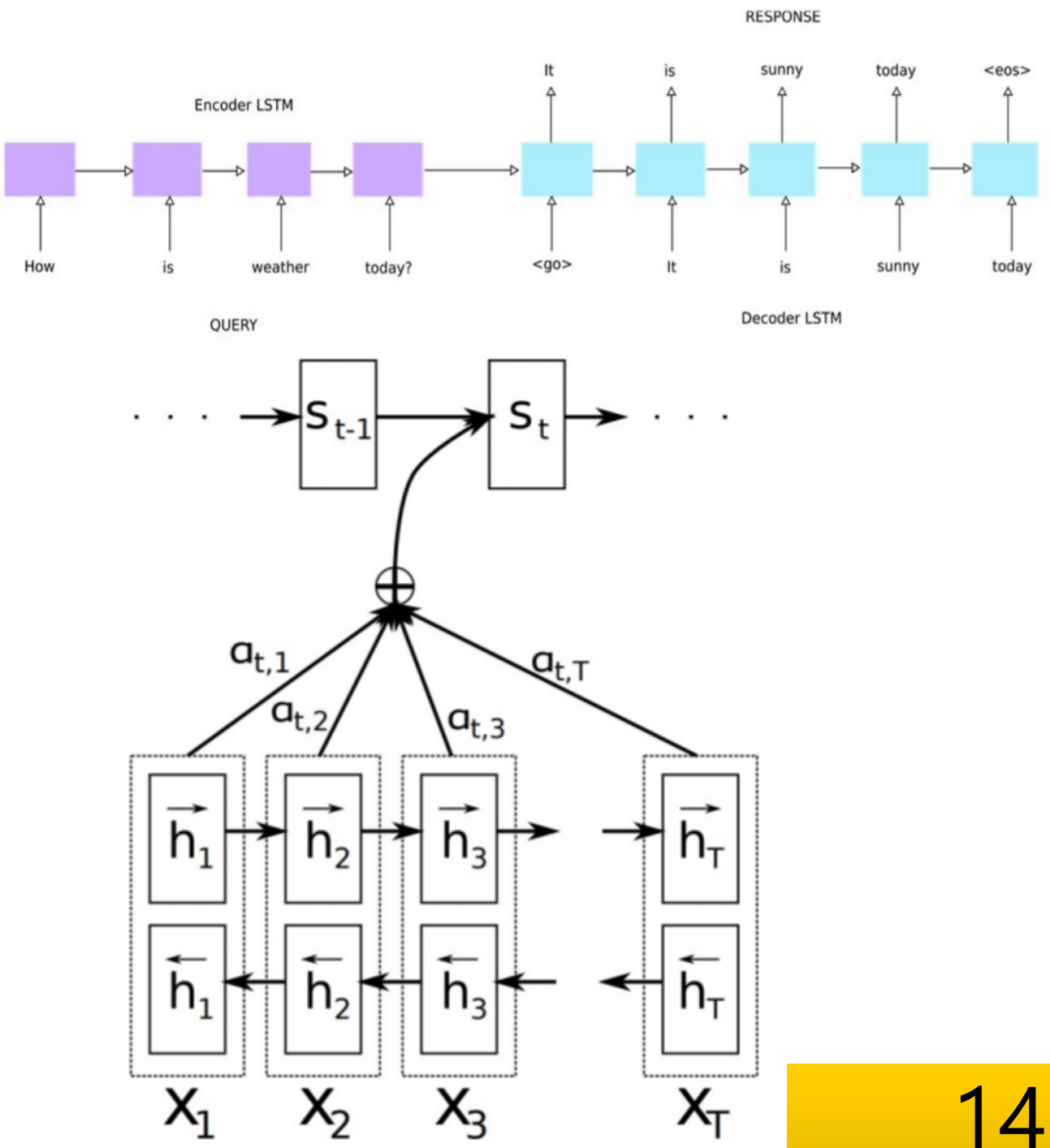
- Attention mechanism
  Neural Machine Translation By Jointly Learning To Align And Translate (ICLR 2015)
- Attention mechanisms are used in conjunction with a recurrent network
- Hidden state s_i
- Annotations h_i
- Context vector c_i
- Energy e_{ij}
- Probability a_{ij}

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

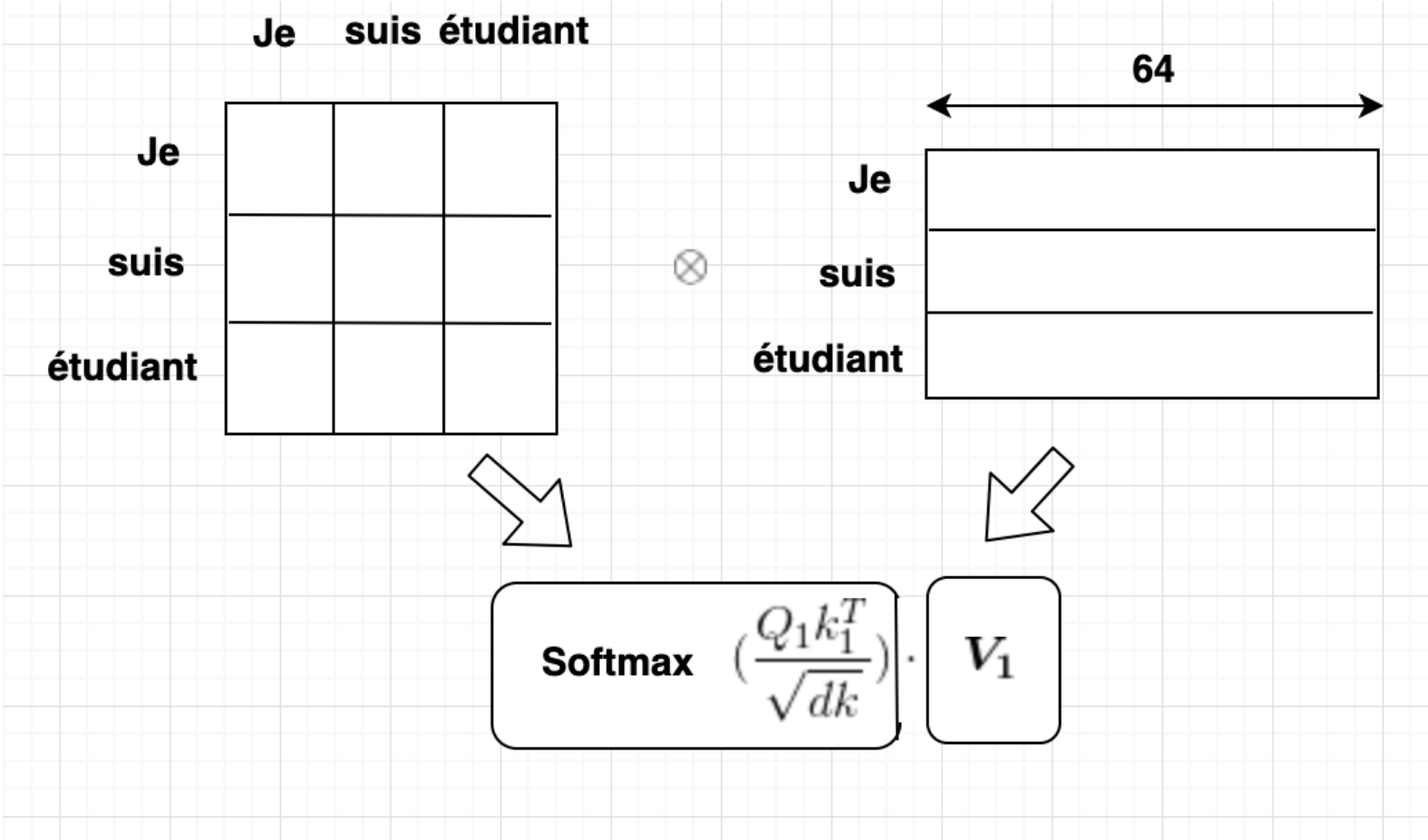
$$e_{ij} = a(s_{i-1}, h_j)$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$



Self-attention

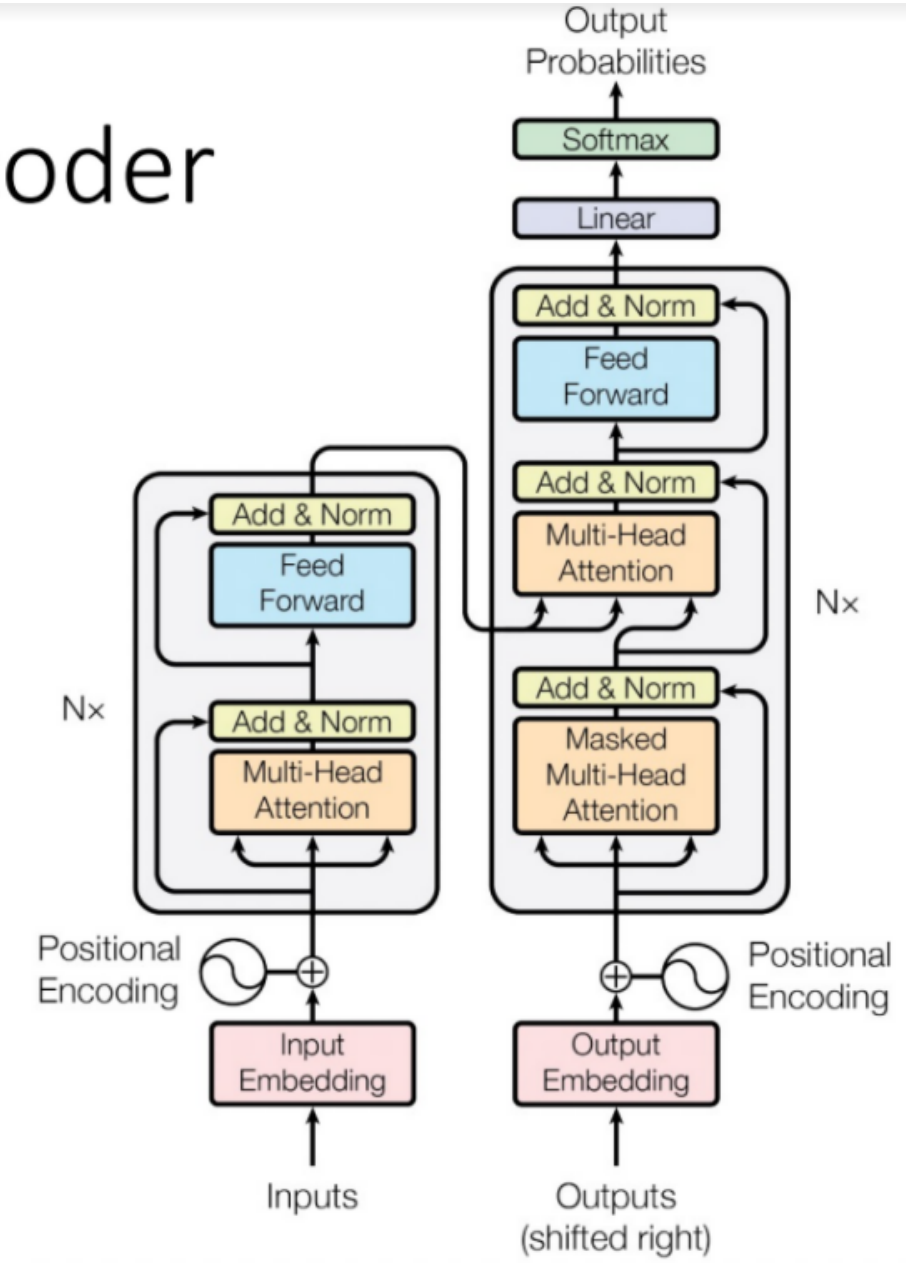
$$Attention(Q, K, V) = softmax\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$



Model Architecture - Encoder

- Composed of a stack of N(=6) identical layers
- Each layer has two sub-layers
  - Multi-head self-attention
  - Simple, positionwise fully connected feed-forward network
- Residual connection
- Layer normalization
- Output of sub-layer =

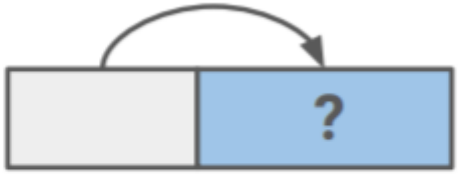
$$LayerNorm(x + Sublayer(x))$$



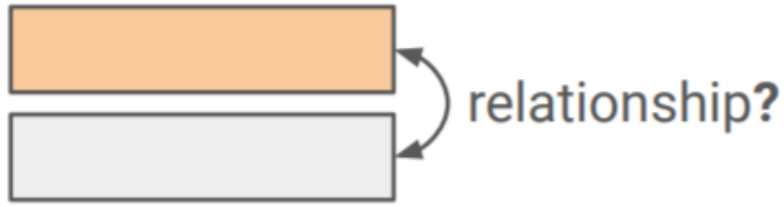
Self-supervised learning

Self-prediction & Contrastive

- Self-prediction
  - Given an individual data sample, the task is to predict one part of the sample given the other part.
- Contrastive
  - Given multiple data samples, the task is to predict the relationship among them.



"Intra-sample" prediction



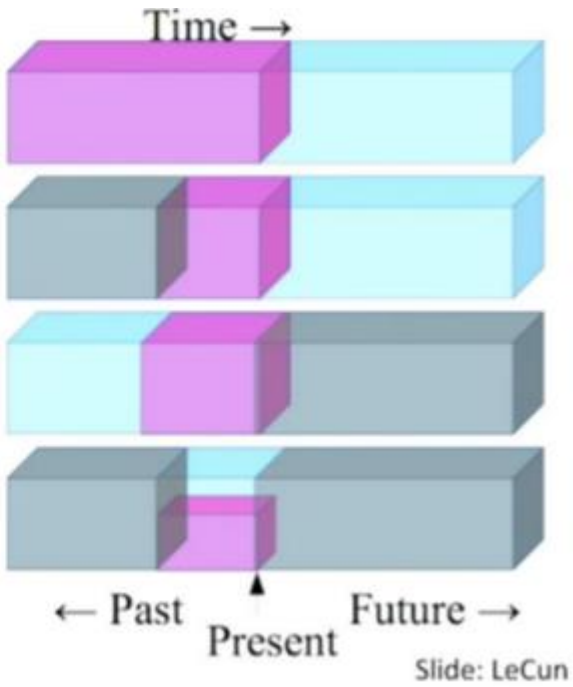
"Inter-sample" prediction

[NeurIPS 2021 Tutorial]

Self-prediction

- to predict a part of the data from the rest while pretending we don't know that part

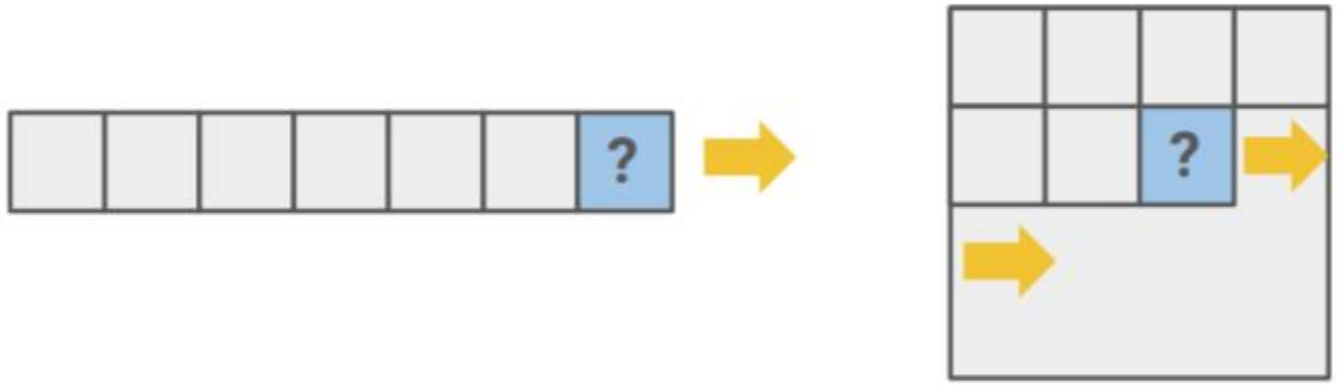
- Predict any part of the input from any other part.
- Predict the future from the past.
- Predict the future from the recent past.
- Predict the past from the present.
- Predict the top from the bottom.
- Predict the occluded from the visible
- Pretend there is a part of the input you don't know and predict that.



Slide: LeCun

Self-prediction: Autoregressive Prediction

The autoregressive model predicts future behavior based on past behavior. Any data that comes with an innate sequential order can be modeled with regression.



Examples:

- Audio (WaveNet, WaveRNN)
- Autoregressive language modeling (GPT, XLNet)
- Images in raster scan (PixelCNN, PixelRNN, iGPT)

[NeurIPS 2021 Tutorial]

Self-prediction: Masked Generation

We mask a random portion of information and pretend it is missing, irrespective of the natural sequence. The model learns to predict the missing portion given other unmasked information.

Examples:

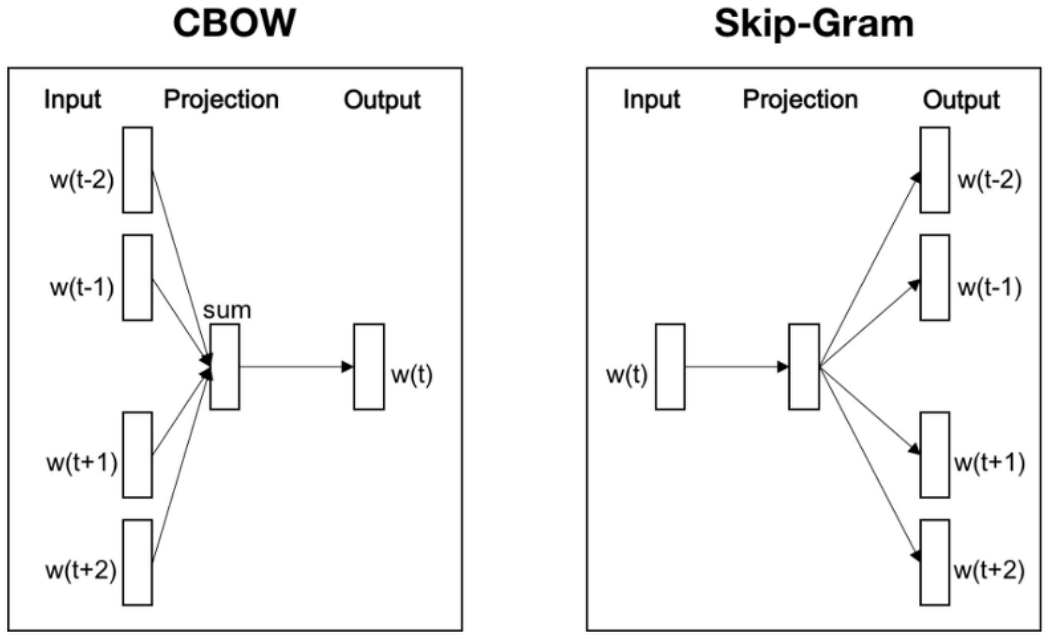
- Masked language modeling (BERT)
- Images with masked patch (denoising autoencoder, context autoencoder, colorization)

[NeurIPS 2021 Tutorial]

Early work

- Word2Vec: Self-Supervised Learning for Language

- Word embeddings to map words to vectors
- CBOW & Skip-gram (Mikolov et al. 2013)
  - Neighboring words → middle word (CBOW)
  - Word → neighboring words (skip-gram)



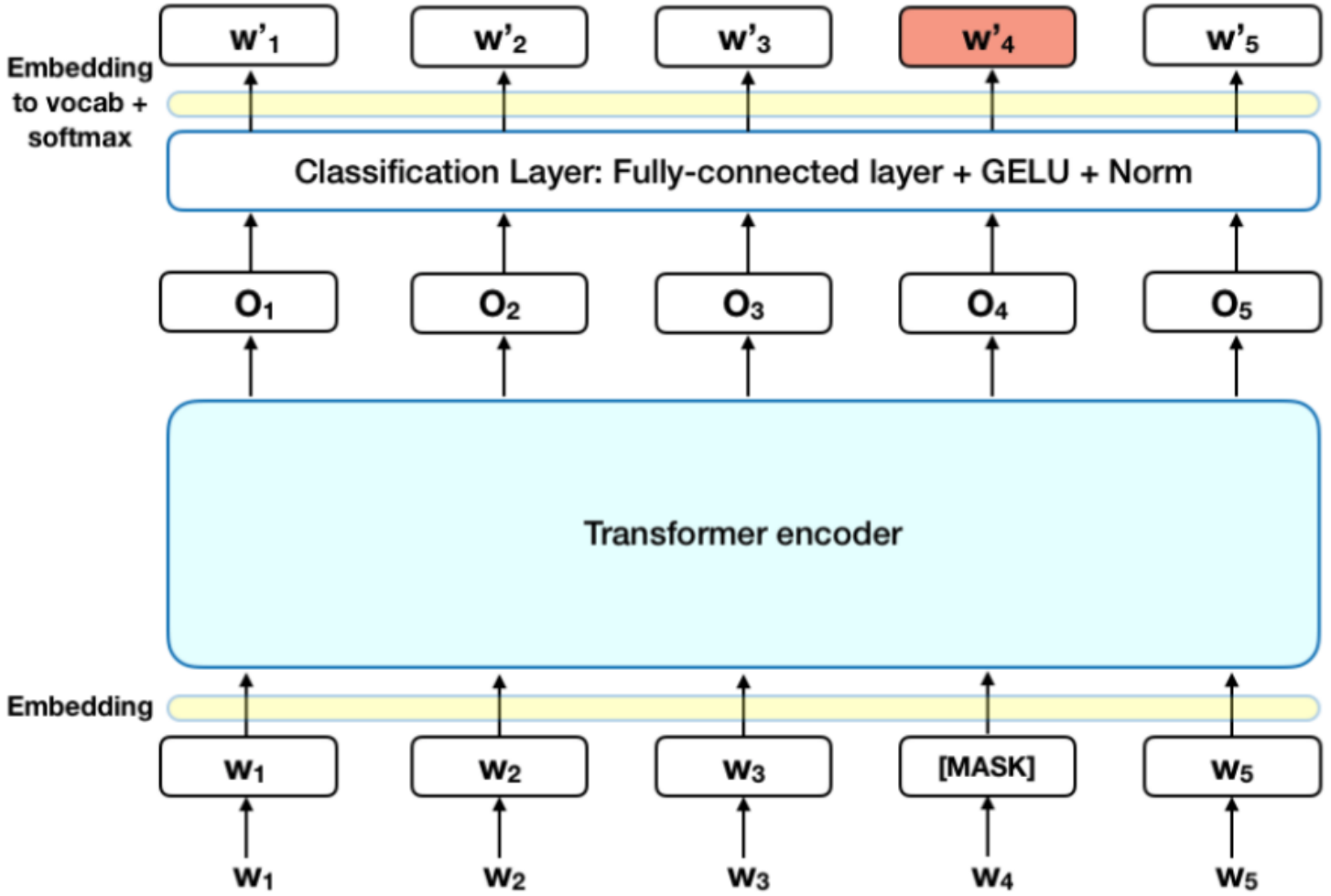
BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

- Embedding methods
- Masked Language Model

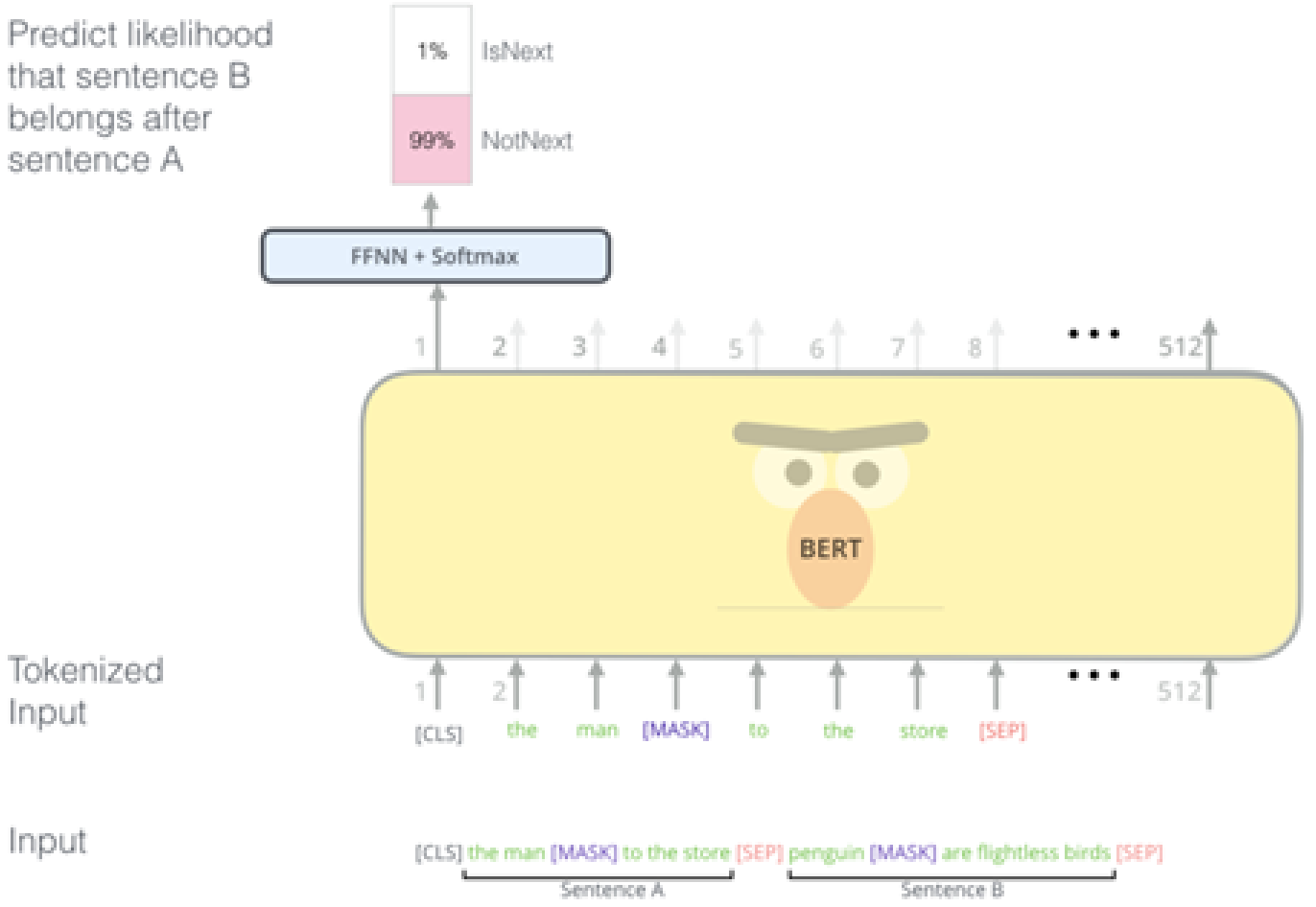
BERT: Training

- Masked Language Model

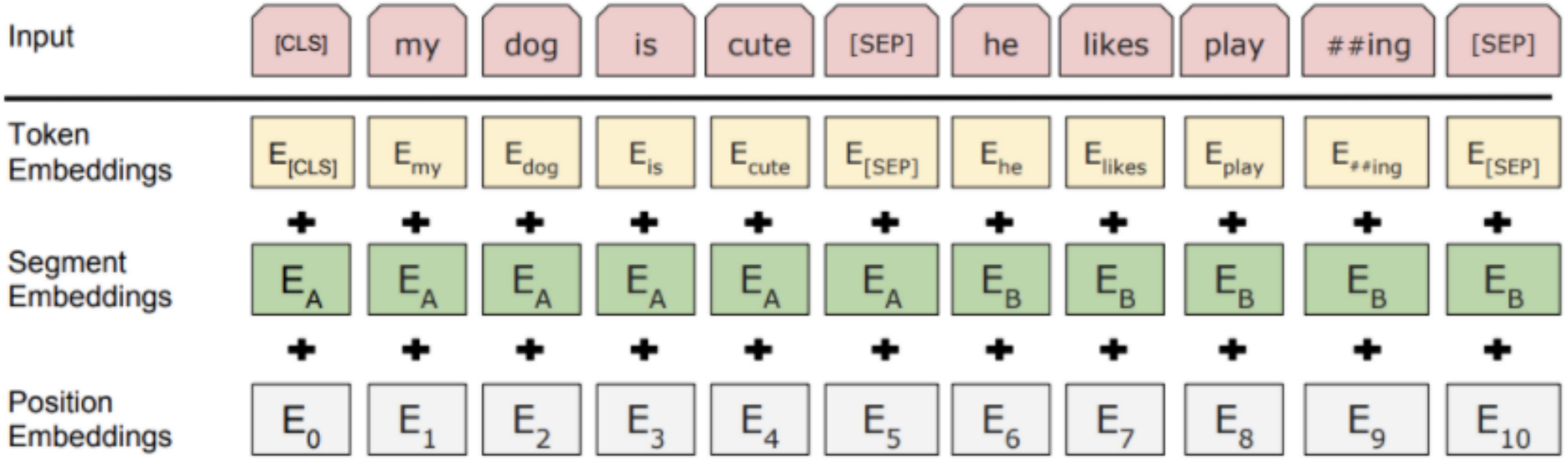


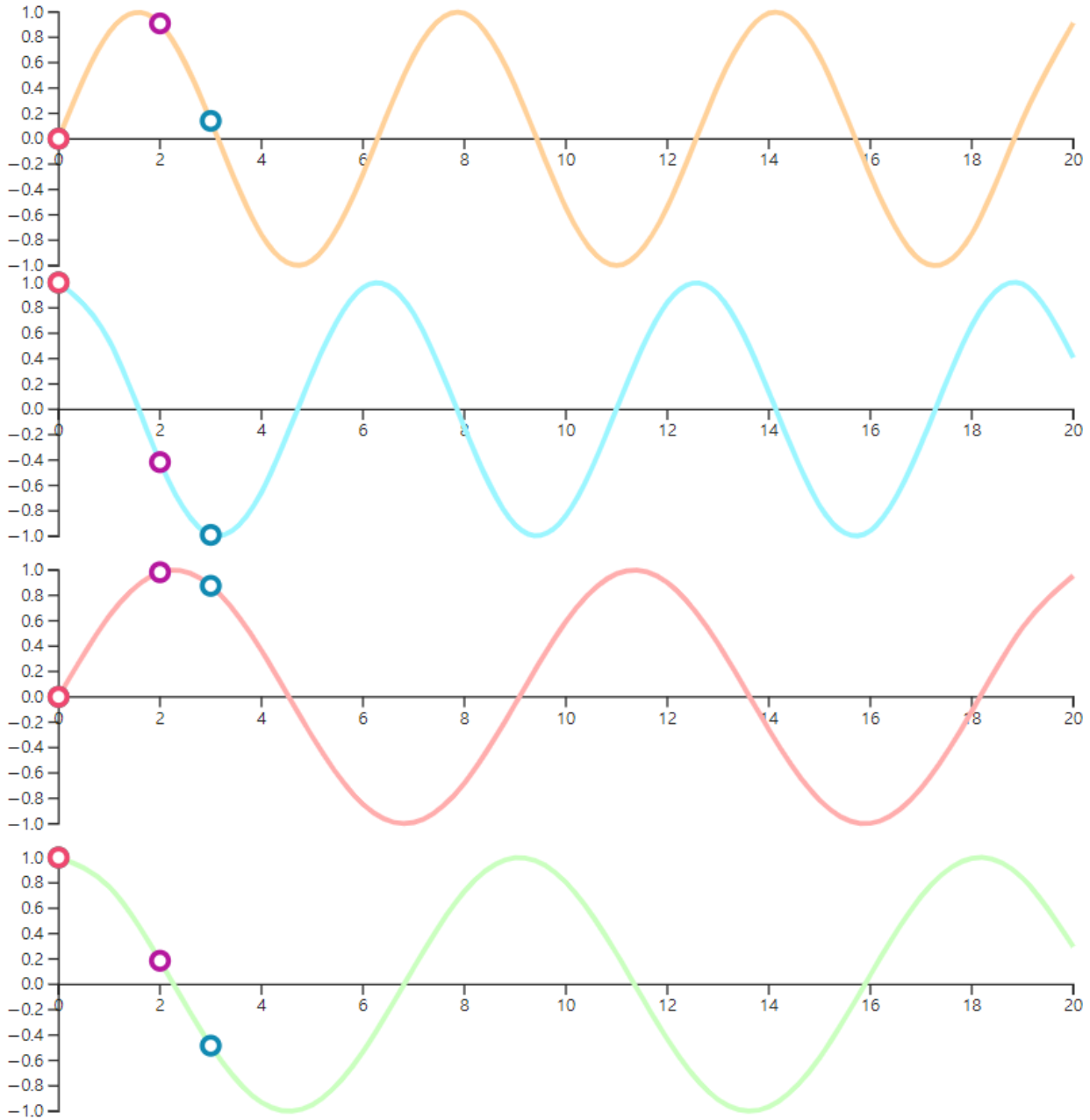
BERT: Training

- Next Sentence Prediction



BERT: Input


Input
[CLS] my dog is cute [SEP] he likes play ##ing [SEP]
Token Embeddings
E[CLS] Emy Edog Eis Ecute E[SEP] Ehe Elikes Eplay E##ing E[SEP]
Segment Embeddings
EA EA EA EA EA EA EB EB EB EB EB
Position Embeddings
E0 E1 E2 E3 E4 E5 E6 E7 E8 E9 E10



p0	p1	p2	p3	
0.000	0.000	0.909	0.141	**i=0**
1.000	1.000	-0.416	-0.990	**i=1**
0.000	0.000	0.983	0.875	**i=2**
1.000	1.000	0.186	-0.484	**i=3**

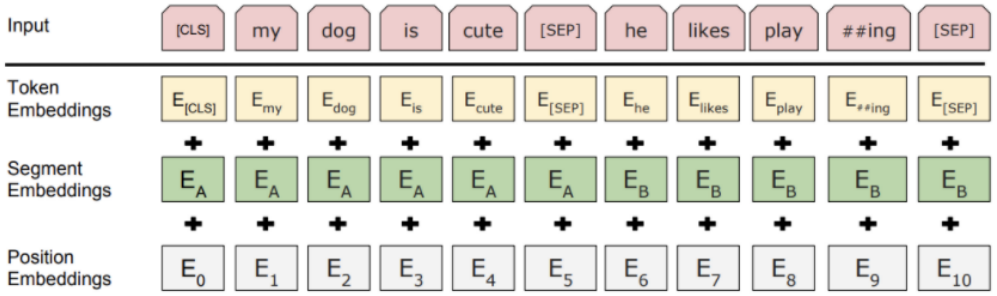
Positional Encoding

$$PE_{(pos,2i)} = sin(\frac{pos}{10000^{2i/d_{\text{model}}}})$$

$$PE_{(pos,2i+1)} = cos(\frac{pos}{10000^{2i/d_{\text{model}}}})$$

Settings: d = 50

The value of each positional encoding depends on the *position* (*pos*) and *dimension* (*d*). We calculate result for every *index* (*i*) to get the whole vector.


Input
[CLS] my dog is cute [SEP] he likes play ##ing [SEP]
Token Embeddings
E[CLS] Emy Edog Eis Ecute E[SEP] Ehe Elikes Eplay E##ing E[SEP]
Segment Embeddings
EA EA EA EA EA EA EB EB EB EB EB
Position Embeddings
E0 E1 E2 E3 E4 E5 E6 E7 E8 E9 E10

Application1
(Text) Question Answering

Harvard_University

The Stanford Question Answering Dataset

Established originally by the Massachusetts legislature and soon thereafter named for John Harvard (its first benefactor), Harvard is the United States' oldest institution of higher learning, and the Harvard Corporation (formally, the President and Fellows of Harvard College) is its first chartered corporation. Although never formally affiliated with any denomination, the early College primarily trained Congregationalist and Unitarian clergy. Its curriculum and student body were gradually secularized during the 18th century, and by the 19th century Harvard had emerged as the central cultural establishment among Boston elites. Following the American Civil War, President Charles W. Eliot's long tenure (1869–1909) transformed the college and affiliated professional schools into a modern research university; Harvard was a founding member of the Association of American Universities in 1900. James Bryant Conant led the university through the Great Depression and World War II and began to reform the curriculum and liberalize admissions after the war. The undergraduate college became coeducational after its 1977 merger with Radcliffe College.

What individual is the school named after?
Ground Truth Answers: John Harvard | John Harvard | John Harvard
Prediction: John Harvard

When did the undergraduate program become coeducational?
Ground Truth Answers: 1977 | 1977 | 1977
Prediction: 1977

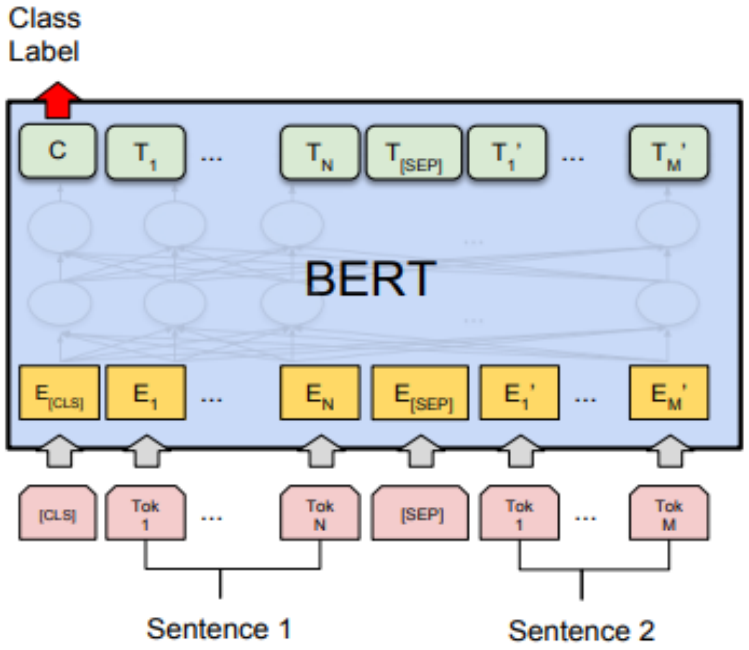
What was the name of the leader through the Great Depression and World War II?
Ground Truth Answers: James Bryant Conant | James Bryant Conant | James Bryant Conant
Prediction: James Bryant Conant

https://rajpurkar.github.io/SQuAD-explorer/

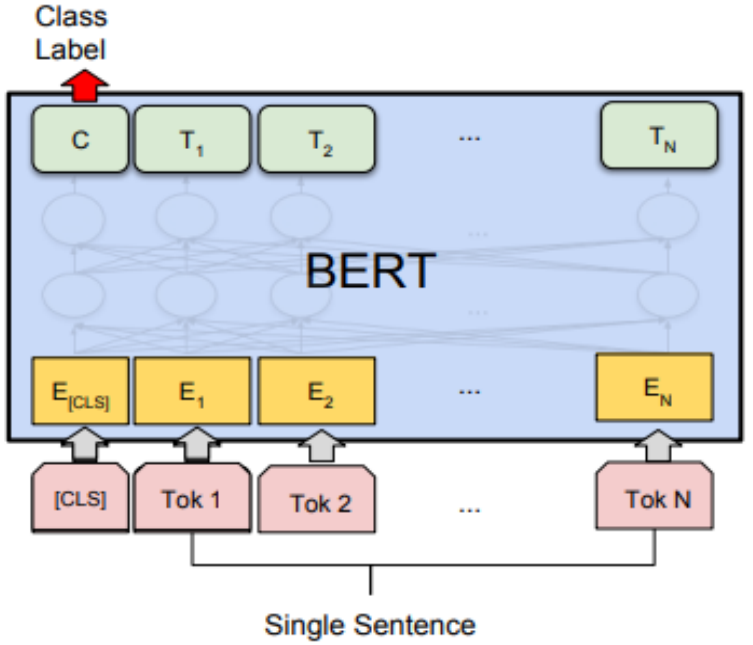
Application2
Natural Language Inference (NLI)

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

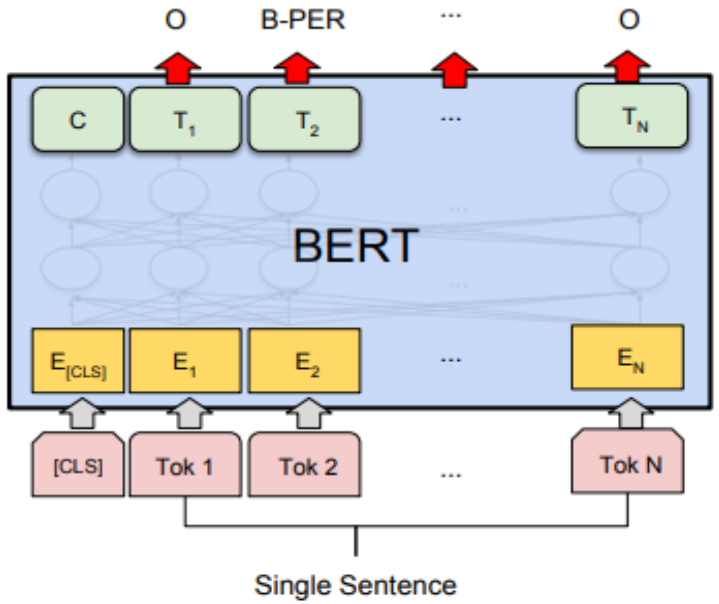
BERT: Experiments



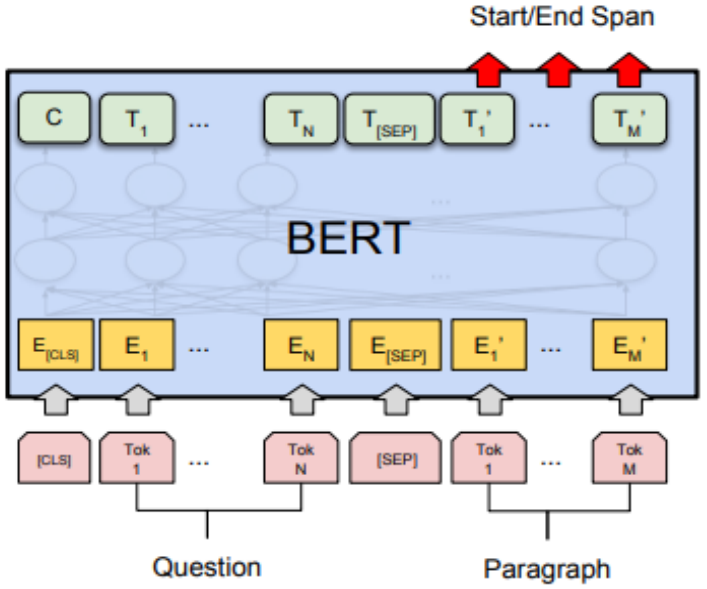
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER



(c) Question Answering Tasks: SQuAD v1.1

Extension
BERT for Images

- An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (ViT)

Published as a conference paper at ICLR 2021

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

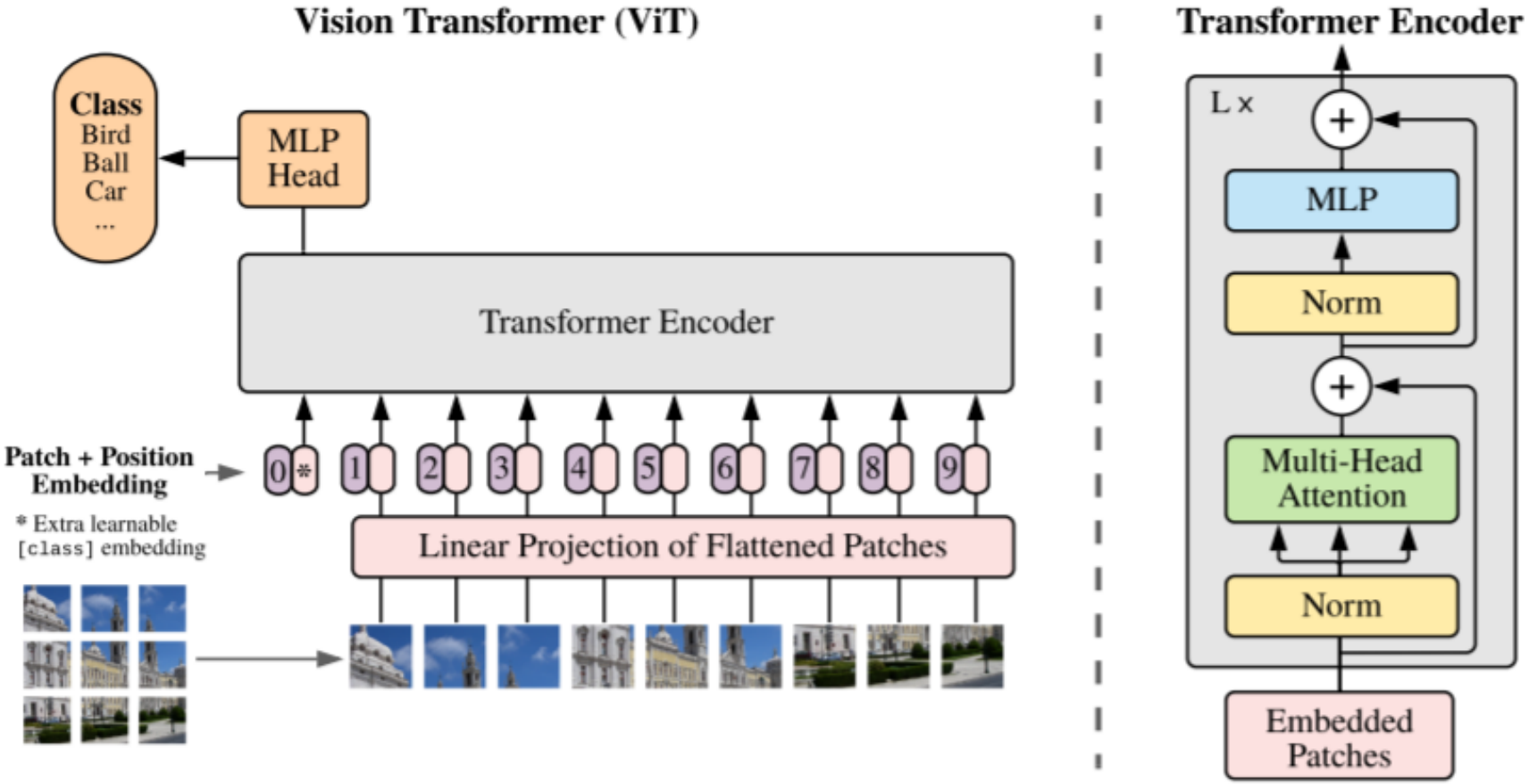
Alexey Dosovitskiy[*,†], **Lucas Beyer**[*], **Alexander Kolesnikov**[*], **Dirk Weissenborn**[*], **Xiaohua Zhai**[*], **Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby**[*,†]

[*]equal technical contribution, [†]equal advising

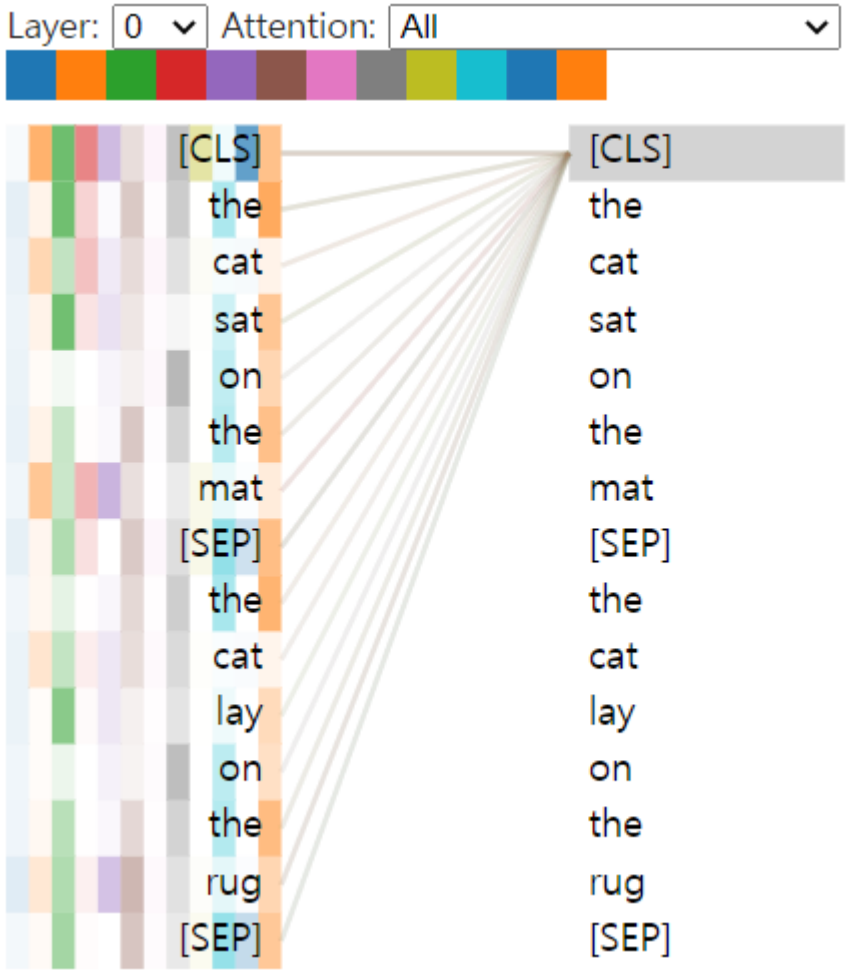
Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

Vision Transformer



Analyze Attention Visualization



https://github.com/jessevig/bertviz

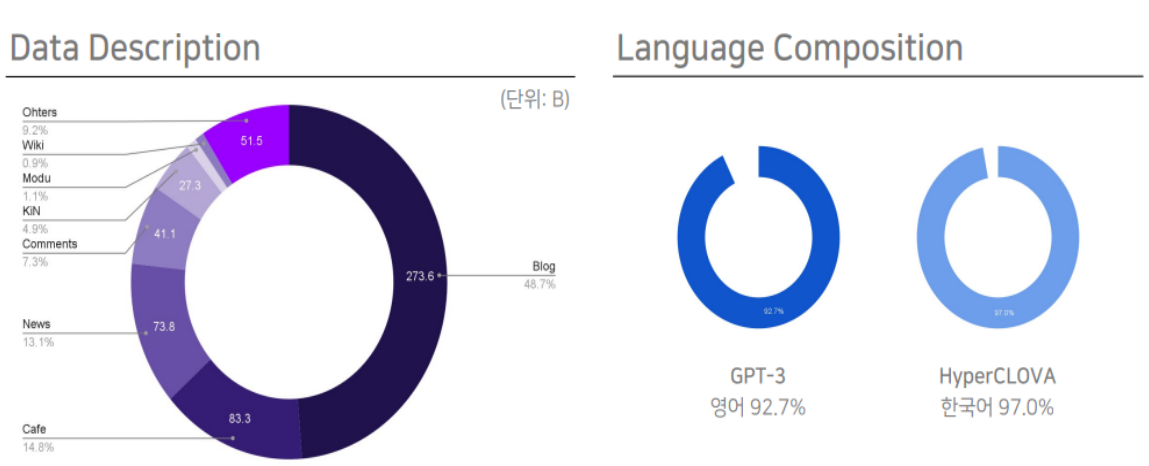
한국어를 위한 BERT

네이버 HyperCLOVA

1.1 HyperCLOVA 언어 모델 개요

DEVIEW 2021

562B token의 다양한 한국어 데이터



김형석, 이상우 CLOVA Conversation / AI Lab – DEVIEW 2021
https://deview.naver.com

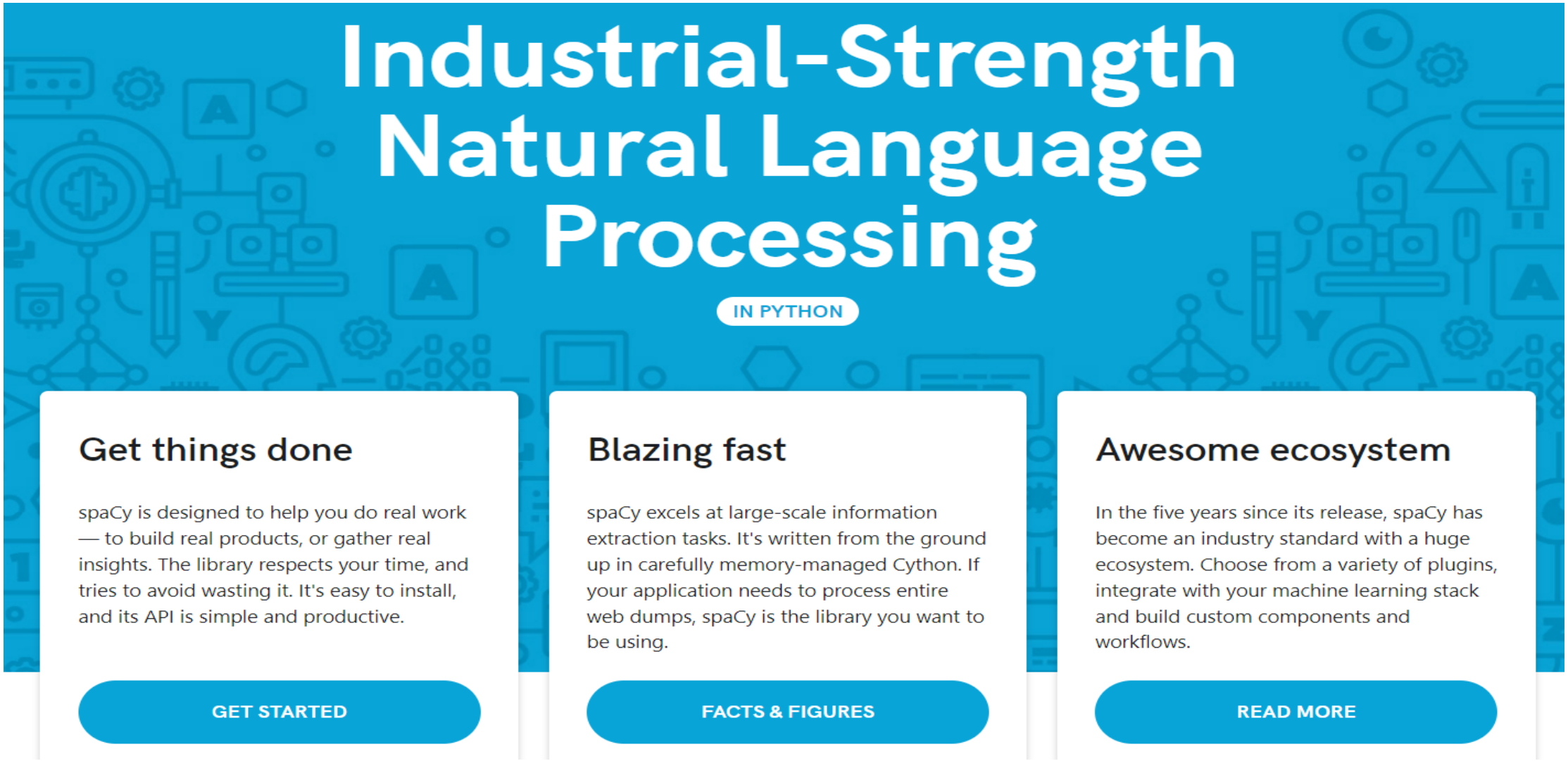
카카오브레인 KoGPT

KoGPT6B-ryan1.5b

- [huggingface][kakaobrain/kogpt][KoGPT6B-ryan1.5b]
- [huggingface][kakaobrain/kogpt][KoGPT6B-ryan1.5b-float16]

Hyperparameter	Value
$n_{parameters}$	6,166,502,400
n_{layers}	28
d_{model}	4,096
d_{ff}	16,384
n_{heads}	16
d_{head}	256
n_{ctx}	2,048
n_{vocab}	64,512
Positional Encoding	Rotary Position Embedding (RoPE)
RoPE Dimensions	64

Minor Tips – NLTK, Spacy

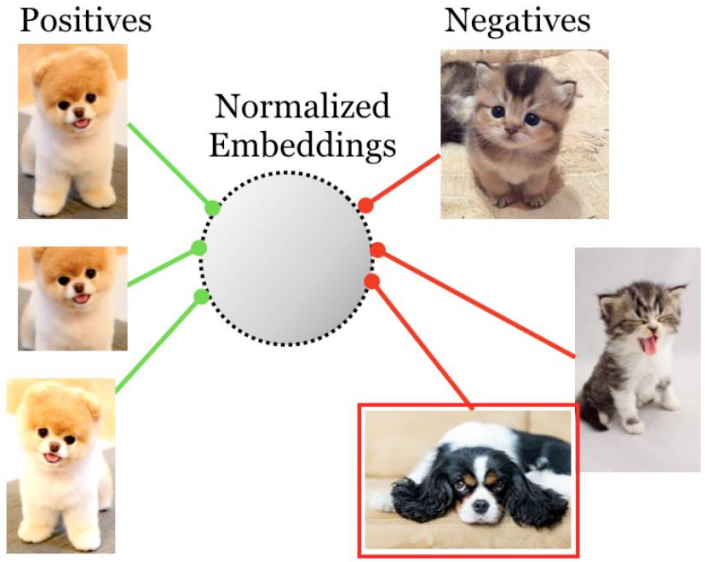


Contrastive Learning

흐름 및 정의

Contrastive Loss

비슷한 데이터들은 feature space 상에서 가깝게 하고
상관없는 데이터들은 feature space 상에서 멀게 한다는
매우 간단한 아이디어!



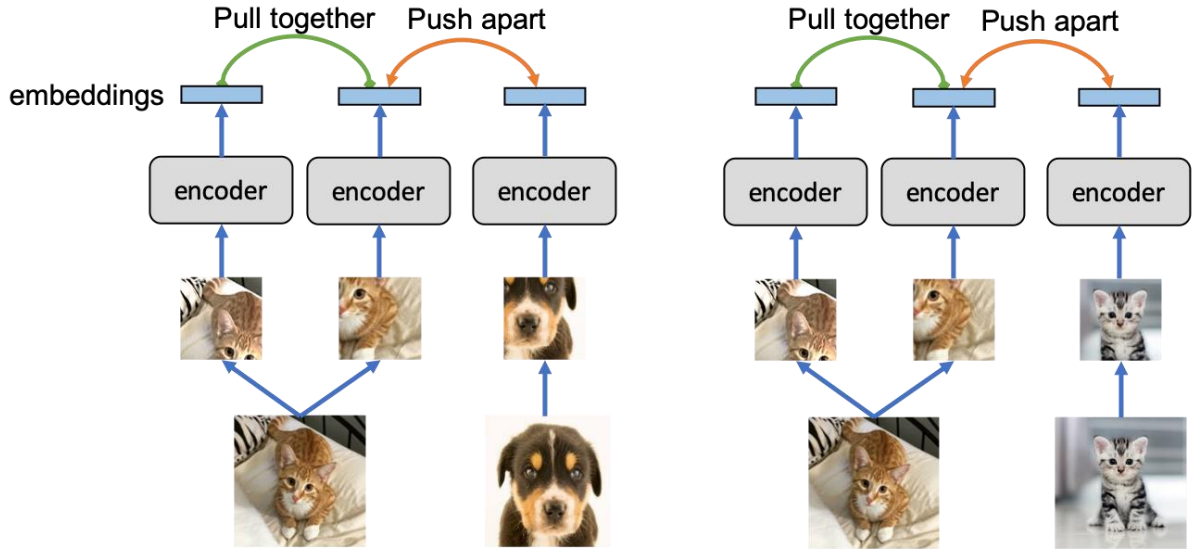
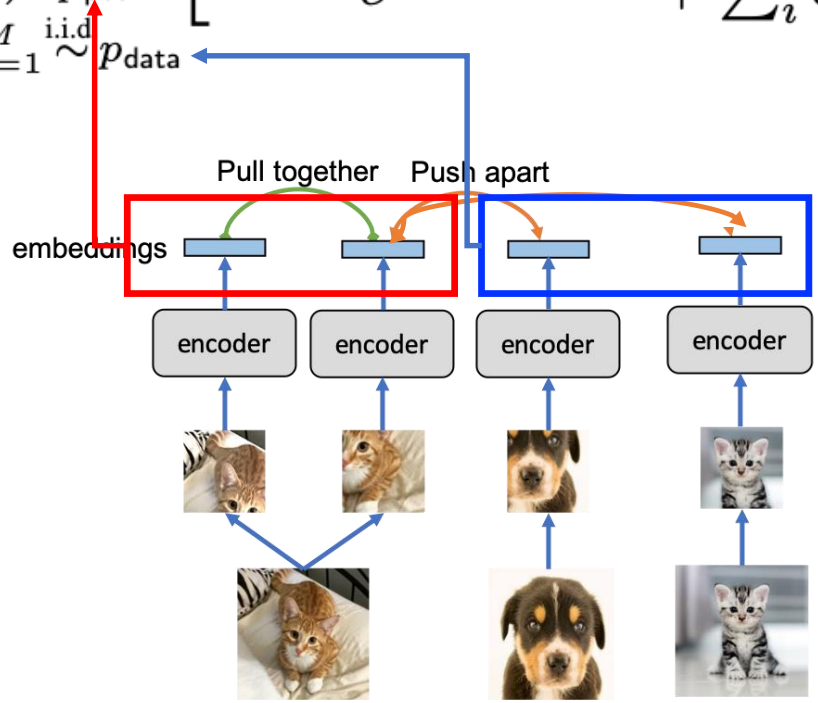


Figure 1 (a) (b)

https://blog.einstein.ai/prototypical-contrastive-learning-pushing-the-frontiers-of-unsupervised-learning/

Contrastive Loss

$$\mathcal{L}_{\text{contrastive}}(f;\tau,M) \triangleq \underset{\substack{(x,y)\sim p_{\text{pos}} \\ \{x_i^-\}_{i=1}^{M} \overset{\text{i.i.d}}{\sim} p_{\text{data}}}}{\mathbb{E}} \left[-\log \frac{e^{f(x)^\top f(y)/\tau}}{e^{f(x)^\top f(y)/\tau} + \sum_i e^{f(x_i^-)^\top f(y)/\tau}} \right]$$



Contrastive Loss

$$\mathcal{L}_{\text{cont}}(\mathbf{x}_i, \mathbf{x}_j, \theta) = \mathbb{1}[y_i = y_j] \|f_\theta(\mathbf{x}_i) - f_\theta(\mathbf{x}_j)\|_2^2 + \mathbb{1}[y_i \neq y_j] \max(0, \epsilon - \|f_\theta(\mathbf{x}_i) - f_\theta(\mathbf{x}_j)\|_2)^2$$

minimize

maximize

사실 그리 새로운 개념은 아닙니다.

Nonlinear Dimensionality Reduction

- 2000, Science
  - Isomap, LLE
  - Local Data points 사이의 정보를 보존
    - Isomap: Geodesic Distance Structure
    - LLE: Inner-product Structure

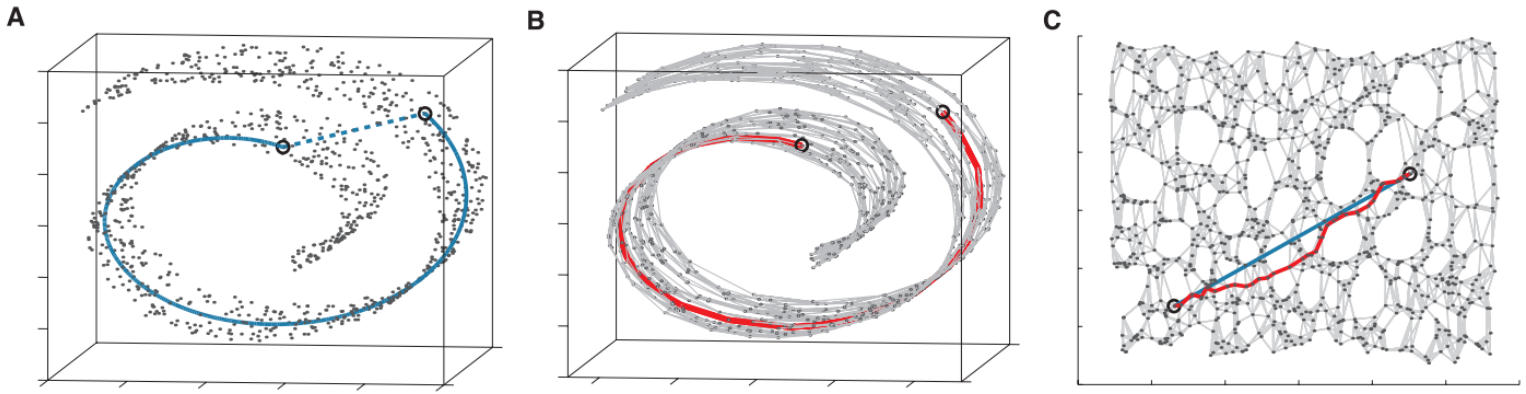
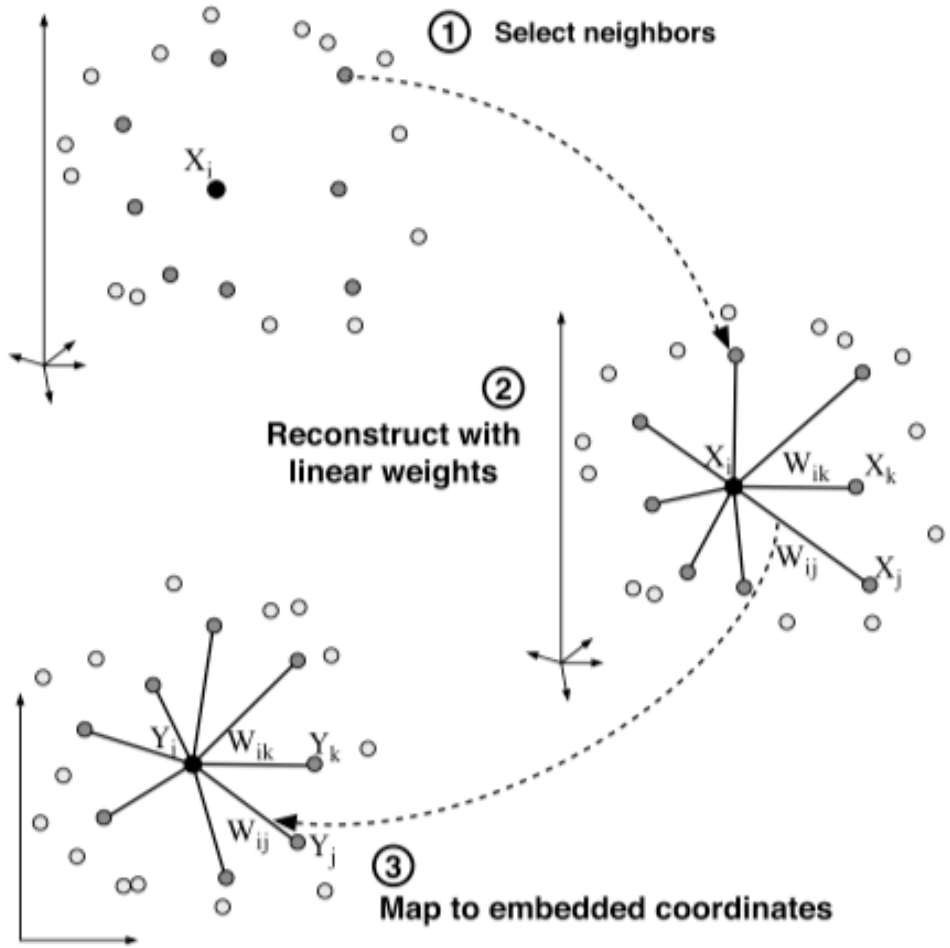
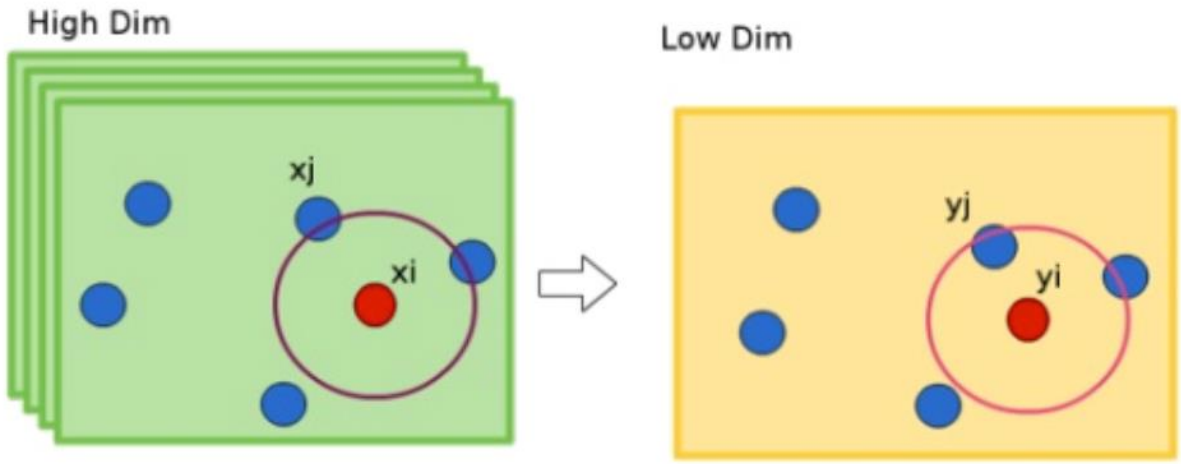


Fig. 3. The "Swiss roll" data set, illustrating how Isomap exploits geodesic paths for nonlinear dimensionality reduction. (**A**) For two arbitrary points (circled) on a nonlinear manifold, their Euclidean distance in the high-dimensional input space (length of dashed line) may not accurately reflect their intrinsic similarity, as measured by geodesic distance along the low-dimensional manifold (length of solid curve). (**B**) The neighborhood graph G constructed in step one of Isomap (with $K = 7$ and $N =$ 1000 data points) allows an approximation (red segments) to the true geodesic path to be computed efficiently in step two, as the shortest path in G. (**C**) The two-dimensional embedding recovered by Isomap in step three, which best preserves the shortest path distances in the neighborhood graph (overlaid). Straight lines in the embedding (blue) now represent simpler and cleaner approximations to the true geodesic paths than do the corresponding graph paths (red).



t-SNE



$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)} \qquad q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)}$$

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

그 동안 많이 사용되지 않았던 이유

Real-world data는 High-dimensional space에 존재하기 때문에
Nearest data points 사이의 거리가 애초에 멀고

Large-scale dataset을 다루는데 있어서
Nearest-neighbor를 찾는 계산량이 큰 문제

Contrast Loss (2006)

- neighbors are pulled together and non-neighbors are pushed apart

Dimensionality Reduction by Learning an Invariant Mapping

Raia Hadsell, Sumit Chopra, Yann LeCun
The Courant Institute of Mathematical Sciences
New York University, 719 Broadway, New York, NY 1003, USA.
http://www.cs.nyu.edu/~yann

(November 2005. To appear in CVPR 2006)

Contrast Loss (2006)

$$D_W(\vec{X}_1, \vec{X}_2) = \|G_W(\vec{X}_1) - G_W(\vec{X}_2)\|_2$$

$$\mathcal{L}(W) = \sum_{i=1}^{P} L(W, (Y, \vec{X}_1, \vec{X}_2)^i)$$

$$L(W, (Y, \vec{X}_1, \vec{X}_2)^i) = (1 - Y)L_S\left(D_W^i\right) + Y L_D\left(D_W^i\right)$$

Step 1: For each input sample $\vec{X}_i$, do the following:

(a) Using prior knowledge find the set of samples $\mathcal{S}_{\vec{X}_i} = \{\vec{X}_j\}_{j=1}^{p}$, such that $\vec{X}_j$ is deemed similar to $\vec{X}_i$.

(b) Pair the sample $\vec{X}_i$ with all the other training samples and label the pairs so that: $Y_{ij} = 0$ if $\vec{X}_j \in \mathcal{S}_{\vec{X}_i}$, and $Y_{ij} = 1$ otherwise.

Combine all the pairs to form the labeled training set.

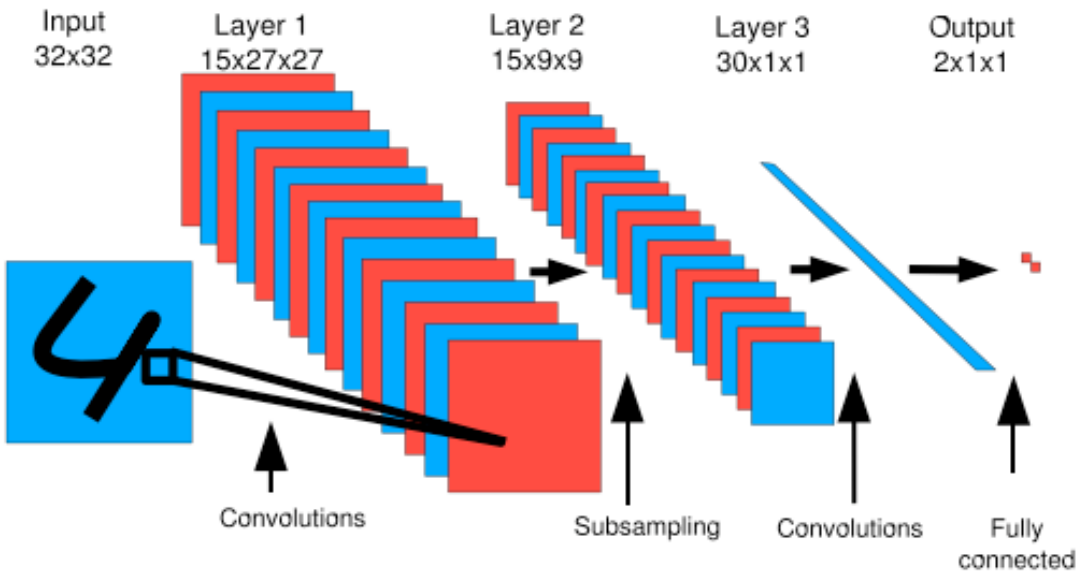
Step 2: Repeat until convergence:

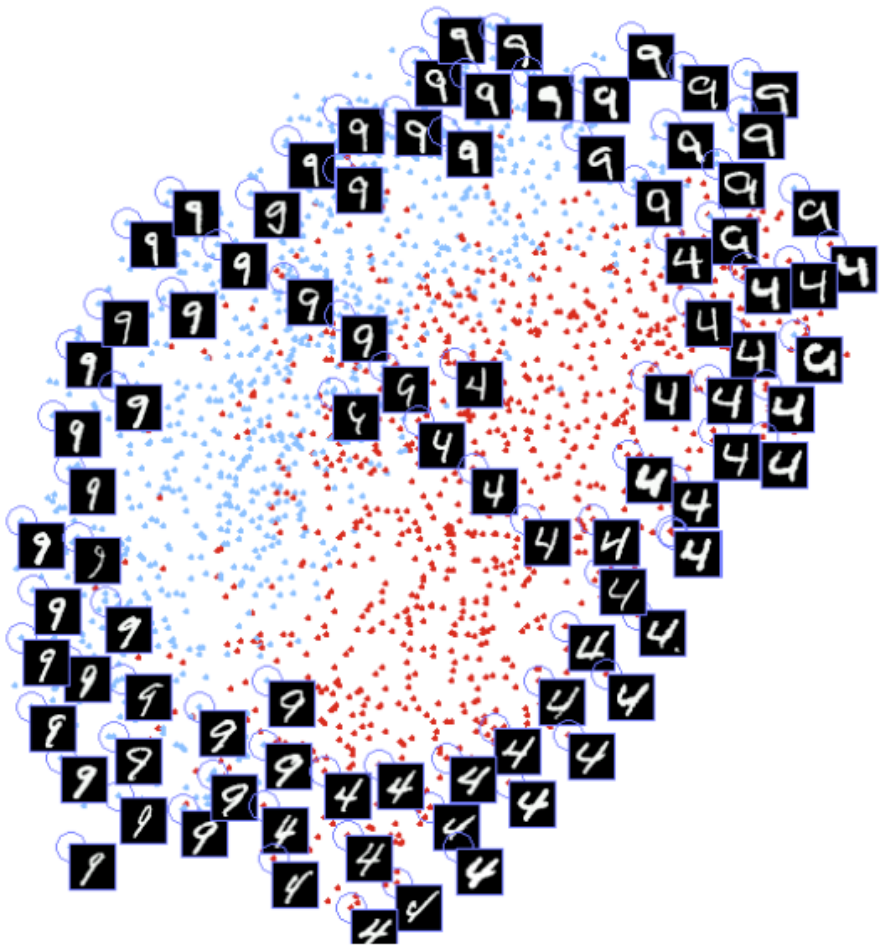
(a) For each pair $(\vec{X}_i, \vec{X}_j)$ in the training set, do

i. If $Y_{ij} = 0$, then update W to decrease $D_W = \|G_W(\vec{X}_i) - G_W(\vec{X}_j)\|_2$

ii. If $Y_{ij} = 1$, then update W to increase $D_W = \|G_W(\vec{X}_i) - G_W(\vec{X}_j)\|_2$

Contrast Loss (2006)

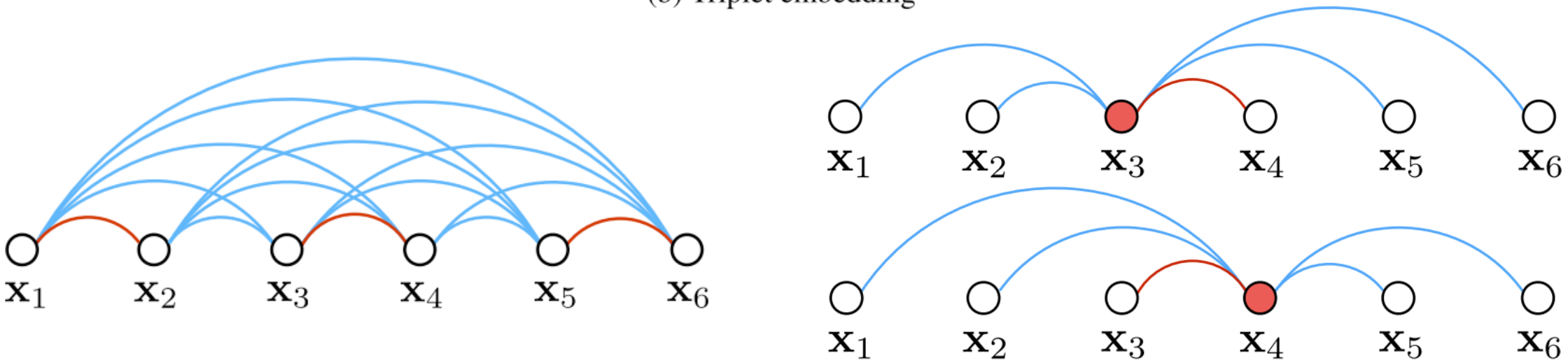




Contrast Loss with Supervision (Labels)

(a) Contrastive embedding

(b) Triplet embedding



(c) Lifted structured embedding

Contrast Loss with Supervision (Labels)

- Contrast loss
  - LeCun et al, 2006

$$J = \frac{1}{m} \sum_{(i,j)}^{m/2} y_{i,j} D_{i,j}^2 + (1 - y_{i,j}) \left[\alpha - D_{i,j}\right]_+^2 ,$$

- Triplet loss
  - 2006

$$J = \frac{3}{2m} \sum_{i}^{m/3} \left[D_{ia,ip}^2 - D_{ia,in}^2 + \alpha\right]_+$$

- Lifted Structured Embedding
  - Song et al, 2016

$$\tilde{J}_{i,j} = \log \left(\sum_{(i,k) \in \mathcal{N}} \exp\{\alpha - D_{i,k}\} + \sum_{(j,l) \in \mathcal{N}} \exp\{\alpha - D_{j,l}\} \right) + D_{i,j}$$

$$\tilde{J} = \frac{1}{2|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \max\left(0,\ \tilde{J}_{i,j}\right)^2 , \qquad (4)$$

Self-supervised Learning with Contrastive Loss

SimCLR을 중심으로

Data Augmentation

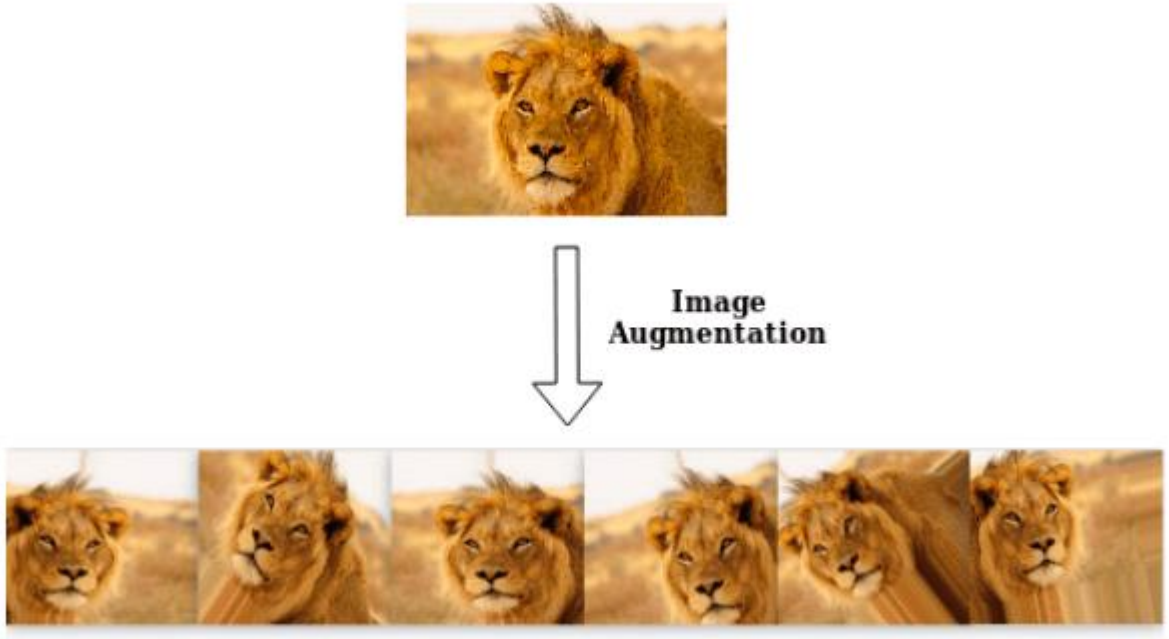
Data Labeling

Data Augmentation

MixUp
CutMix
Fast AA

Data Labeling

Data augmentation



Momentum Contrast for Unsupervised Visual Representation Learning

Kaiming He Haoqi Fan Yuxin Wu Saining Xie Ross Girshick

Facebook AI Research (FAIR)

A Simple Framework for Contrastive Learning of Visual Representations

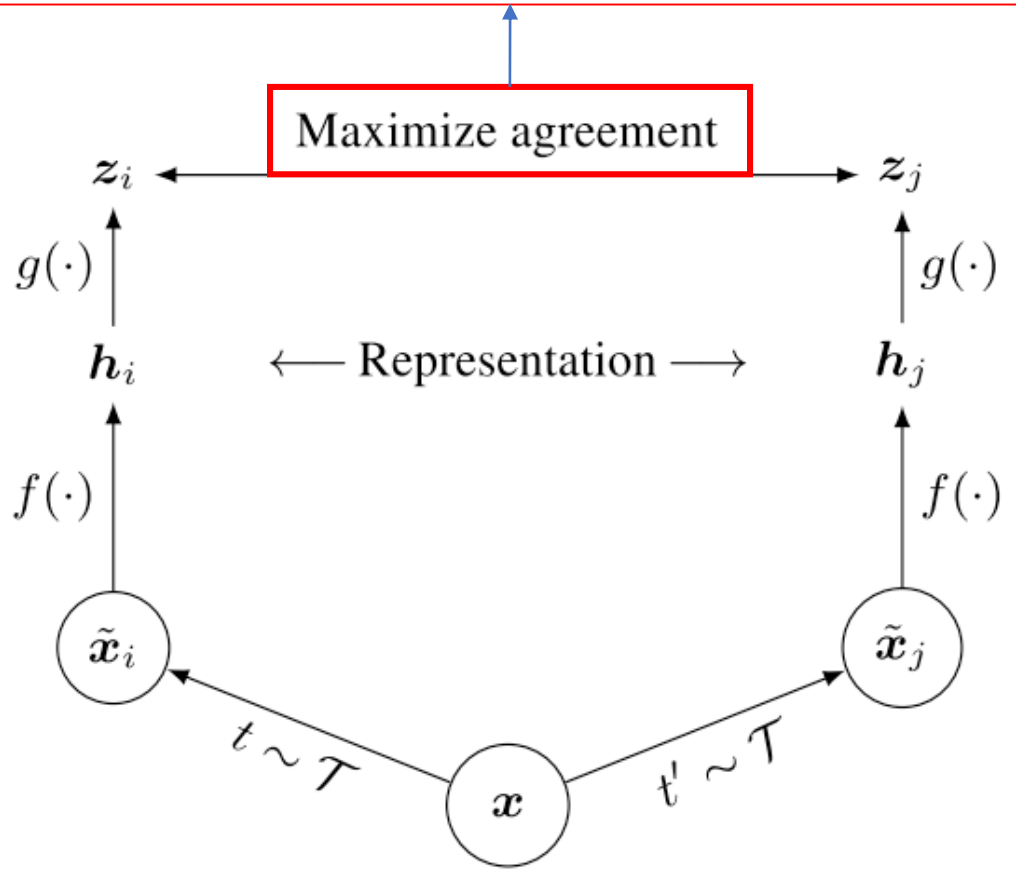
---

Ting Chen [1] **Simon Kornblith** [1] **Mohammad Norouzi** [1] **Geoffrey Hinton** [1]

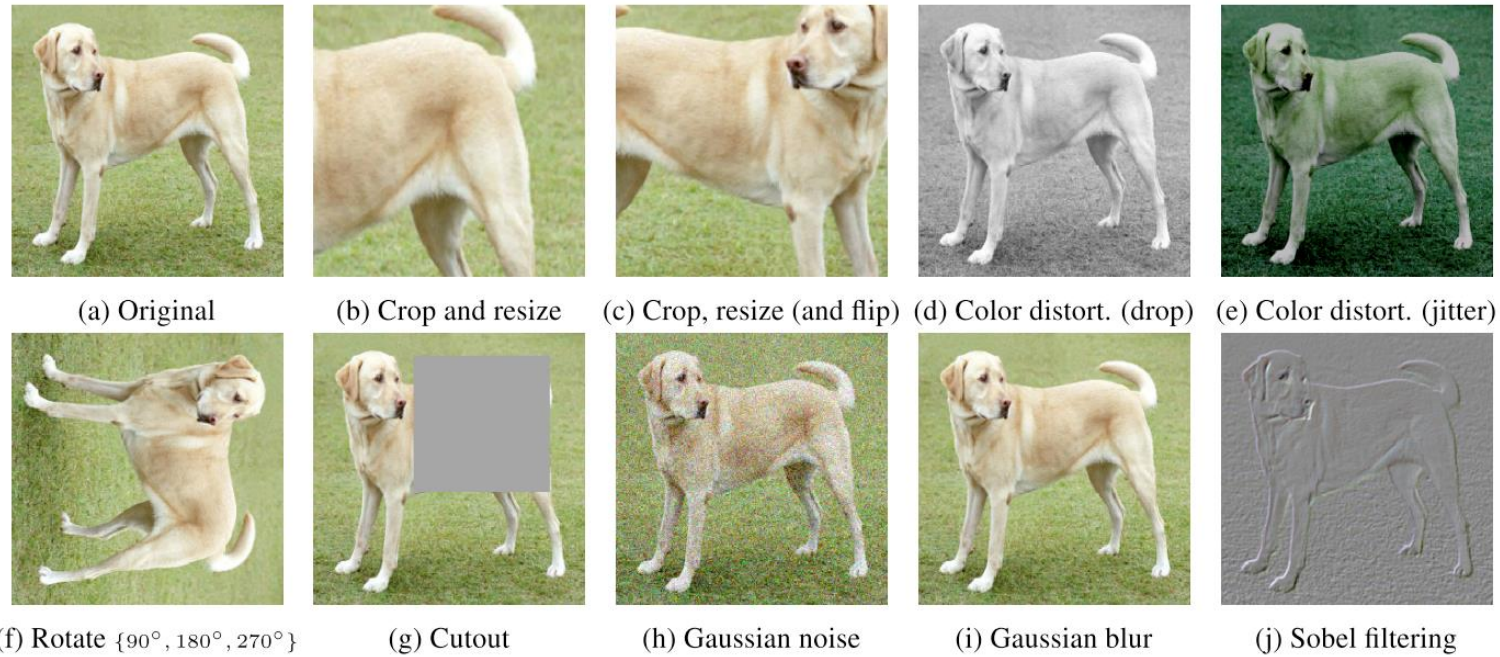
SimCLR

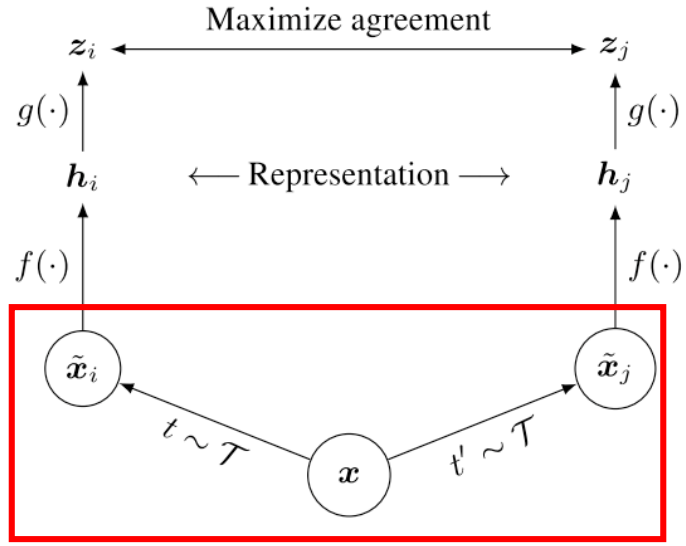
- 매우 직관적인 framework
- Contrastive Loss를 활용한 Unsuvervised Learning이 promising 함
- 다음의 요소가 중요함
  - Multiview 를 만들 때
    - Random Crop + Resize
    - Color distortion
    - Guassian Blur
  - h와 z의 분리!

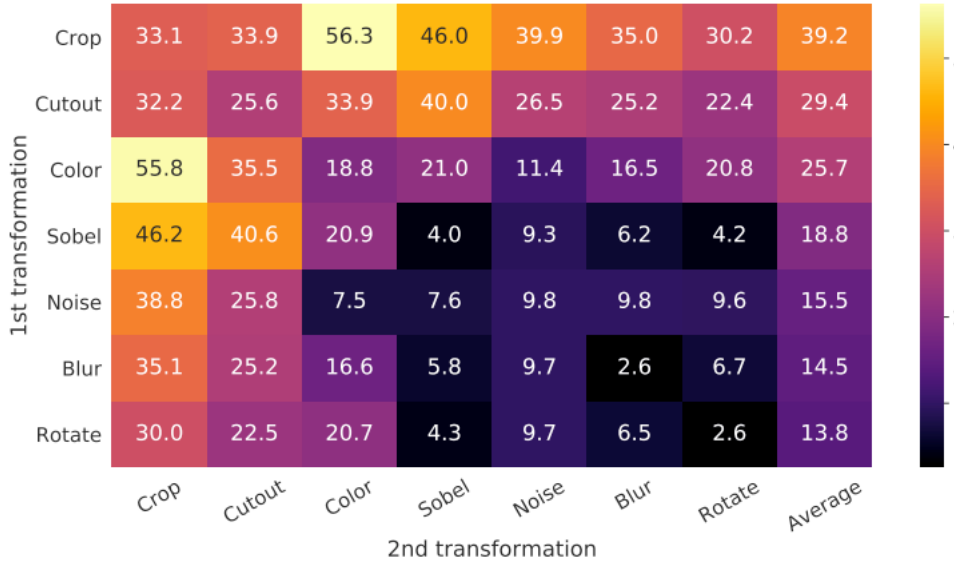
$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$



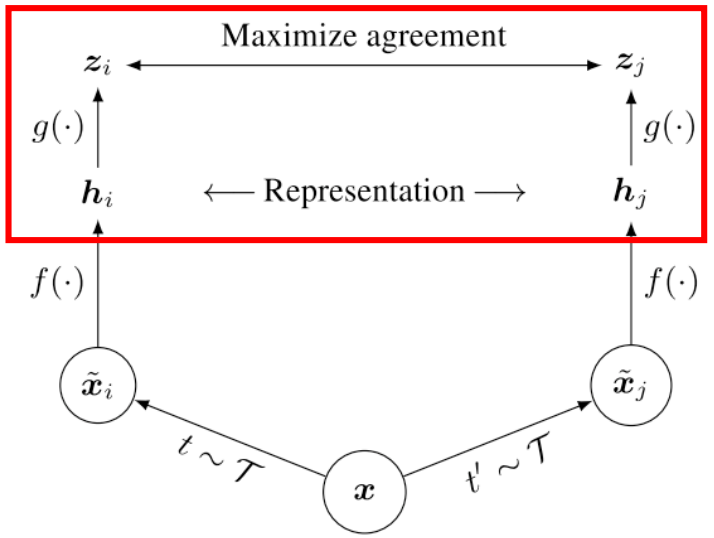
SimCLR

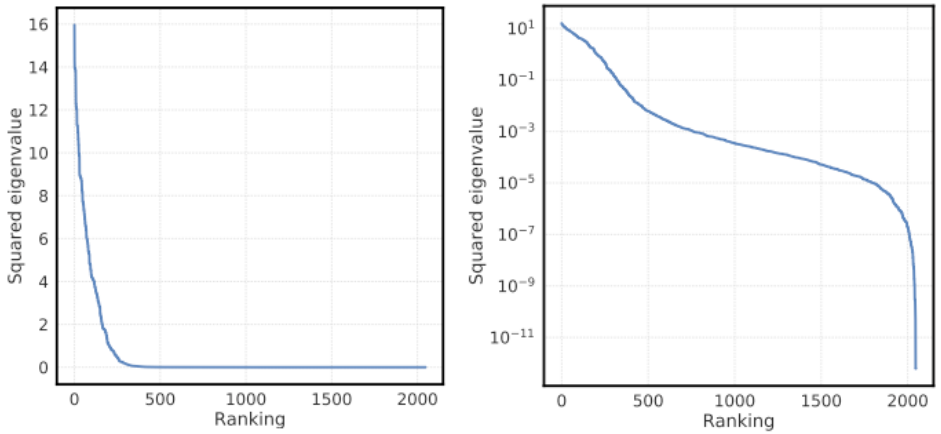

(a) Original
(b) Crop and resize
(c) Crop, resize (and flip)
(d) Color distort. (drop)
(e) Color distort. (jitter)
(f) Rotate $\{90°, 180°, 270°\}$
(g) Cutout
(h) Gaussian noise
(i) Gaussian blur
(j) Sobel filtering


Maximize agreement
z_i
z_j
$g(\cdot)$
$g(\cdot)$
h_i
Representation
h_j
$f(\cdot)$
$f(\cdot)$
$\tilde{x}_i$
$\tilde{x}_j$
$t \sim \mathcal{T}$
$t' \sim \mathcal{T}$
x


Crop 33.1 33.9 56.3 46.0 39.9 35.0 30.2 39.2
Cutout 32.2 25.6 33.9 40.0 26.5 25.2 22.4 29.4
Color 55.8 35.5 18.8 21.0 11.4 16.5 20.8 25.7
Sobel 46.2 40.6 20.9 4.0 9.3 6.2 4.2 18.8
Noise 38.8 25.8 7.5 7.6 9.8 9.8 9.6 15.5
Blur 35.1 25.2 16.6 5.8 9.7 2.6 6.7 14.5
Rotate 30.0 22.5 20.7 4.3 9.7 6.5 2.6 13.8
1st transformation
Crop
Cutout
Color
Sobel
Noise
Blur
Rotate
Average
2nd transformation
50
40
30
20
10

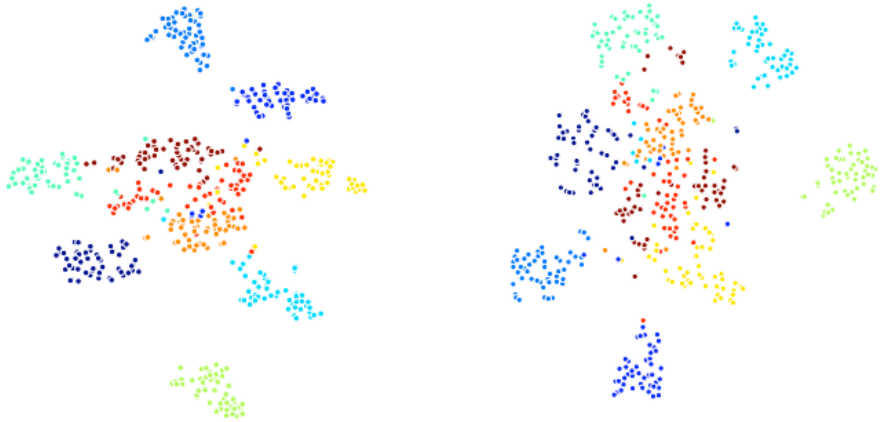
SimCLR





(a) Y-axis in uniform scale. (b) Y-axis in log scale.

Figure B.1. Squared real eigenvalue distribution of linear projection matrix $W \in R^{2048\times2048}$ used to compute $g(\boldsymbol{h}) = W\boldsymbol{h}$.

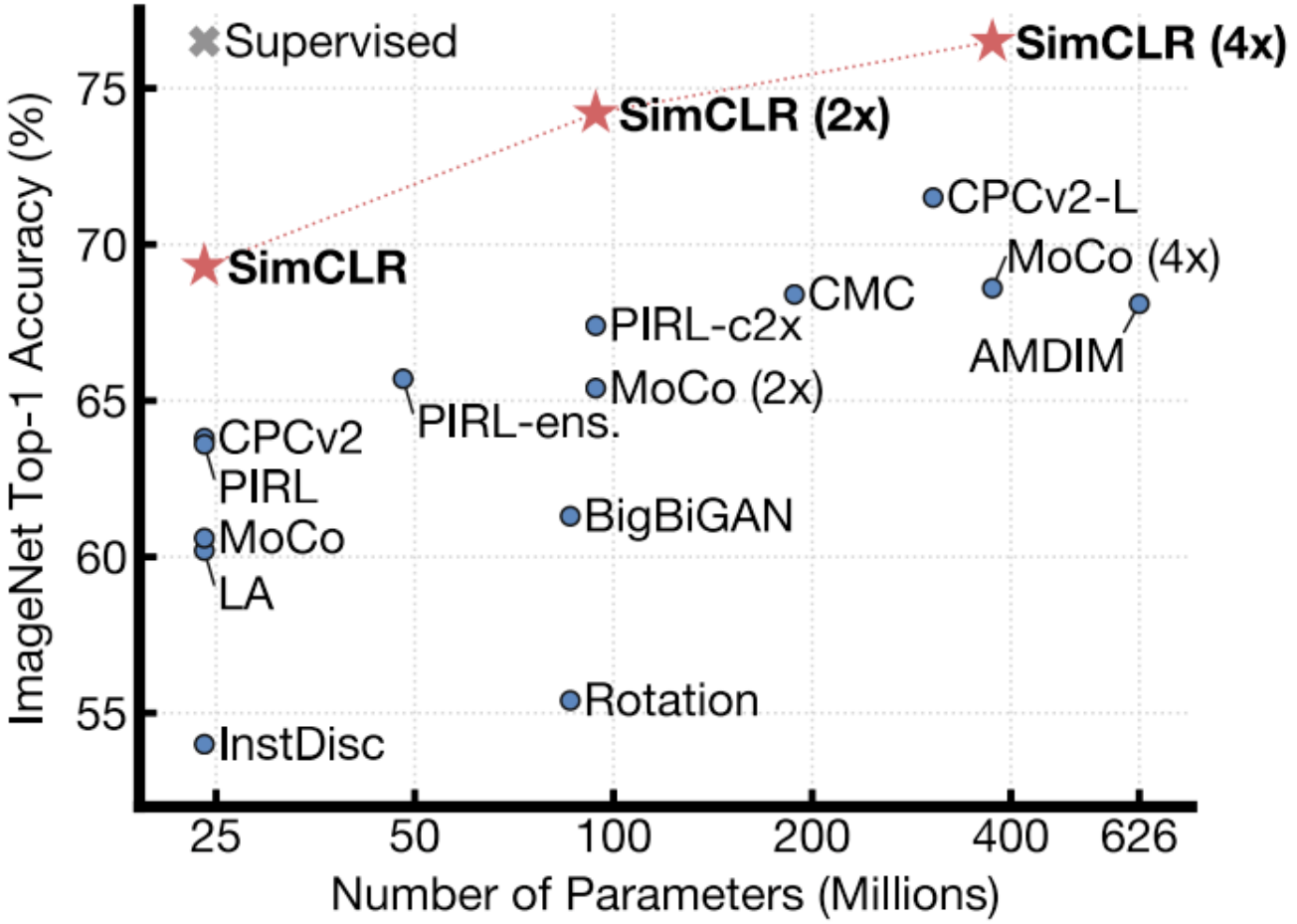


(a) $\boldsymbol{h}$ (b) $\boldsymbol{z} = g(\boldsymbol{h})$

Figure B.2. t-SNE visualizations of hidden vectors of images from a randomly selected 10 classes in the validation set.

SimCLR

- Image Classification
  - ImageNet dataset



Applications

CLIP

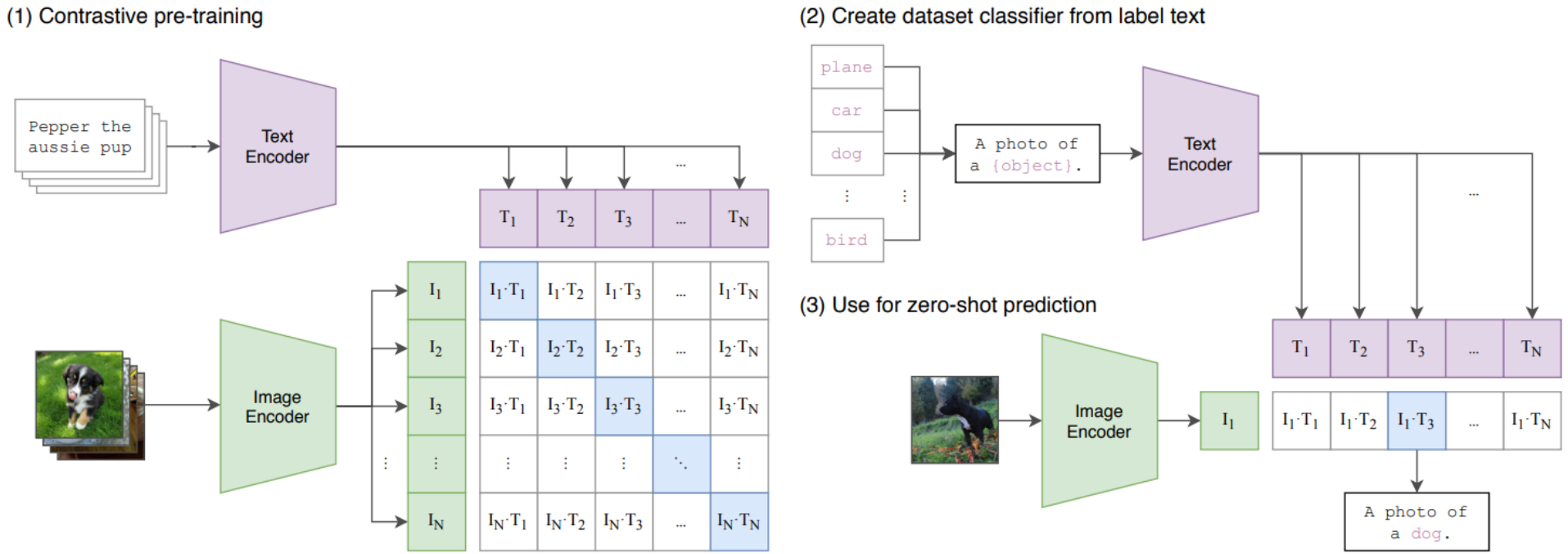
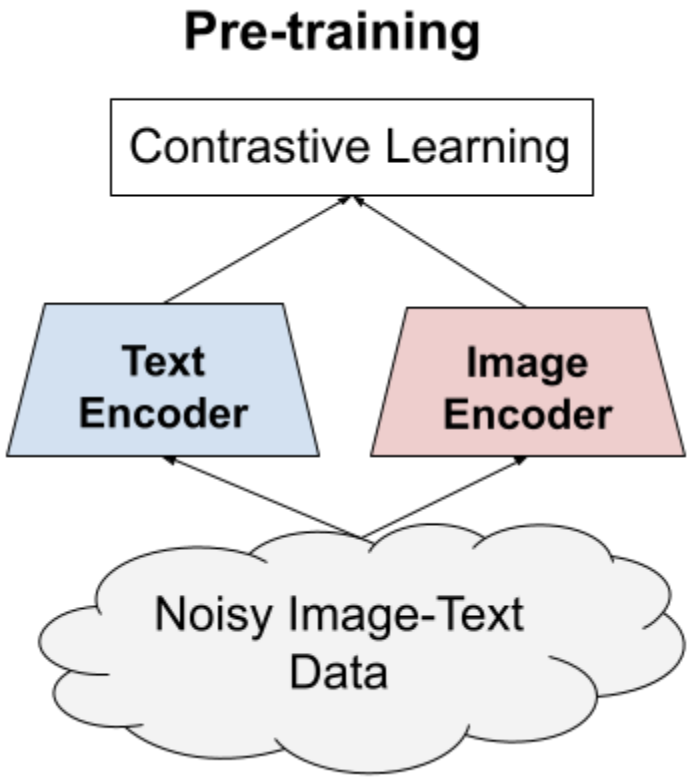


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

ALIGN: A Large-scale ImaGe and Noisy-Text Embedding



감사합니다!

eunsolkim@hanyang.ac.kr