자바지로학습 + 언어 모델을 통한 대규모 사전학습 모델들

Self-Supervised Learning and Large-Scale Pre-Trained Models

주재걸 교수

KAIST 김재철AI대학원

DAVIAN

Data and Visual Analytics Lab

What is Self-Supervised Learning? 자가지도학습?

- Given unlabeled data, hide part of the data and train the model so that it can predict such a hidden part of data, given the remaining data.
- Image in-painting example:

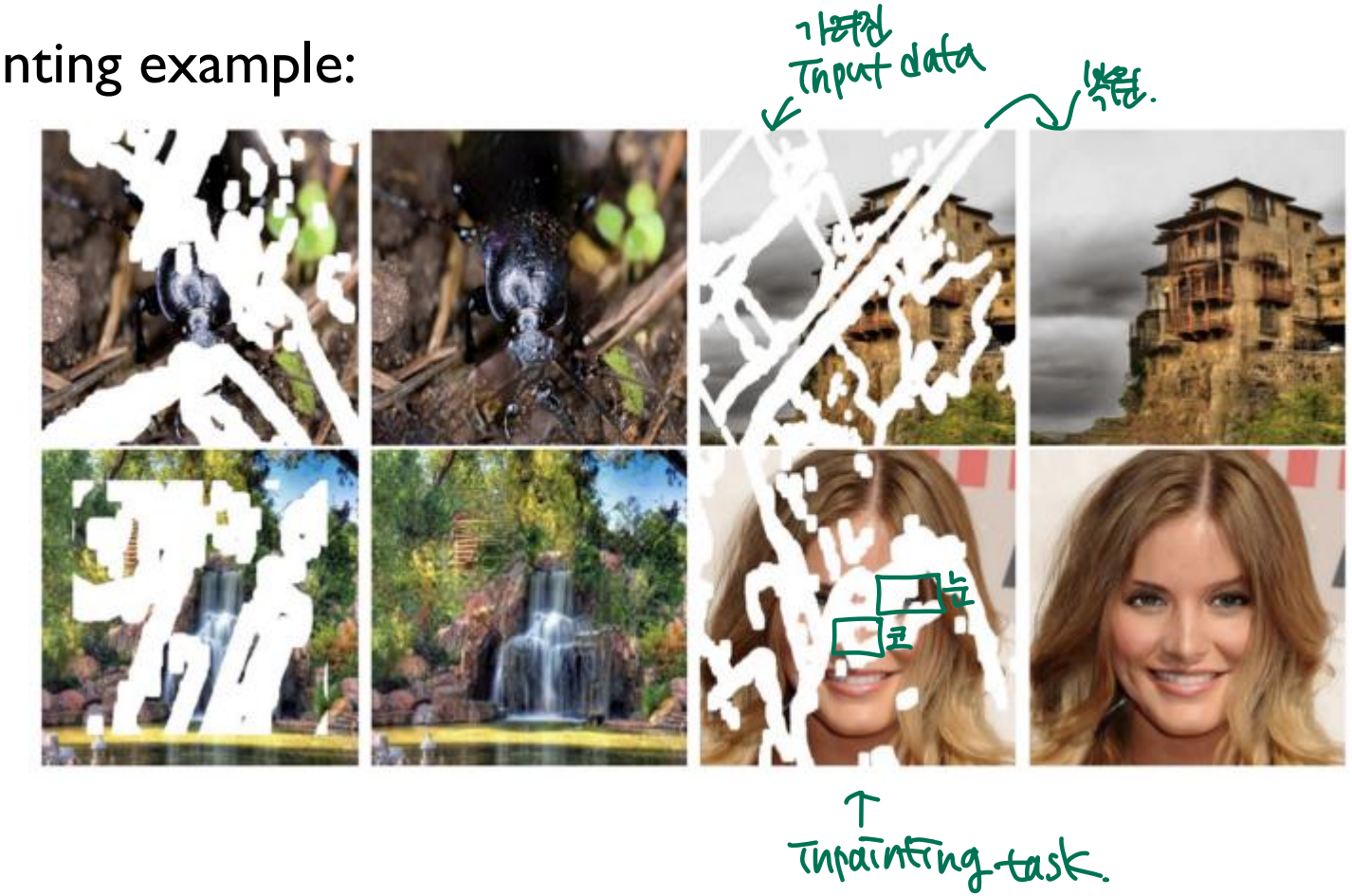


Image Courtesy: https://towardsdatascience.com/pushing-the-limits-of-deep-image-inpainting-using-partial-convolutions-ed5520775ab4

What is Self-Supervised Learning?

- Given unlabeled data, hide part of the data and train the model so that it can predict such a hidden part of data, given the remaining data.
- Solving a zigsaw puzzle:

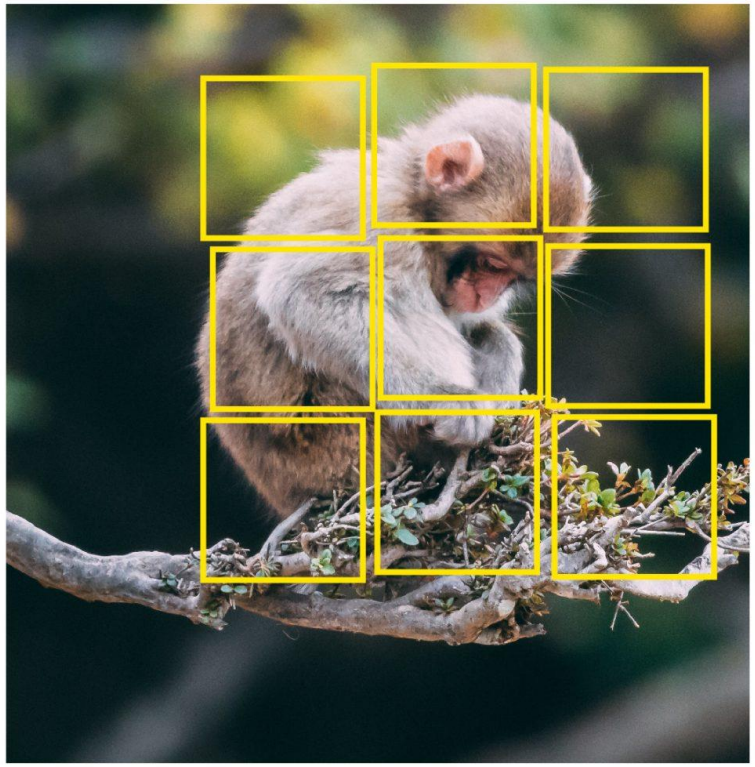


Image Courtesy: https://perfectial.com/blog/self-supervised-learning/

Transfer Learning from Self-Supervised Pre-trained Model

- Pre-trained models using a particular self-supervised learning can be fine-tuned to improve the accuracy of a given target task.

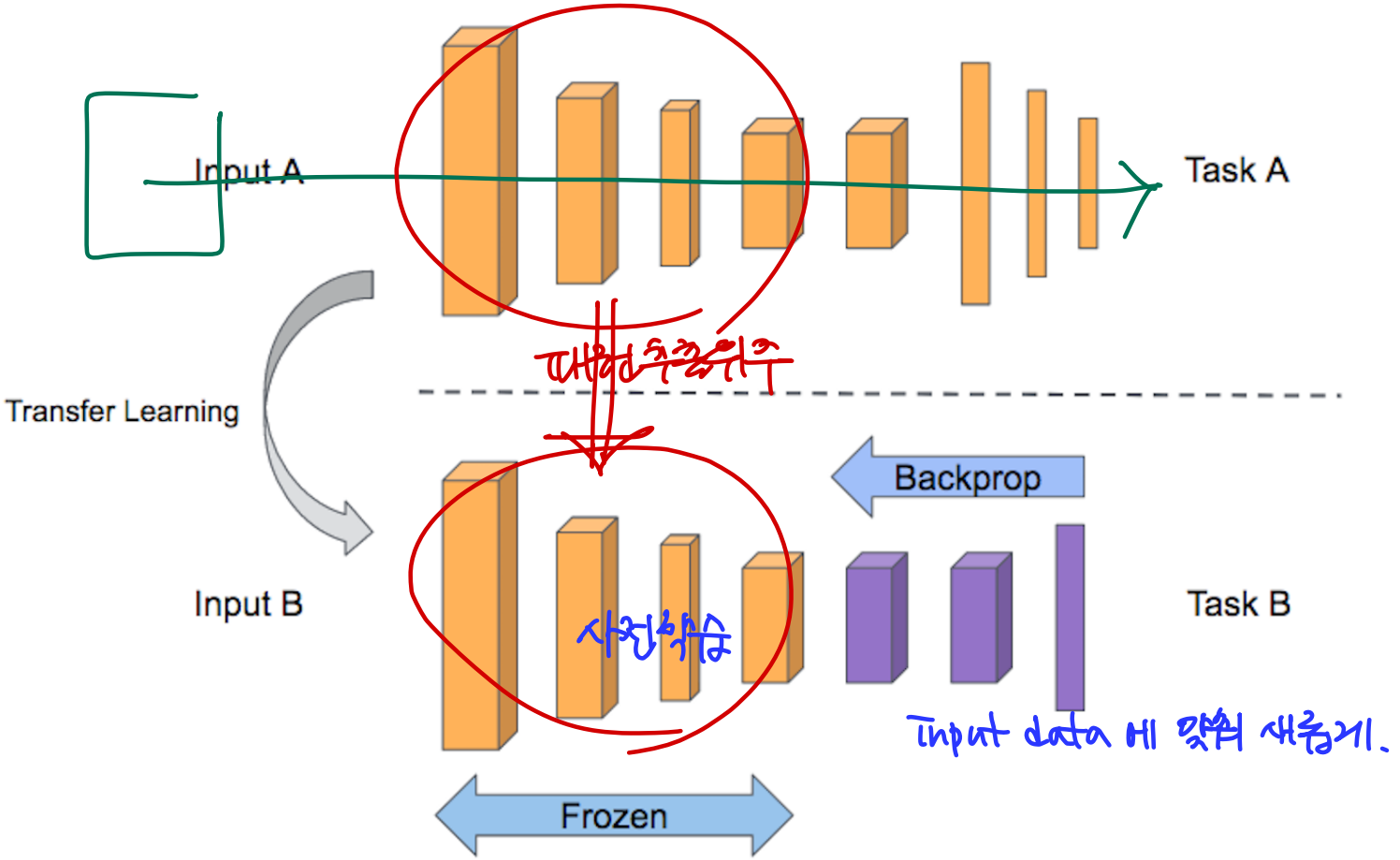


Image Courtesy: https://medium.com/@subodh.malgonde/transfer-learning-using-tensorflow-52a4f6bcde3e

BERT

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
  - Learn through masked language modeling (MLM) and next-sentence prediction (NSP) tasks
  - Use large-scale data and large-scale model

Overview of BERT

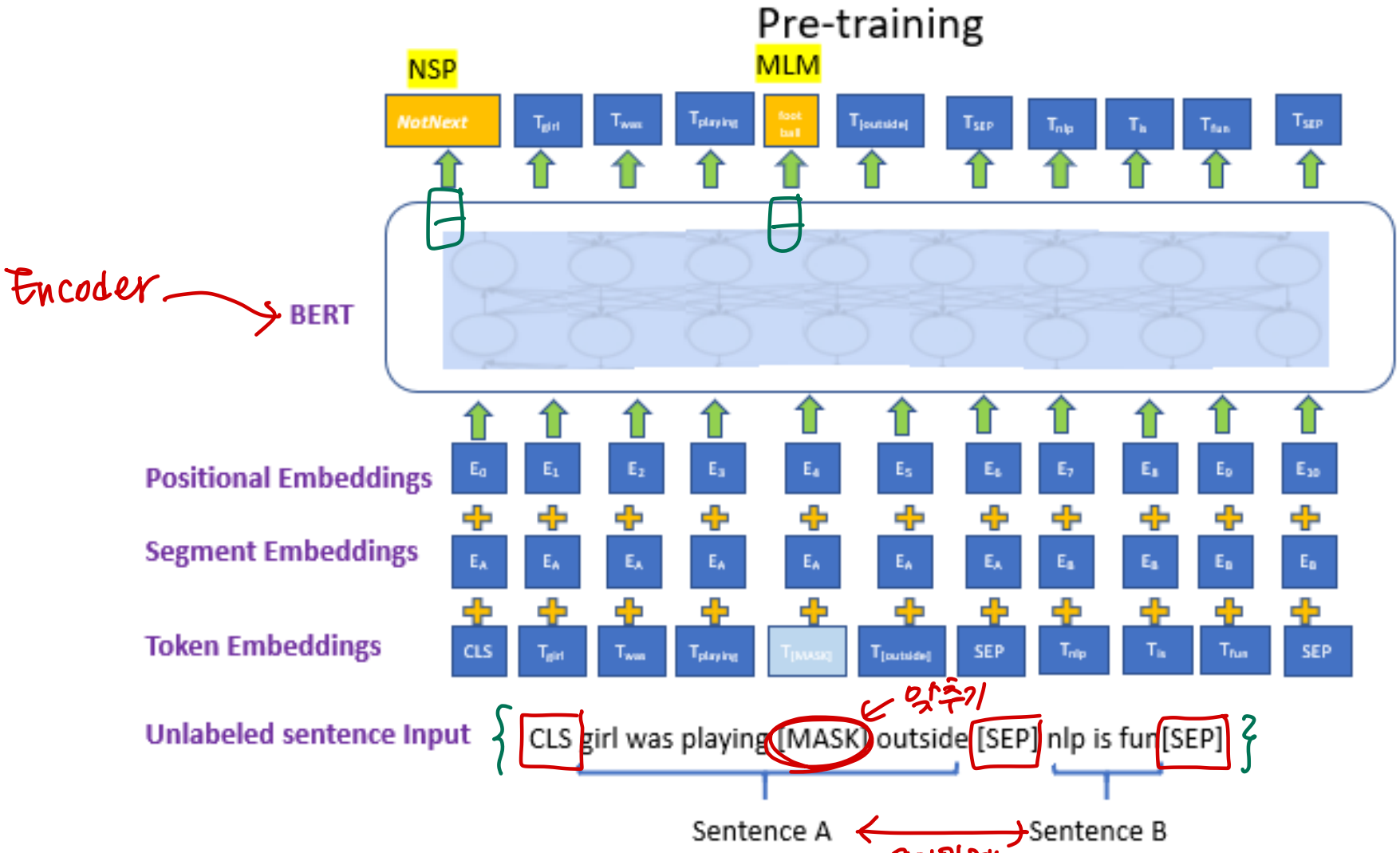
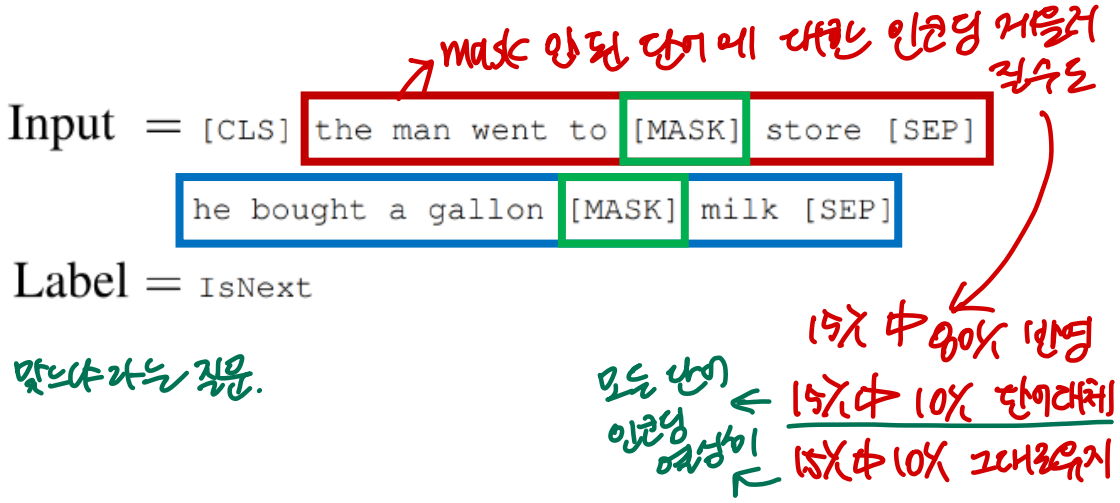


Image courtesy: https://towardsdatascience.com/intuitive-explanation-of-bert-bidirectional-transformers-for-nlp-cdc1efd696c1e

Pre-Training Tasks of BERT

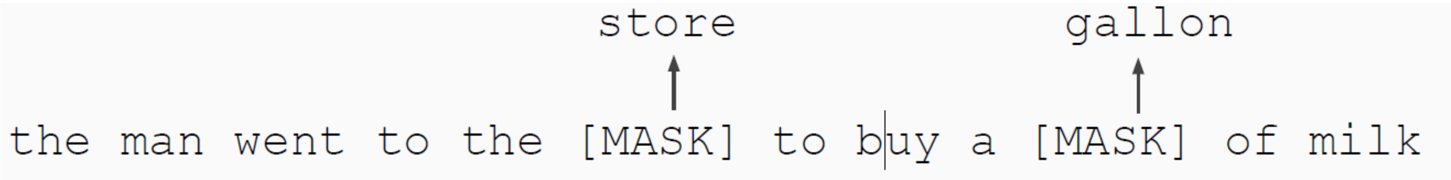
- Masked Language Model (MLM)
  - Mask some percentage of the input tokens at random, and then predict those masked tokens.
  - 15% of the words to predict
    - 80% of the time, replace with [MASK]
    - 10% of the time, replace with a random word
    - 10% of the time, keep the sentence as same



- Next Sentence Prediction (NSP)
  - Predict whether Sentence B is an actual sentence that proceeds Sentence A, or a random sentence

Masked Language Modeling (MLM) Task

- Mask out *k*% of the input words, and then predict the masked words.
- BERT uses *k*= 15%
  - Too little masking : Too expensive to train.
  - Too much masking : Not enough to capture the given context. → 빠르지만 예측어렵.

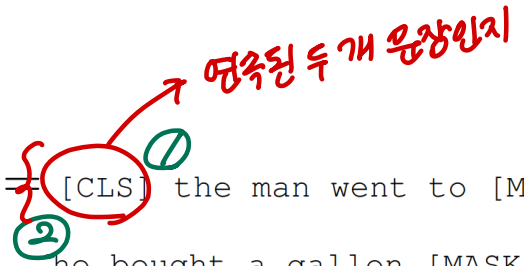


- Out of *k*= 15% selected tokens,
  - For 80% of the time, replace them with [MASK] token
    - went to the store → went to the [MASK]
  - For 10% of the time, replace them with a random word
    - e.g., went to the store → went to the running
  - For 10% of the time, keep them as it is,
    - e.g., went to the store → went to the store

https://nlp.stanford.edu/seminar/details/jdevlin.pdf

Next Sentence Prediction (NSP) Task

- To learn the relationships among sentences, predict whether Sentence B is an actual sentence that follows Sentence A, or a random, irrelevant sentence



Input = [CLS] the man went to [MASK] store [SEP]
he bought a gallon [MASK] milk [SEP]

Label = IsNext

Input = [CLS] the man [MASK] to the store [SEP]
penguin [MASK] are flight ##less birds [SEP]

Label = NotNext

https://nlp.stanford.edu/seminar/details/jdevlin.pdf

Further Details of BERT

1. Model Architecture
   - BERT BASE: L = 12, H = 768, A = 12
   - BERT LARGE: L = 24, H = 1024, A = 16
2. Input Representation
   - WordPiece embeddings (30,000 WordPiece)
   - Learned positional embedding
   - [CLS] – Classification embedding
   - Packed sentence embedding [SEP]
   - Segment Embedding
3. Pre-training Tasks
   - Masked LM
   - Next Sentence Prediction

Segment Embedding and Position Embedding

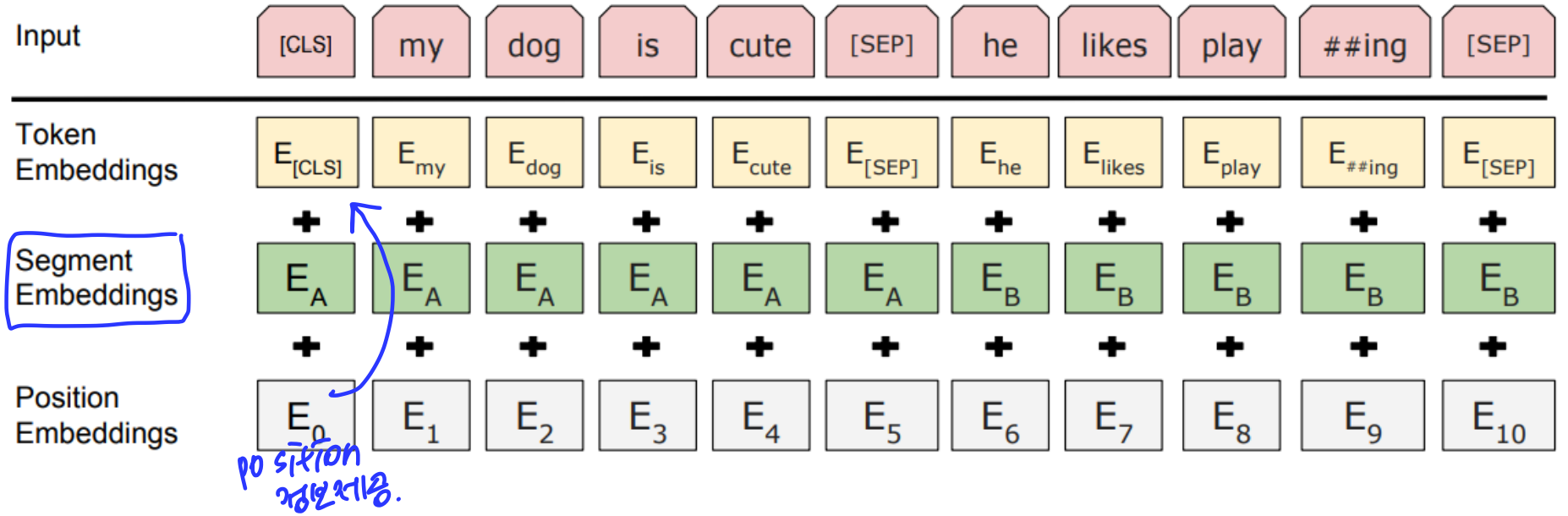
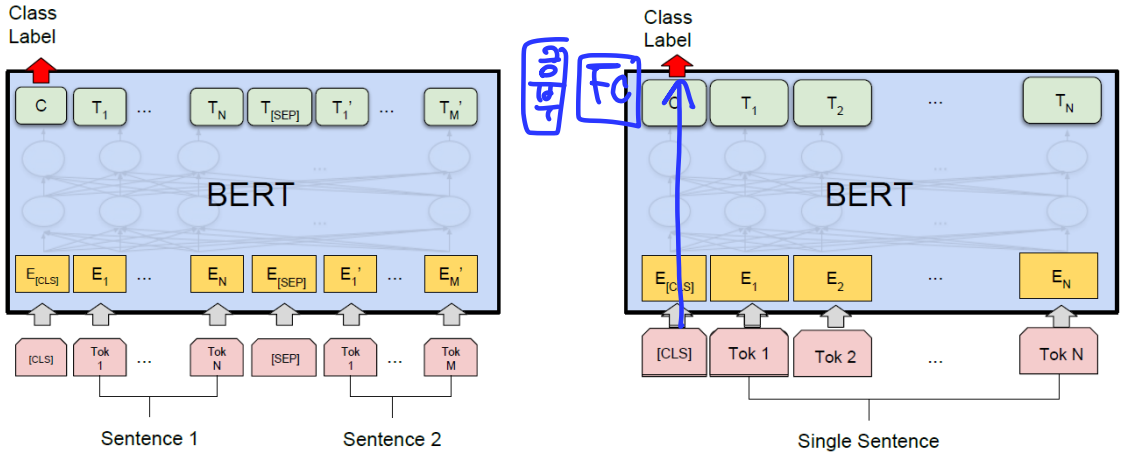


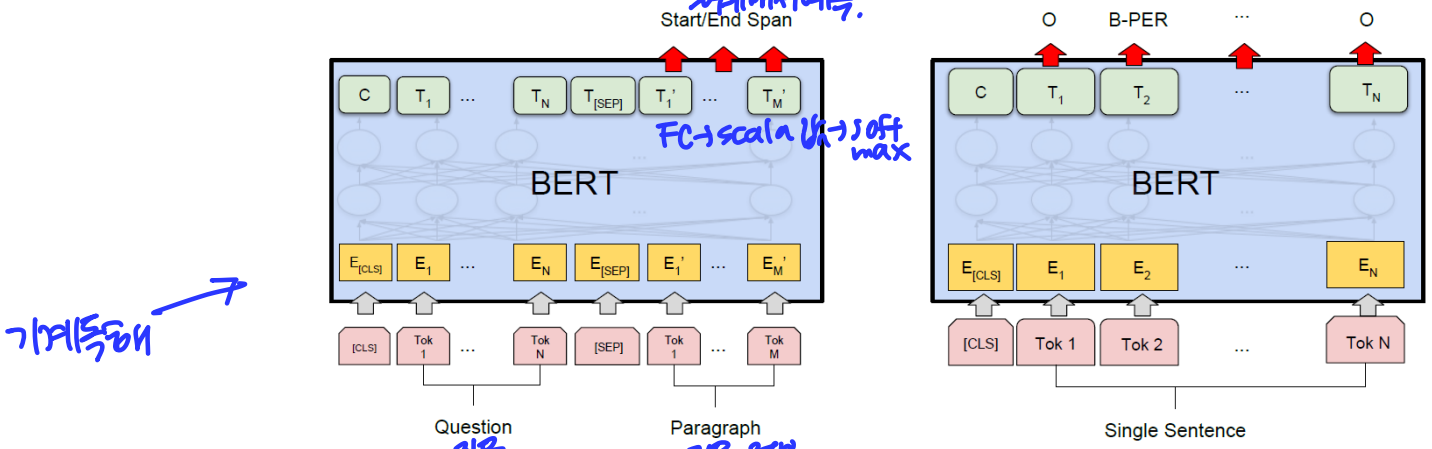
Figure 2: BERT input representation. The input embeddings is the sum of the token embeddings, the segmentation embeddings and the position embeddings.

Various Fine-Tuning Approaches



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

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



(c) Question Answering Tasks: SQuAD v1.1

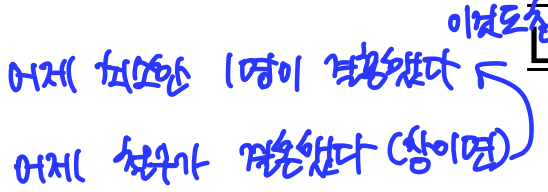
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

BERT: GLUE Benchmark Results

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

- MultiNLI
  - Premise: Hills and mountains are especially sanctified in Jainism.
  - Hypothesis: Jainism hates nature.
  - Label: Contradiction

- CoLa
  - Sentence: The wagon rumbled sown the road.
    Label: Acceptable
  - Sentence: The car honked down the road.
    Label: Unacceptable



GPT-1/2/3: Generative Pre-Trained Transformer

- Generative Pre-Training Task
  - In other words, this task is called Language Modeling.
  - From another perspective, this task is called an Auto-Regressive Model, in a sense that the predicted output at the current time step will be used as an input at the next time step.

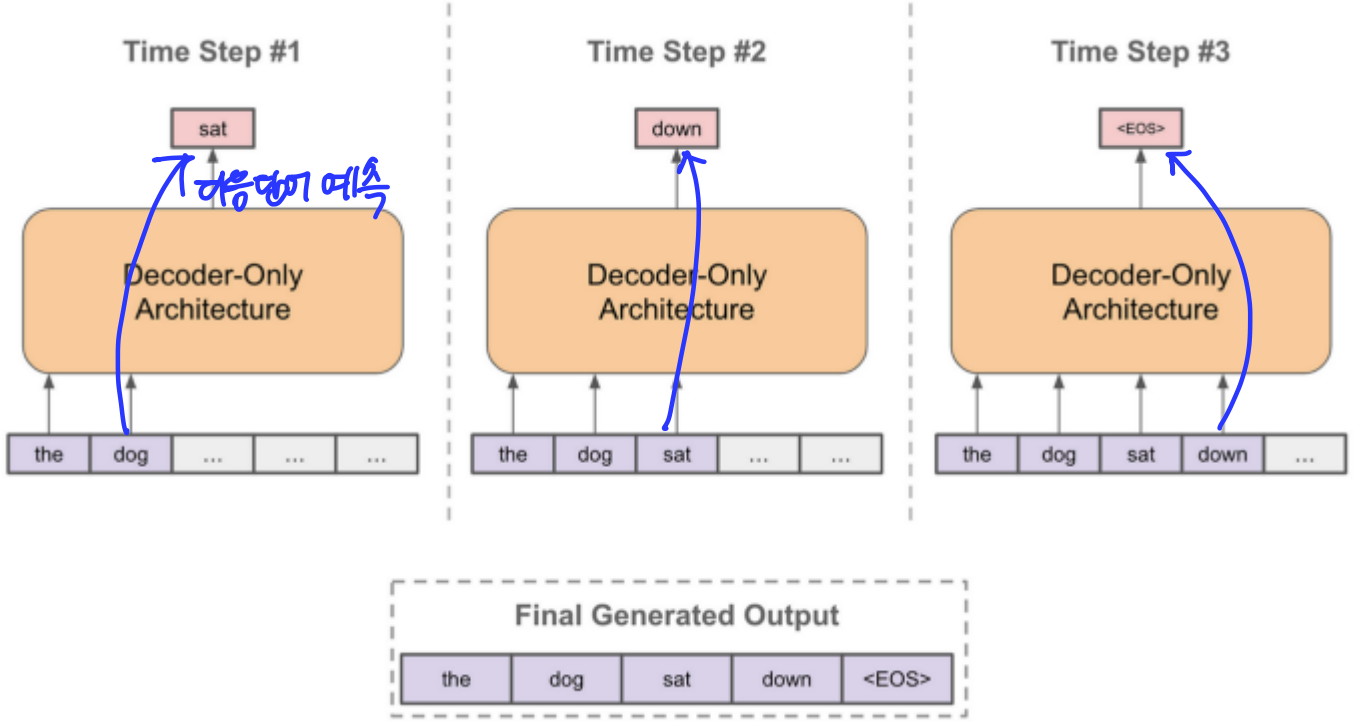
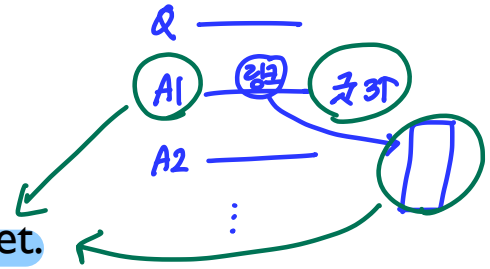


Image courtesy: https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html

GPT-2: Language Models are Unsupervised Multi-task Learners

- Just a really big transformer-based language model
  - Trained on 40GB of text
    - A large amount of efforts have been put to secure a high-quality dataset.
    - Take webpages from reddit links with at least 3 karma (up-vote)
  - Language models can perform down-stream tasks in a zero-shot setting – without any parameter or architecture modification



https://openai.com/blog/better-language-models/

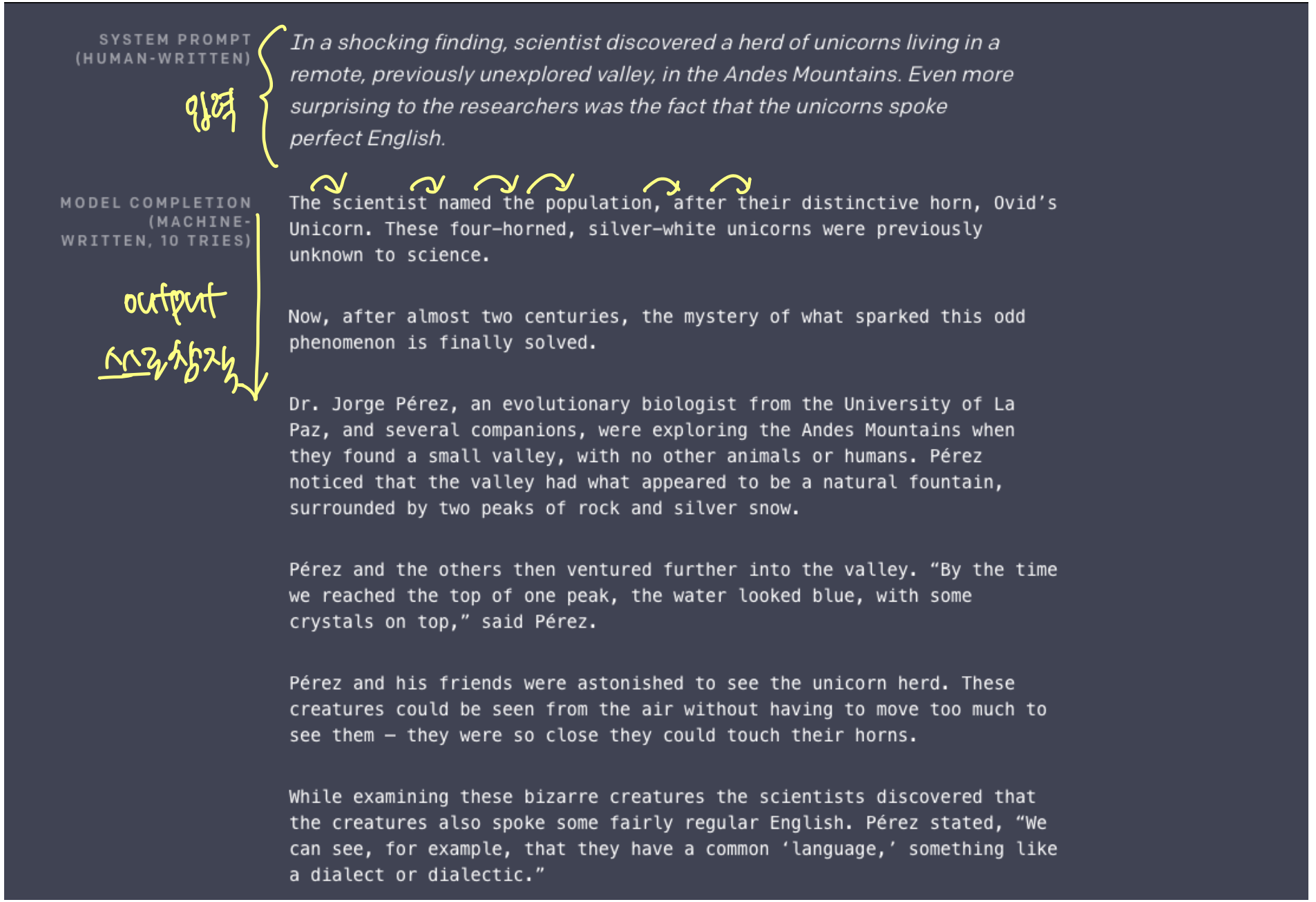


Image courtesy: https://blog.floydhub.com/GPT-2/

GPT-2: Zero-Shot Summarization

- CNN and Daily Main dataset
  - Add text TL;DR: after the article and generate 100 tokens (TL;DR: Too long, didn't read)

Language Models are Unsupervised Multitask Learners

Article: Prehistoric man sketched an incredible array of prehistoric beasts on the rough limestone walls of a cave in modern day France 36,000 years ago.
Now, with the help of cutting-edge technology, those works of art in the Chauvet-Pont-d'Arc Cave have been reproduced to create the biggest replica cave in the world.
The manmade cavern named the Caverne du Pont-d'Arc has been built a few miles from the original site in Vallon-Pont-D'arc in Southern France and contains 1,000 painstakingly-reproduced drawings as well as around 450 bones and other features...
Cavemen and women sketched an incredible array of prehistoric beasts on the rough limestone walls of a cave 36,000 years ago and now a replica has been created (pictured)
...

TL;DR ← 마지막 단어

GPT-2: The original site in Vallon-Pont-D'arc in Southern France is a Unesco World Heritage site and is the oldest known and the best preserved cave decorated by man. The replica cave was built a few miles from the original site in Vallon-Pont-D'Arc in Southern France. The cave contains images of 14 different species of animals including woolly rhinoceros, mammoths, and big cats.

Reference: Cave mimics famous Caverne du Pont-d'Arc in France, the oldest cave decorated by man and the best preserved. The replica contains all 1,000 paintings which include 425 such as a woolly rhinoceros and mammoths. Minute details were copied using 3D modelling and anamorphic techniques, often used to shoot widescreen images. The modern cave also includes replica paw prints of bears, bones and details preserved in the original cave.

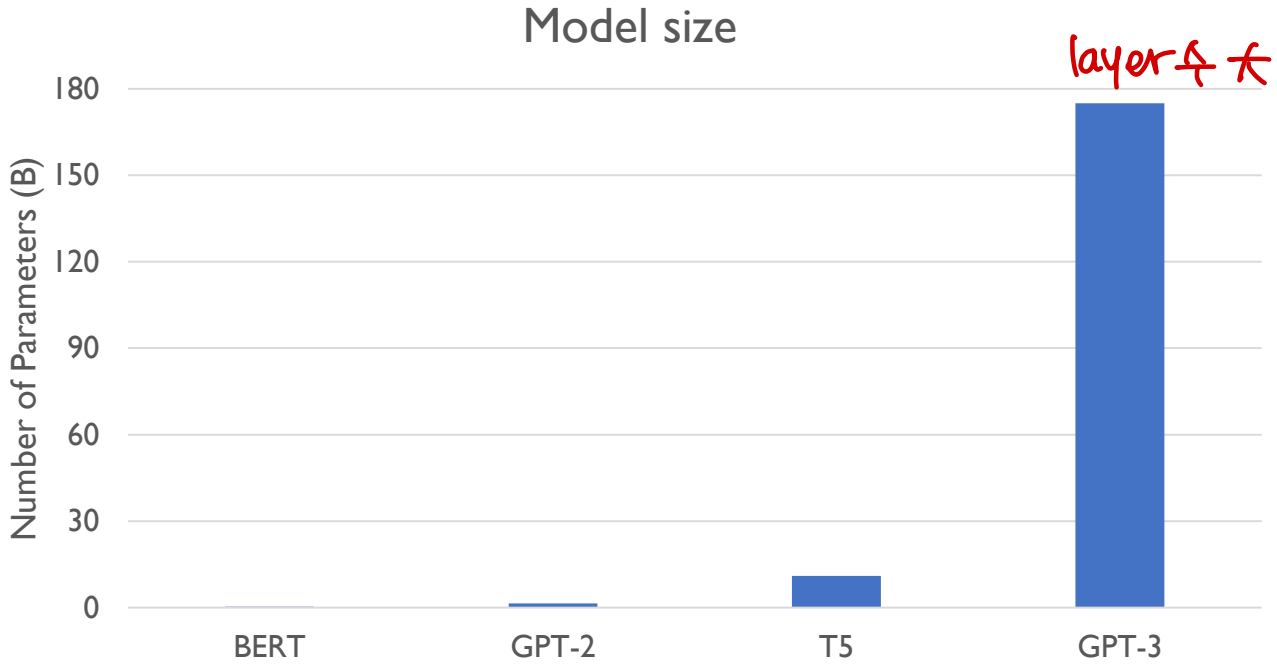
	R-1	R-2	R-L	R-AVG
Bottom-Up Sum	**41.22**	**18.68**	**38.34**	**32.75**
Lede-3	40.38	17.66	36.62	31.55
Seq2Seq + Attn	31.33	11.81	28.83	23.99
GPT-2 `TL;DR:`	29.34	8.27	26.58	21.40
Random-3	28.78	8.63	25.52	20.98
GPT-2 no hint	21.58	4.03	19.47	15.03

Table 4. Summarization performance as measured by ROUGE F1 metrics on the CNN and Daily Mail dataset. Bottom-Up Sum is the SOTA model from (Gehrmann et al., 2018)

https://openai.com/blog/better-language-models/

GPT-3: Language Models are Few-Shot Learners

- Scaling up language models greatly improves task-agnostic, few-shot performance
- An autoregressive language model with 175 billion parameters in the few-shot setting
- 96 attention layers, batch size of 3.2M, 175B parameters

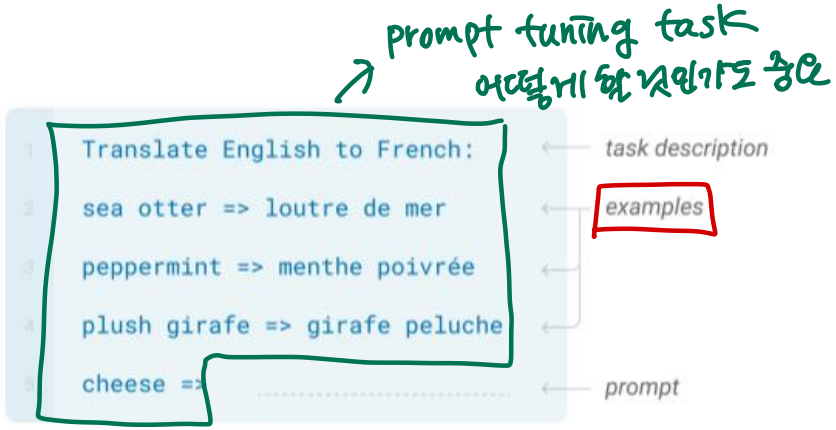


Few-Shot Learning Example of GPT-3

- **Prompt**: the prefix given to the model
- **Zero-shot:** Predict the answer given only a natural language description of the task
- **One-shot**: See a single example of the task in addition to the task description
- **Few-shot**: See a few examples of the task

Zero-shot

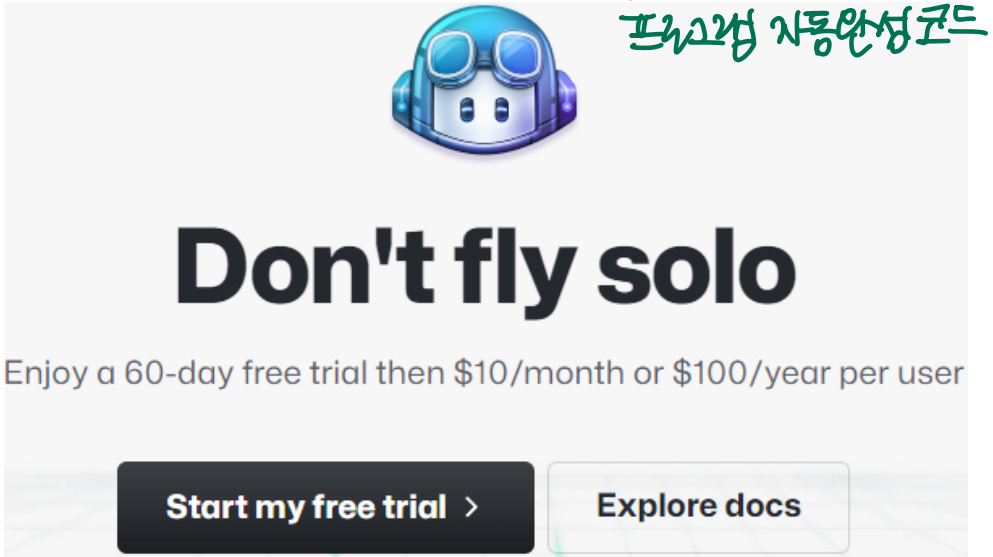
One-shot learning task

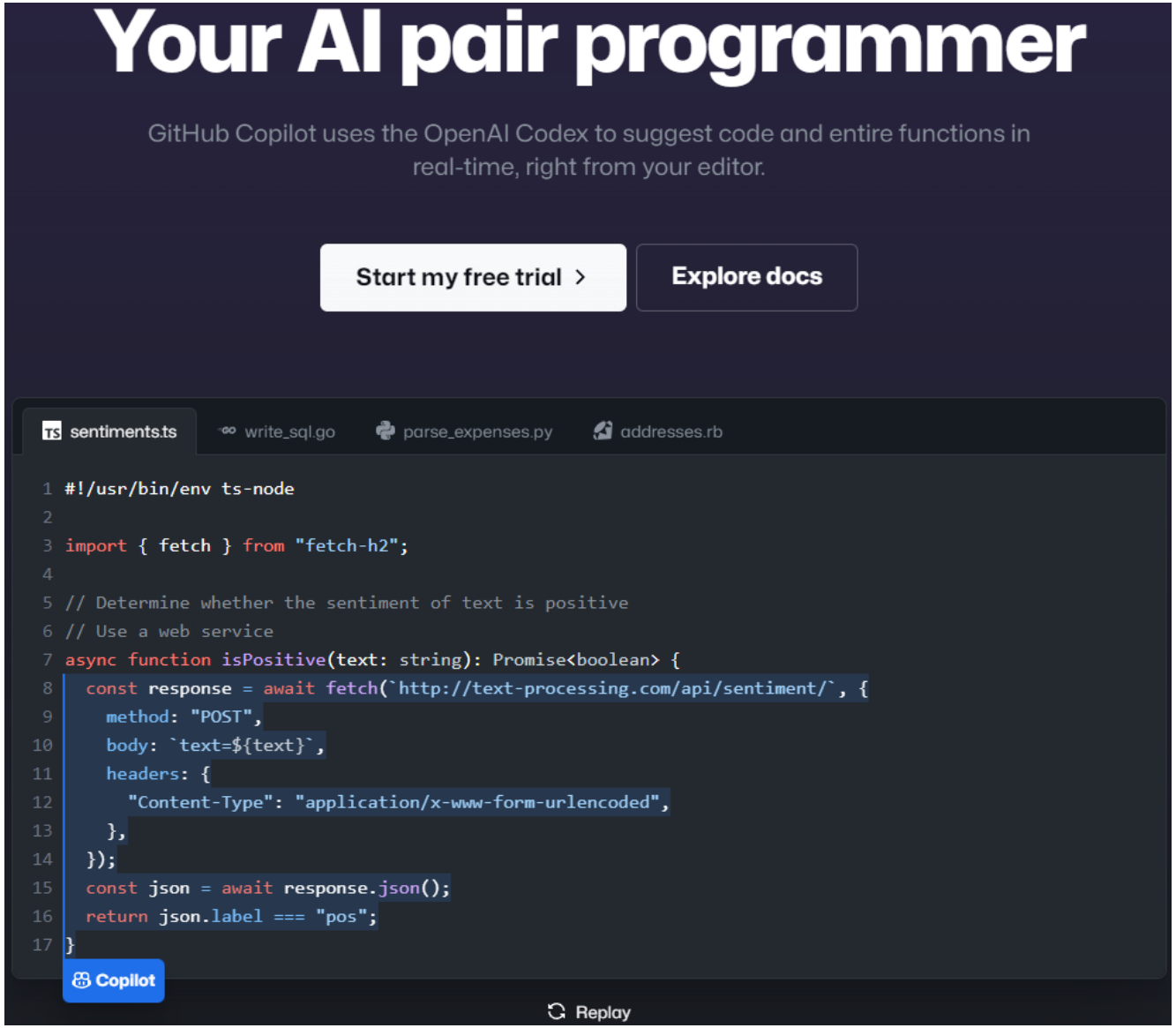


Few-shot ← 성능 좋음.

Copilot: Program Auto-Completion based on GPT-3

프로그램 자동완성코드



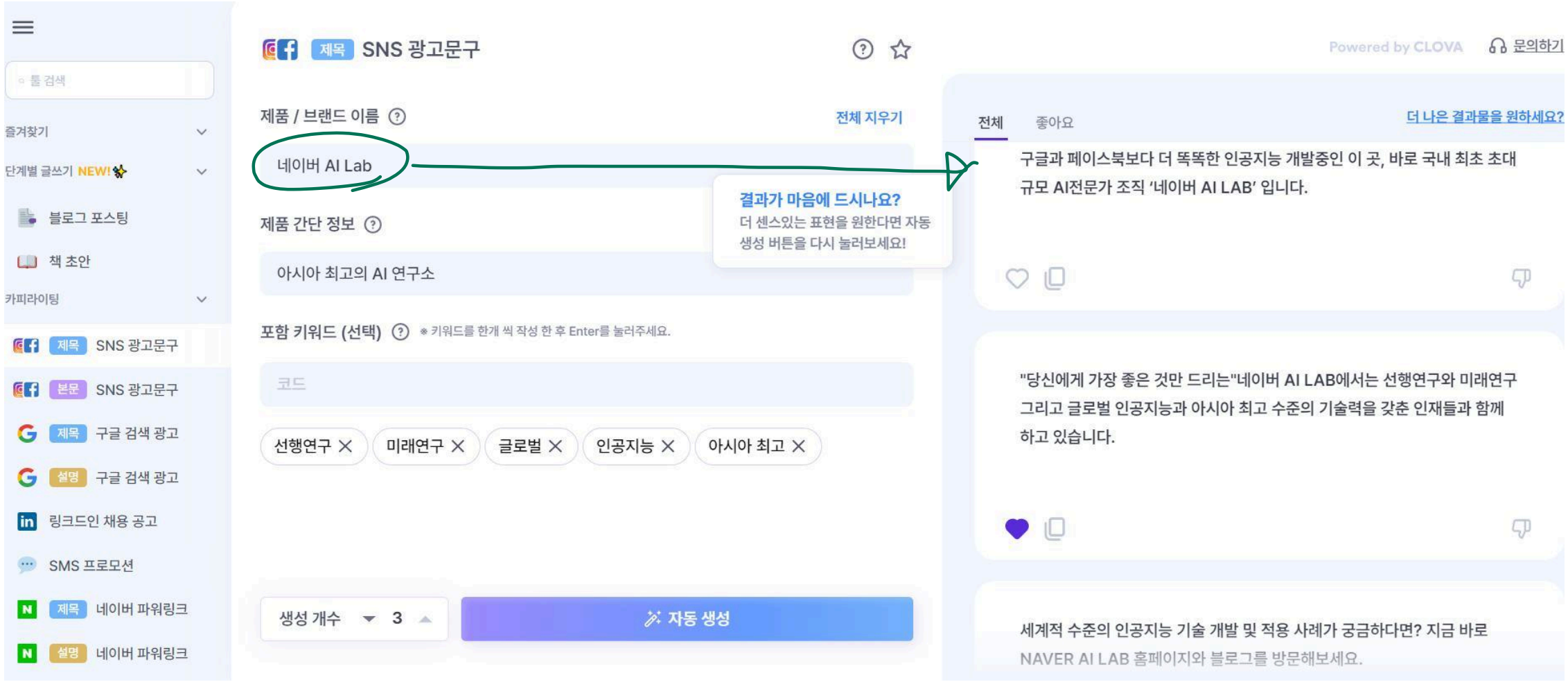


https://github.com/features/copilot

HyperCLOVA (Korean Version of GPT-3) of NAVER

https://www.ncloud.com/product/aiService/clovaStudio

WRTN: Document Auto-Completion based on HyperCLOVA



https://wrtn.io/

Summary

- Models are getting bigger and bigger.
- Owing to self-supervised learning techniques, the language generation capability is getting better and better.
- We are getting closer to artificial general intelligence.