CLIP: Connecting Text and Images

DMQA Open Seminar

2022.05.20

유이경

발표자 소개

❖ 유이경

- 고려대학교 산업경영공학과
- Data Mining & Quality Analytics Lab
- M.S. Student (2021.03 ~)

❖ Research Interest

- Machine Learning Algorithms
- Multi-task learning
- Multi-modal learning

❖ Contact

- E-mail: ylk0801@korea.ac.kr

Contents

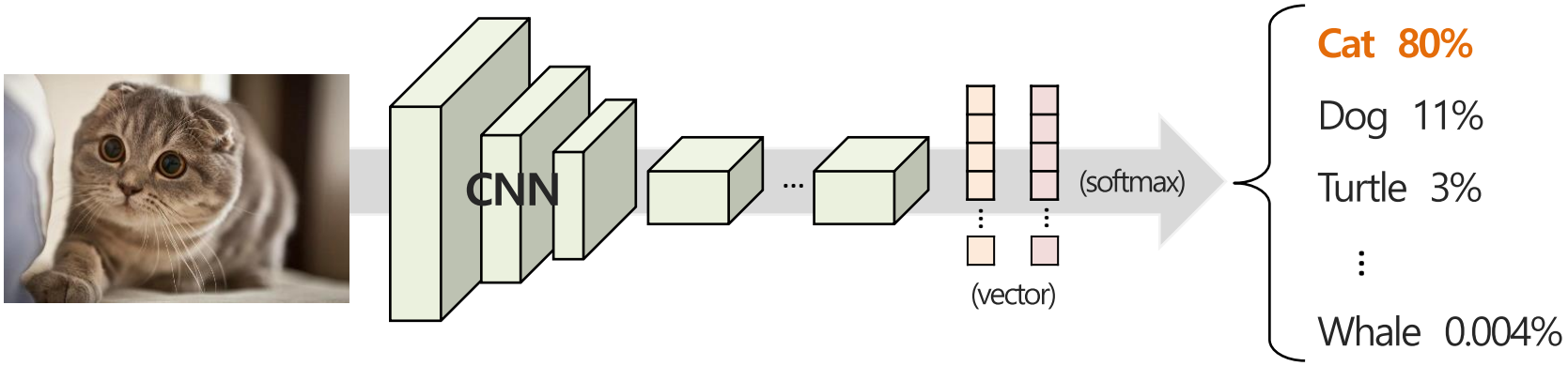
1. Introduction

Introduction

Background

❖ Image classification

- 입력 이미지를 사전에 정의한 클래스 중 하나로 분류하는 문제
- 일반적으로 이미지를 입력, 클래스 레이블을 출력으로 하는 지도학습 적용

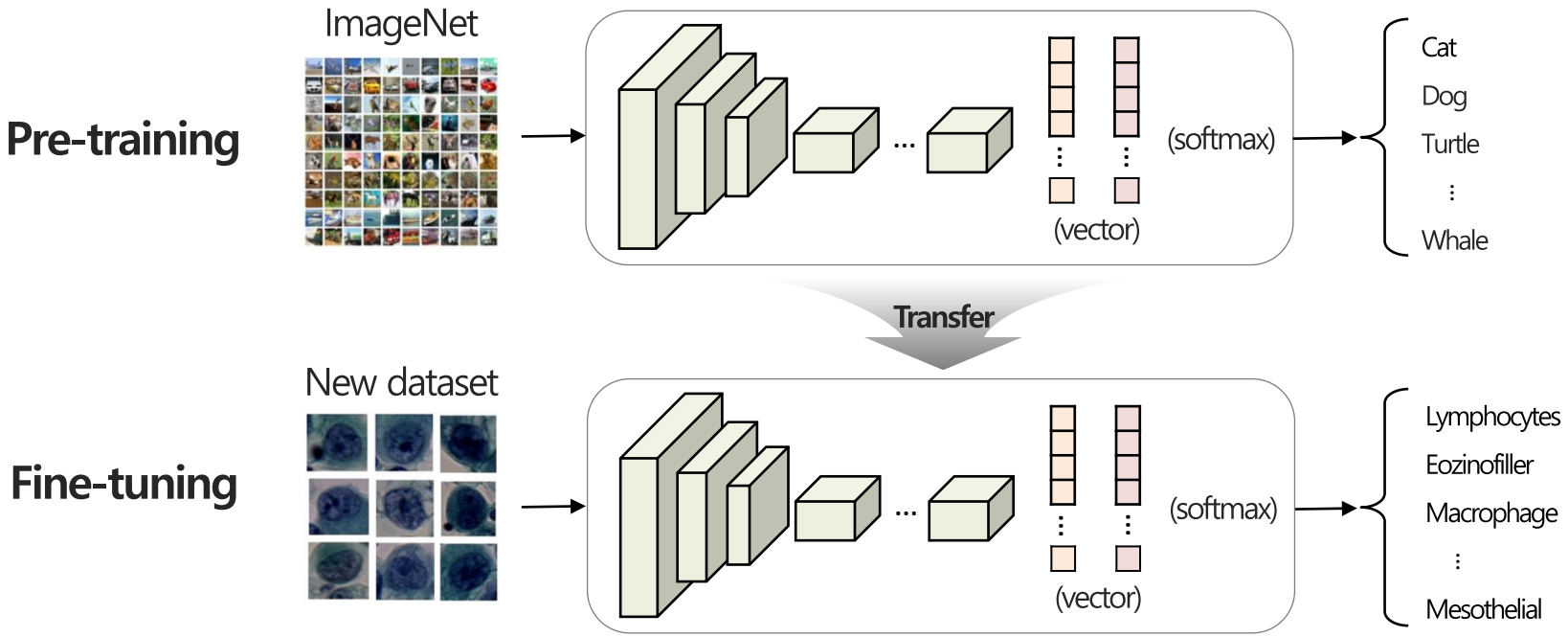


Introduction

Background

❖ **Traditional supervised image classifiers**

- 대규모 데이터셋을 기반으로 모델을 pre-training (ResNet, VGGNet, EfficientNet, …)
- 학습된 모델을 fine-tuning하여 다양한 downstream task에 활용



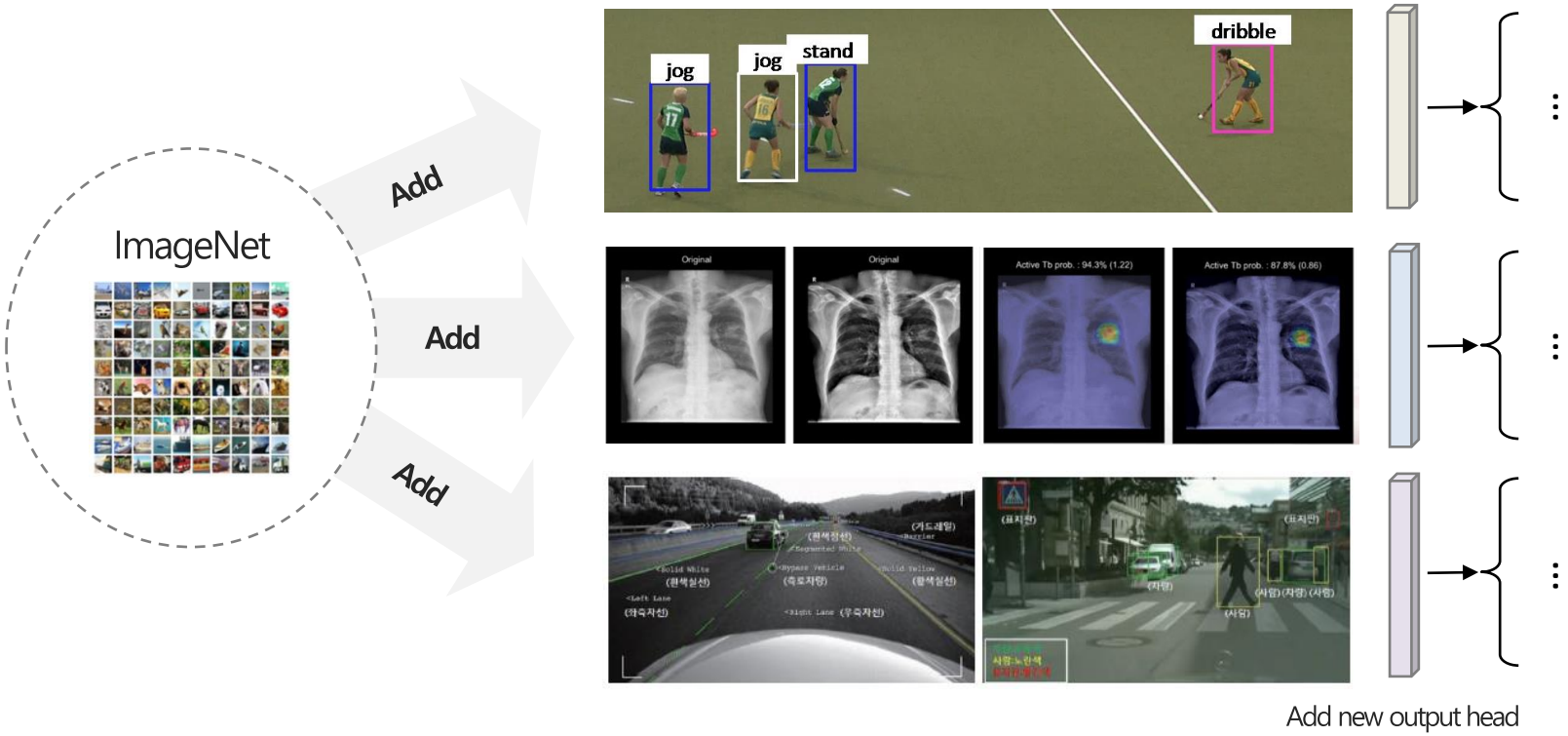
Baykal, E., Dogan, H., Ercin, M. E., Ersoz, S., & Ekinci, M. (2020). Transfer learning with pre-trained deep convolutional neural networks for serous cell classification. Multimedia Tools and Applications, 79(21), 15593-15611.

Introduction

Background

❖ Limitation(1)

- Fine-tuning 없이 새로운 downstream task에 적용하기 어려움 = 모델의 일반화↓



https://medium.com/@akr.rali76/human-activity-recognition-d141baed2ac
https://www.etnews.com/20210315000122?m=1, http://www.aitimes.com/news/articleView.html?idxno=140424

Introduction

Background

❖ Limitation(2)

- 새로운 downstream task에 적합한 다량의 이미지와 레이블링 작업을 요함
- 이미지 수집 및 정답 레이블 생성에 많은 인력과 비용이 요구됨

Data labeling process

라벨러 및 검수자 교육

데이터 레이블링

데이터 검수

Introduction

Background

❖ **Limitation (3)**

- 벤치마크 데이터셋 성능과 실제 현실에서 수집한 데이터셋 성능과는 차이 존재
- 벤치마크 데이터셋에 최적화되어 그 외 데이터셋에서는 저조한 성능을 보임 = 모델의 강건성↓

ImageNet

gap

Real-world dataset

https://openai.com/blog/clip/

Borji, A. (2021). Contemplating real-world object classification. arXiv preprint arXiv:2103.05137.

Introduction

Background

❖ **What we need is …**

Pre-training 시,

✓ Fine-tuning이 필요 없는 **일반화된 모델**

✓ 이미지 수집 및 정답 레이블 생성에 **적은 노력이 드는 모델**

✓ 벤치마크 데이터셋 외 여러 현실 데이터셋에서도 좋은 성능을 보이는 **강건한 모델**

Introduction

Background

❖ **What we need is …**

Pre-training 시,

✓ Fine-tuning이 필요 없는 **일반화된 모델**

✓ 이미지 수집 및 정답 레이블 생성에 **적은 노력이 드는 모델**

✓ 벤치마크 데이터셋 외 여러 현실 데이터셋에서도 좋은 성능을 보이는 **강건한 모델**

CLIP (Contrastive Language-Image Pre-training)

2. CLIP (Contrastive Language-Image Pre-training)

CLIP (Contrastive Language-Image Pre-Training)

Paper

❖ **Learning transferable visual models from natural language supervision (ICML 2021)**

- OpenAI에서 발표, 2022년 5월 17일 기준 991회 인용
- Web-based image-text pair를 기반으로 visual representation을 사전학습하는 방법론, CLIP을 제안

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

Abstract

SOTA computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any other visual concept. Learning directly from raw text about images is a promising alternative which leverages a much broader source of supervision. We demonstrate that the simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to learn SOTA image representations from scratch on a dataset of 400 million (image, text) pairs collected from the internet. After pre-training, natural language is used to reference learned visual concepts (or describe new ones) enabling zero-shot transfer of the model to downstream tasks. We study performance on over 30 different computer vision datasets, spanning tasks such as OCR, action recognition in videos, geo-localization, and many types of fine-grained object classification. The model transfers non-trivially to most tasks and is often competitive with a fully supervised baseline without the need for any dataset specific training. For instance, we match the accuracy of the original ResNet50 on ImageNet zero-shot without needing to use any of the 1.28 million training examples it was trained on. We release our code and pre-trained model weights at https://github.com/OpenAI/CLIP.

interface (McCann et al., 2018; Radford et al., 2019; Raffel et al., 2019) has enabled task-agnostic architectures to zero-shot transfer to downstream datasets. Flagship systems like GPT-3 (Brown et al., 2020) are now competitive across many tasks with bespoke models while requiring little to no dataset specific training data.

These results suggest that the aggregate supervision accessible to modern pre-training methods within web-scale collections of text surpasses that of high-quality crowd-labeled NLP datasets. However, in other fields such as computer vision it is still standard practice to pre-train models on crowd-labeled datasets such as ImageNet (Deng et al., 2009). Could scalable pre-training methods which learn directly from web text result in a similar breakthrough in computer vision? Prior work is encouraging.

Joulin et al. (2016) demonstrated that CNNs trained to predict words in image captions can learn representations competitive with ImageNet training. Li et al. (2017) then extended this approach to predicting phrase n-grams in addition to individual words and demonstrated the ability of their system to zero-shot transfer to other image classification datasets. Adopting more recent architectures and pre-training approaches, VirTex (Desai & Johnson, 2020), ICMLM (Bulent Sariyildiz et al., 2020), and ConVIRT (Zhang et al., 2020) have recently demonstrated the potential of transformer-based language modeling, masked language modeling, and contrastive objectives to learn image representations from text.

However, the aforementioned models still under-perform

CLIP (Contrastive Language-Image Pre-Training)

Approach

❖ **Using Image-Text pair**

- 기존 일반적인 분류 모델은 이미지의 의미론적 정보를 학습하지 못함
- 반면, CLIP은 이미지와 이미지를 설명하는 텍스트를 결합한 image-text pair를 입력으로 사용
- 이미지와 언어에 대한 representation을 함께 학습하여 일반화된 특징 학습 가능

Visual representation

vs.

Visual representation + Semantic information

CLIP (Contrastive Language-Image Pre-Training)

Approach

❖ **Creating a sufficiently large dataset**

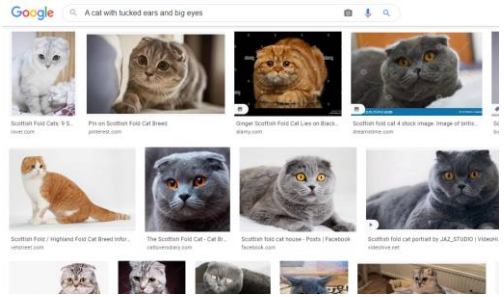
- 인터넷으로부터 레이블링이 필요없는 약 4억개의 Image-Text pair 데이터를 수집
- 다양한 분야의 텍스트 및 이미지를 수집하기 위해 총 50만건의 검색을 수행
- 이미지의 균형을 대략적으로 맞추기 위해 각 검색어당 이미지는 최대 2만개로 조절

ImageNet

✓ 규모: 1천 400만개
✓ 범주: 2만 2천개
✓ 레이블링 인력: 약 2만 5천명

vs.

WebImageText



✓ 규모: 4억개
✓ 범주: 50만개
✓ 레이블링 인력: -

CLIP (Contrastive Language-Image Pre-Training)

Approach

❖ **Selecting an efficient pre-training method**

- Image-Text pair를 사용한 pre-training 기법은 이전부터 존재 (ex. Image captioning)
- 그러나 모델 사이즈가 크며, 학습 및 예측 시간이 길어 비효율적임

Image captioning (Transformer-based)

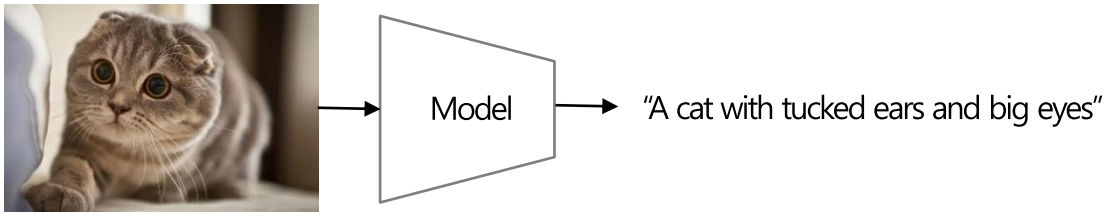
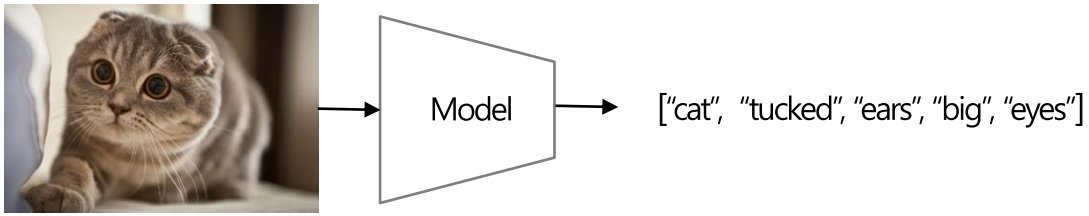


Image captioning (Bag of Words-based)



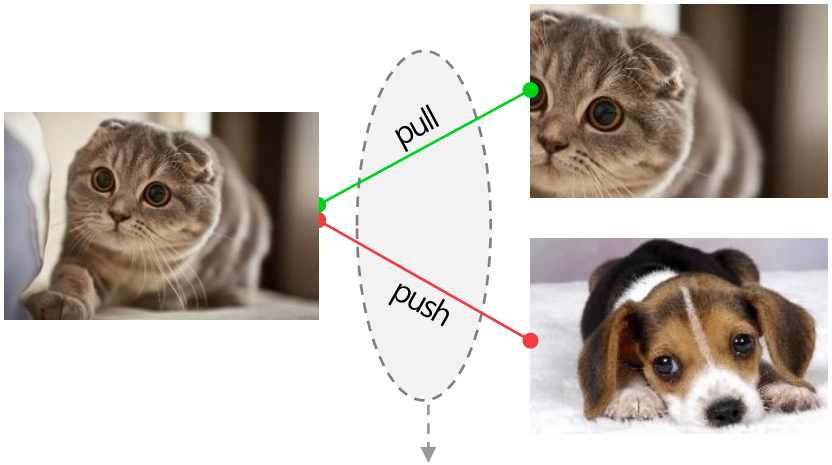
CLIP (Contrastive Language-Image Pre-Training)

Approach

❖ **Selecting an efficient pre-training method**

- 기존의 방식보다 효율적인 contrastive learning을 적용하여 pre-training 진행
- Zero-shot prediction에서도 상대적으로 가장 우수한 성능을 보임
  - ➢ Zero-shot prediction = 특정 하위 문제의 데이터셋 없이 pre-training한 모델 그대로 사용하여 예측

Contrastive learning = "데이터 내 positive & negative samples 간의 관계를 학습"



Positive - 특정 이미지에 대해 비슷하게 만들고 싶은 sample

Negative - 특정 이미지에 대해 다르게 만들고 싶은 sample

샘플을 각각 임베딩하여 positive간 유사도는 크게,
negative간 유사도는 작게 학습

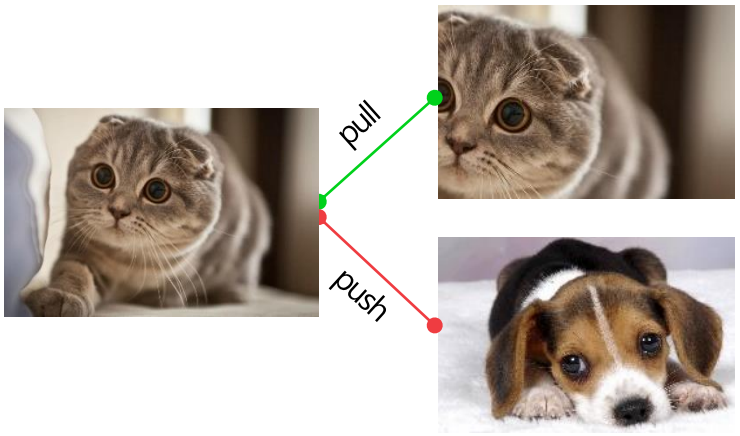
CLIP (Contrastive Language-Image Pre-Training)

Approach

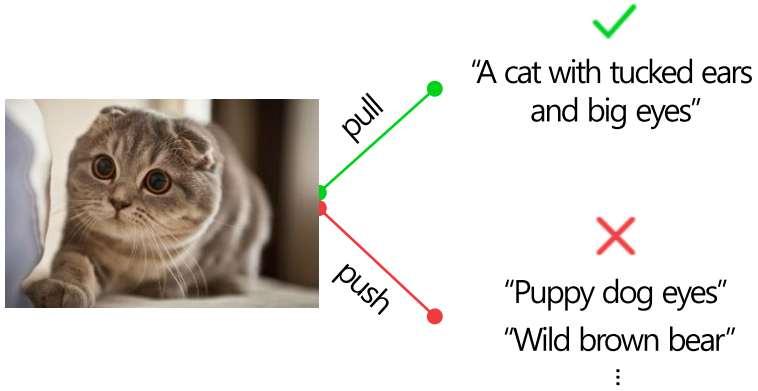
❖ **Selecting an efficient pre-training method**

- 기존의 방식보다 효율적인 contrastive learning을 적용하여 pre-training 진행
- Zero-shot prediction에서도 상대적으로 가장 우수한 성능을 보임
  - ➢ Zero-shot prediction = 한 번도 본 적 없는 특정 하위 문제의 데이터셋에 대해 예측 수행

Contrastive learning = "데이터 내 positive & negative samples 간의 관계를 학습"



Contrastive learning (Image-Image pair)



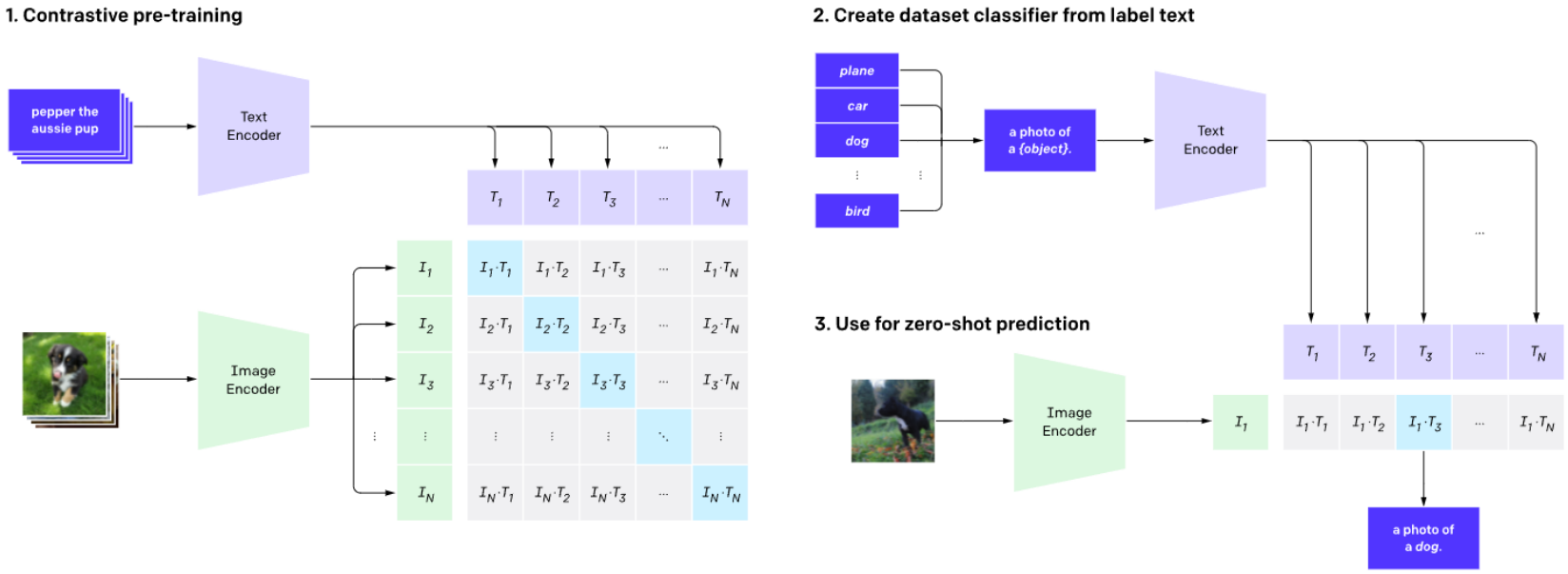
Contrastive learning (Image-Text pair)
→ **Connecting Text and Images**

CLIP (Contrastive Language-Image Pre-Training)

Methodology

❖ Overview

1. Contrastive pre-training
2. Create dataset classifier from label text
3. Use for zero-shot prediction

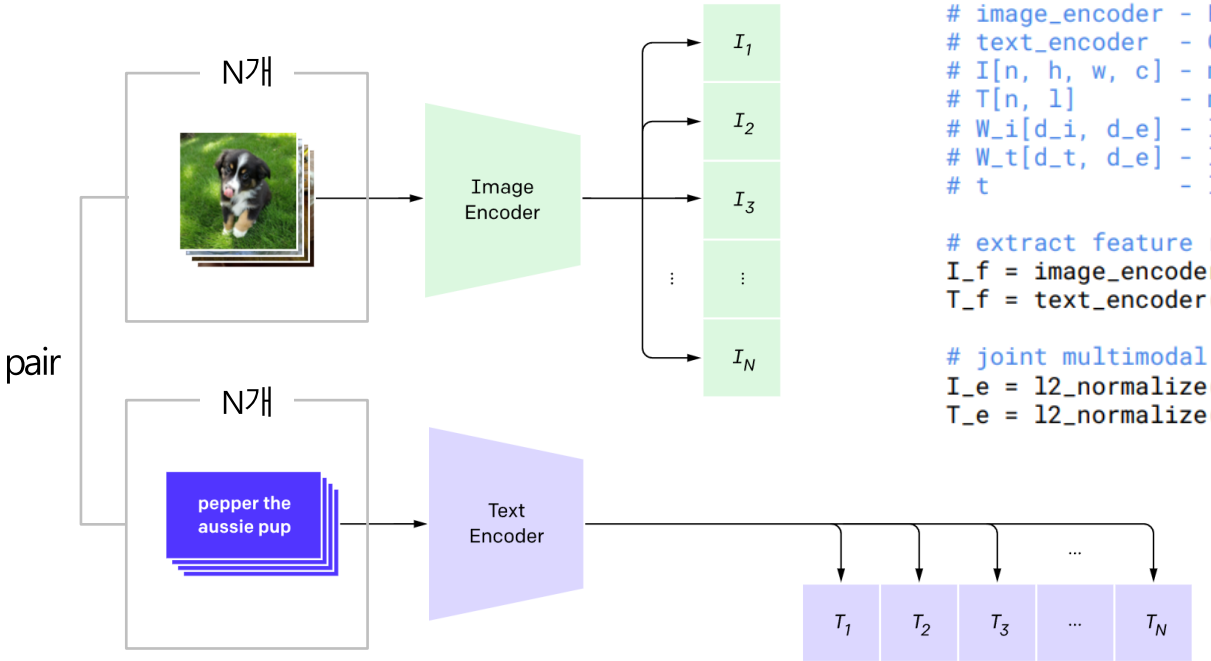


CLIP (Contrastive Language-Image Pre-Training)

Methodology

1. Contrastive pre-training

- 배치 단위로 이루어진 N개의 이미지와 텍스트를 각각 인코더에 통과시켜 임베딩 벡터 산출
  - ➢ Image encoder : Modified ResNet / Vision Transformer
  - ➢ Text encoder : Transformer



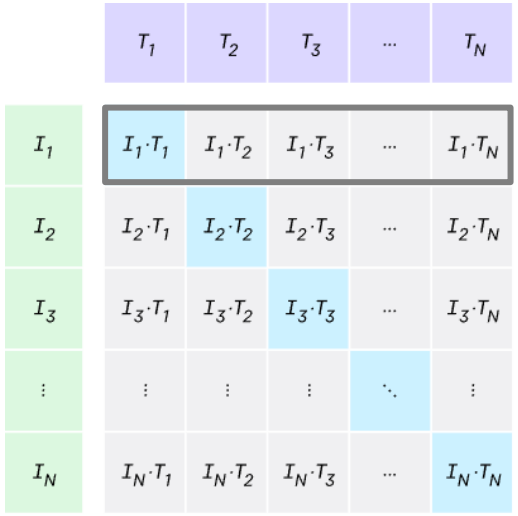

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)
```

CLIP (Contrastive Language-Image Pre-Training)

Methodology

1. Contrastive pre-training

- 이미지와 텍스트 벡터간의 내적을 통해 코사인 유사도 계산
- Pair에 해당하지 않는 이미지 혹은 텍스트는 서로 다르다는 관계하에 cross-entropy loss 계산



총 N^2개의 쌍
- N개는 positive pair
- (N^2-N)개는 negative pair

```
# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

i 번째 이미지의 $loss$

$$= \sum_{k=1}^{C} I(k=i) \cdot \left(-\log \frac{\exp(I_i \cdot T_K)}{\sum_{p=1}^{N} \exp(I_i \cdot T_p)} \right) = -\log \frac{\exp(I_i \cdot T_K)}{\sum_{p=1}^{N} \exp(I_i \cdot T_p)}$$

softmax

$$I(k=i) = \begin{cases} 1 & if\ k=i \\ 0 & otherwise \end{cases}, k = 1, \dots, N, i = 1, \dots, N$$

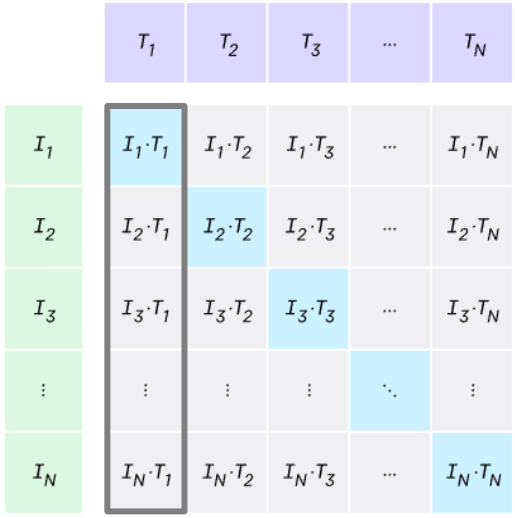
클래스 레이블, 인덱스

CLIP (Contrastive Language-Image Pre-Training)

Methodology

1. Contrastive pre-training

- 이미지와 텍스트 벡터간의 내적을 통해 코사인 유사도 계산
- Pair에 해당하지 않는 이미지 혹은 텍스트는 서로 다르다는 관계하에 cross-entropy loss 계산



총 N^2개의 쌍 { N개는 correct pairing / (N^2-N)개는 incorrect pairing

```
# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

i 번째 텍스트의 $loss$

$$= \sum_{k=1}^{C} T(k=i) \cdot \left(-\log \frac{\exp(T_i \cdot I_K)}{\sum_{p=1}^{N} \exp(T_i \cdot I_p)} \right) = -\log \frac{\exp(T_i \cdot I_K)}{\sum_{p=1}^{N} \exp(T_i \cdot I_p)}$$

softmax

$$\mathrm{T}(k=i) = \begin{cases} 1 & if\ k=i \\ 0 & otherwise \end{cases}, k=1,\dots,N, i=1,\dots,N$$

CLIP (Contrastive Language-Image Pre-Training)

Methodology

1. Contrastive pre-training

- 이미지와 텍스트 벡터간의 내적을 통해 코사인 유사도 계산
- Pair에 해당하지 않는 이미지 혹은 텍스트는 서로 다르다는 관계하에 cross-entropy loss 계산

i 번째 이미지의 $loss$

$$= \sum_{k=1}^{C} I(k=i) \cdot \left(-\log \frac{\exp(I_i \cdot T_K)}{\sum_{p=1}^{N} \exp(I_i \cdot T_p)} \right) = -\log \frac{\exp(I_i \cdot T_K)}{\sum_{p=1}^{N} \exp(I_i \cdot T_p)}$$

i 번째 텍스트의 $loss$

$$= \sum_{k=1}^{C} T(k=i) \cdot \left(-\log \frac{\exp(T_i \cdot I_K)}{\sum_{p=1}^{N} \exp(T_i \cdot I_p)} \right) = -\log \frac{\exp(T_i \cdot I_K)}{\sum_{p=1}^{N} \exp(T_i \cdot I_p)}$$

최종 $loss$

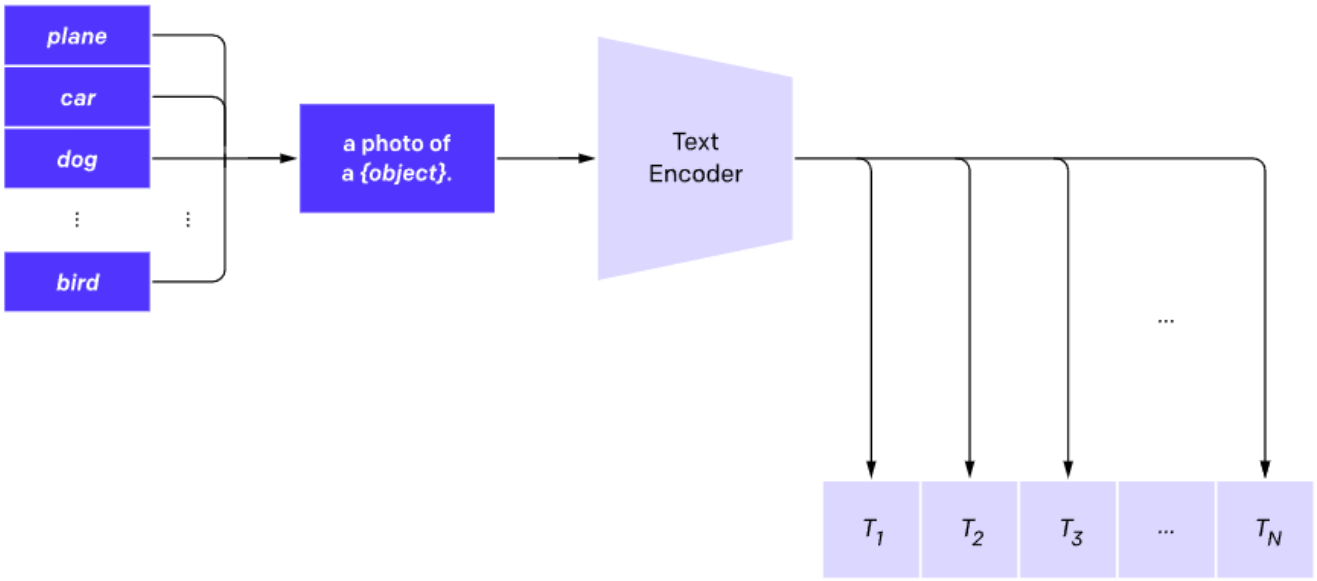
$$= \frac{\text{N개 이미지의 } loss + \text{N개텍스트의 } loss}{2}$$

CLIP (Contrastive Language-Image Pre-Training)

Methodology

2. Create dataset classifier from label text

- 적용하고자 하는 특정 하위 문제의 데이터셋 레이블을 텍스트로 변환
  - ➢ 단순 단어가 아닌 "a photo of { }"에 해당하는 구로 변환
  - ➢ 단어에서 구로 변환하여 인코더 입력 시 성능 향상되었다는 실험 결과
- 이후 학습된 텍스트 인코더에 통과시켜 텍스트 임베딩 벡터값 산출

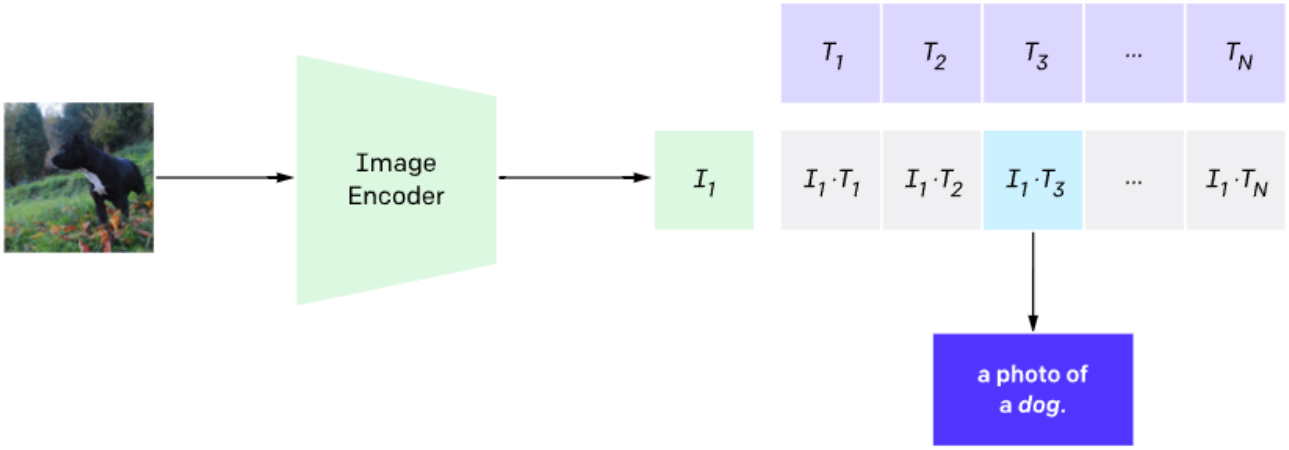


CLIP (Contrastive Language-Image Pre-Training)

Methodology

3. Use for a zero-shot prediction

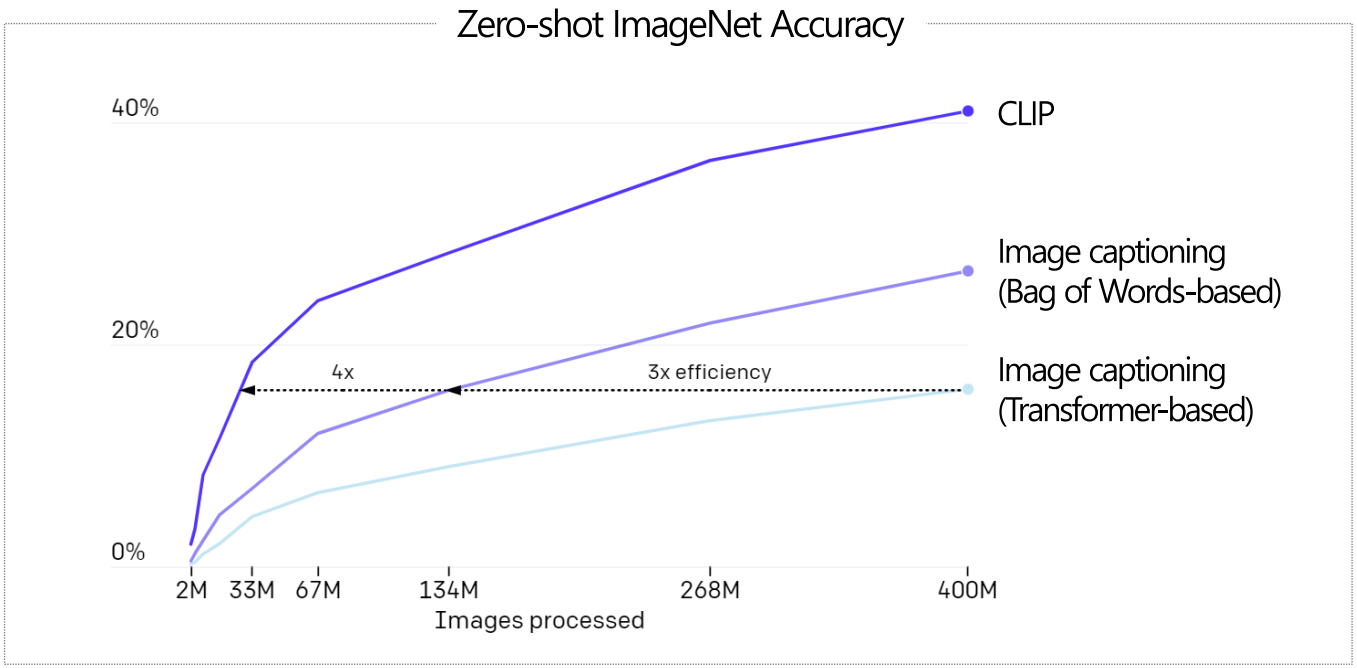
- 예측하고자 하는 이미지를 학습된 이미지 인코더에 통과시켜 이미지 임베딩 벡터값 산출
- 텍스트 임베딩 벡터와 코사인 유사도를 계산하여 상대적으로 높은 값을 갖는 텍스트 선택
- Fine-Tuning 과정을 거치지 않고서도, 처음 보는 이미지에 대해 예측 가능



CLIP (Contrastive Language-Image Pre-Training)

Experiments

❖ **Zero-shot CLIP vs. Zero-shot image captioning**

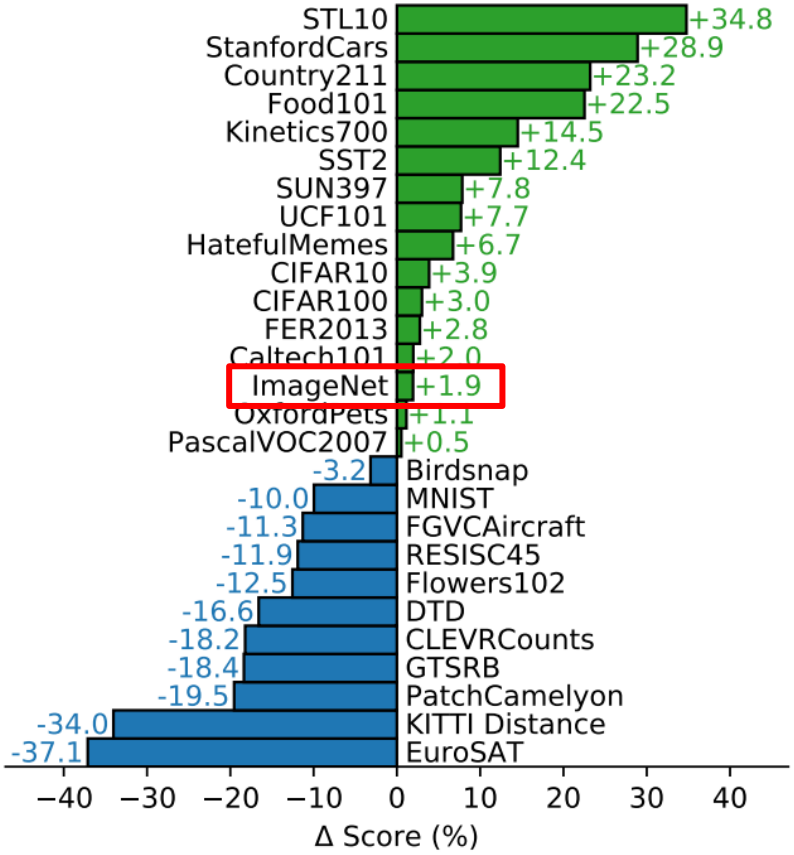


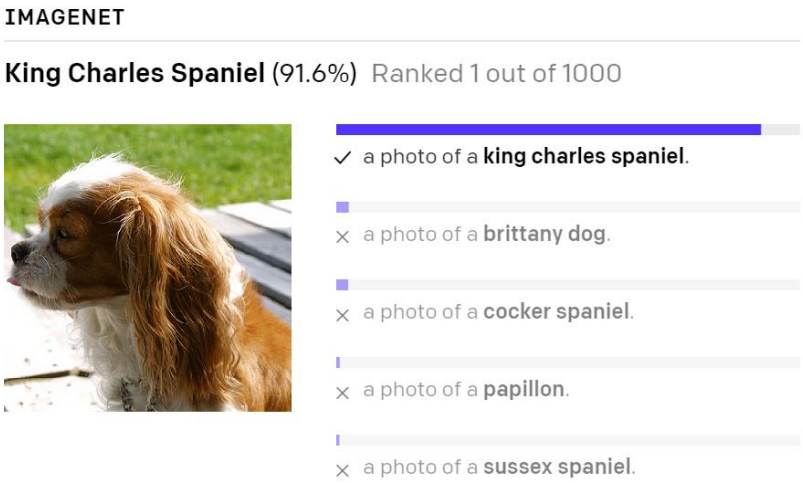
→ 적은 데이터로 가장 높은 예측 정확도를 보이며 효율적인 pre-training 가능

CLIP (Contrastive Language-Image Pre-Training)

Experiments

❖ **Zero-shot CLIP vs. Fully supervised ResNet50**

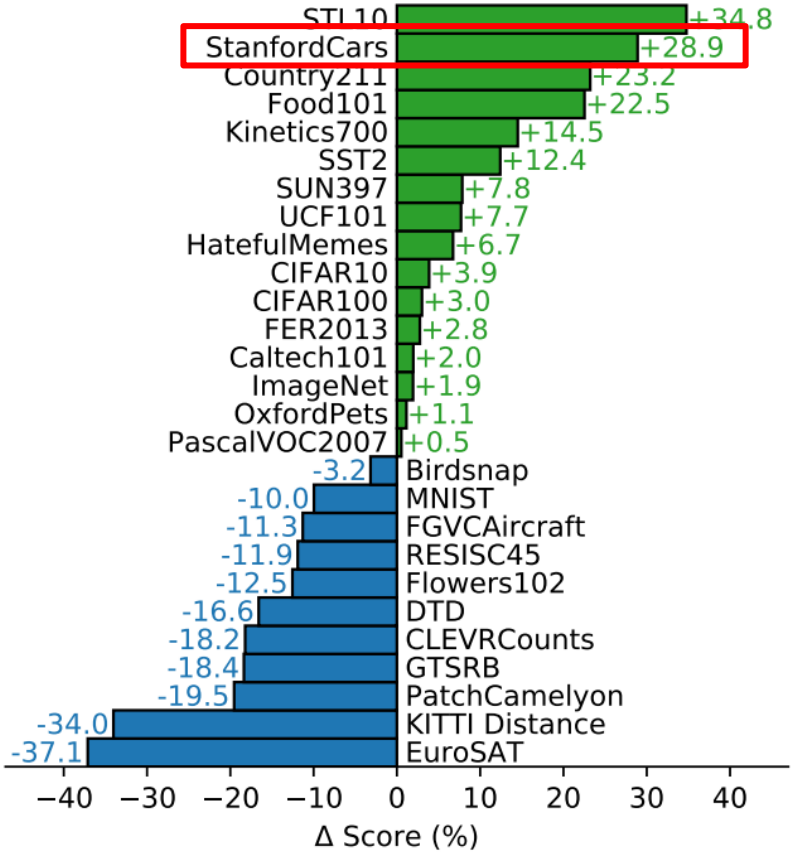


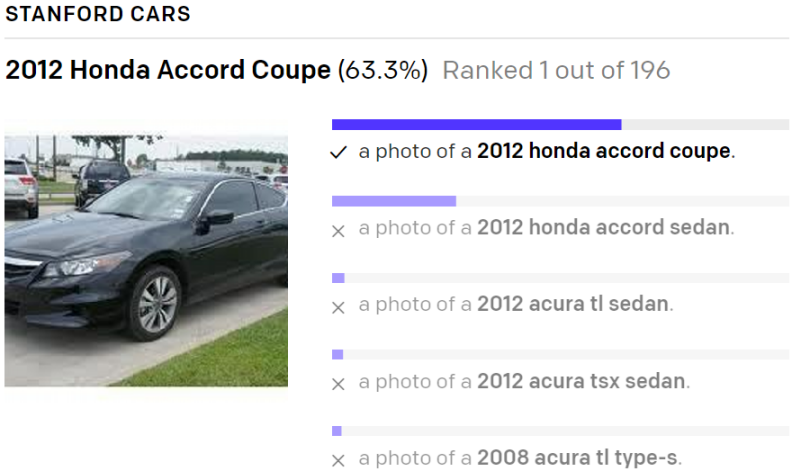


CLIP (Contrastive Language-Image Pre-Training)

Experiments

❖ **Zero-shot CLIP vs. Fully supervised ResNet50**

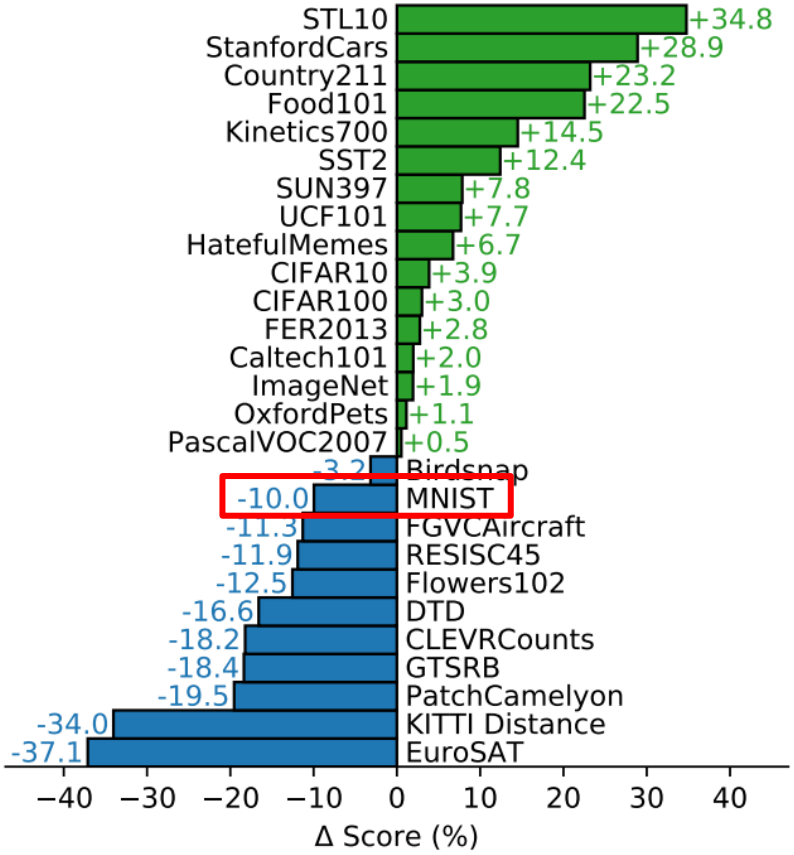


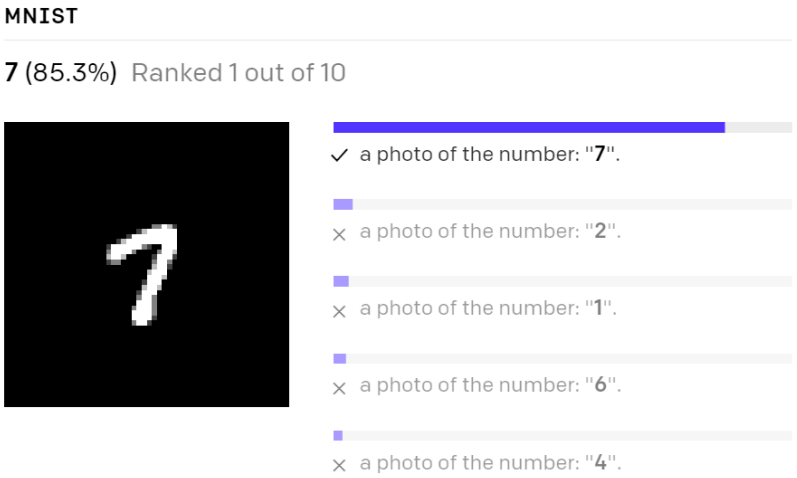


CLIP (Contrastive Language-Image Pre-Training)

Experiments

❖ **Zero-shot CLIP vs. Fully supervised ResNet50**



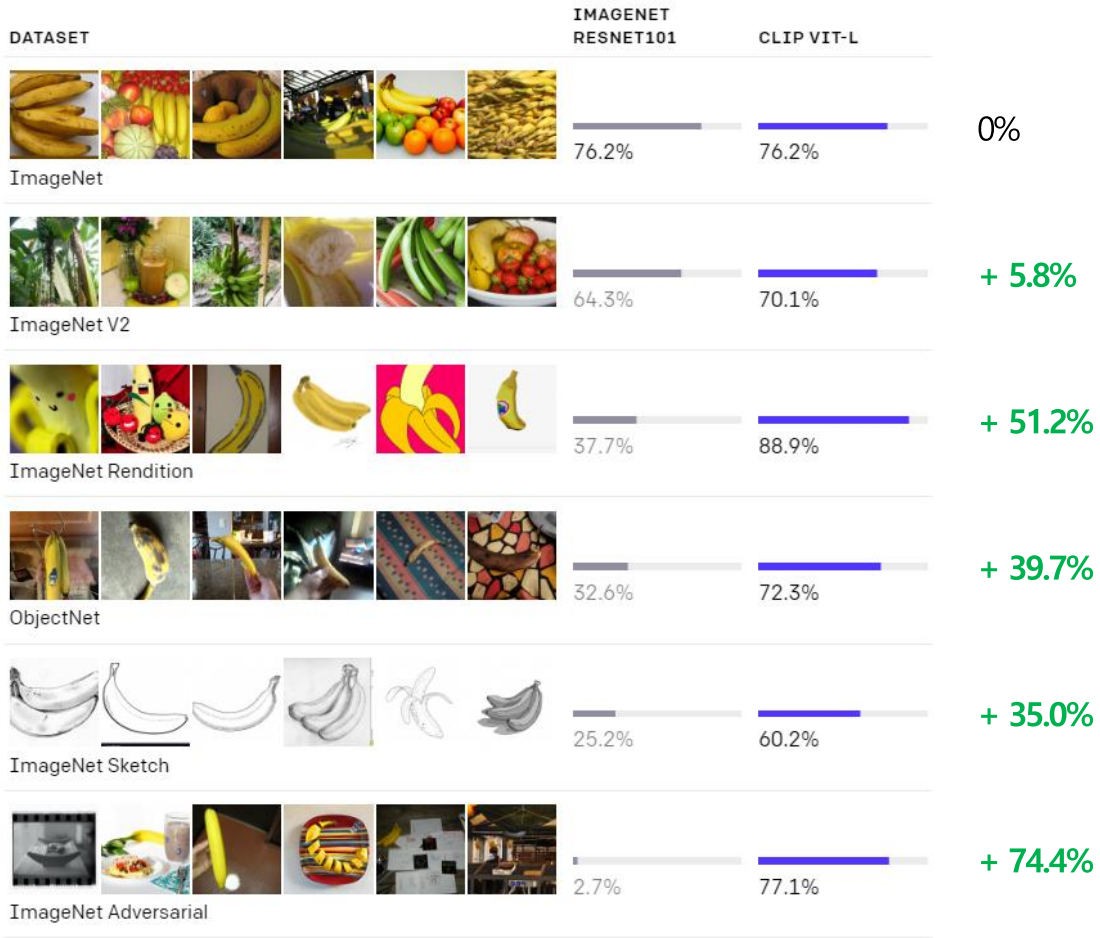


https://openai.com/blog/clip/

CLIP (Contrastive Language-Image Pre-Training)

Experiments

❖ **Robust on natural distribution shift**



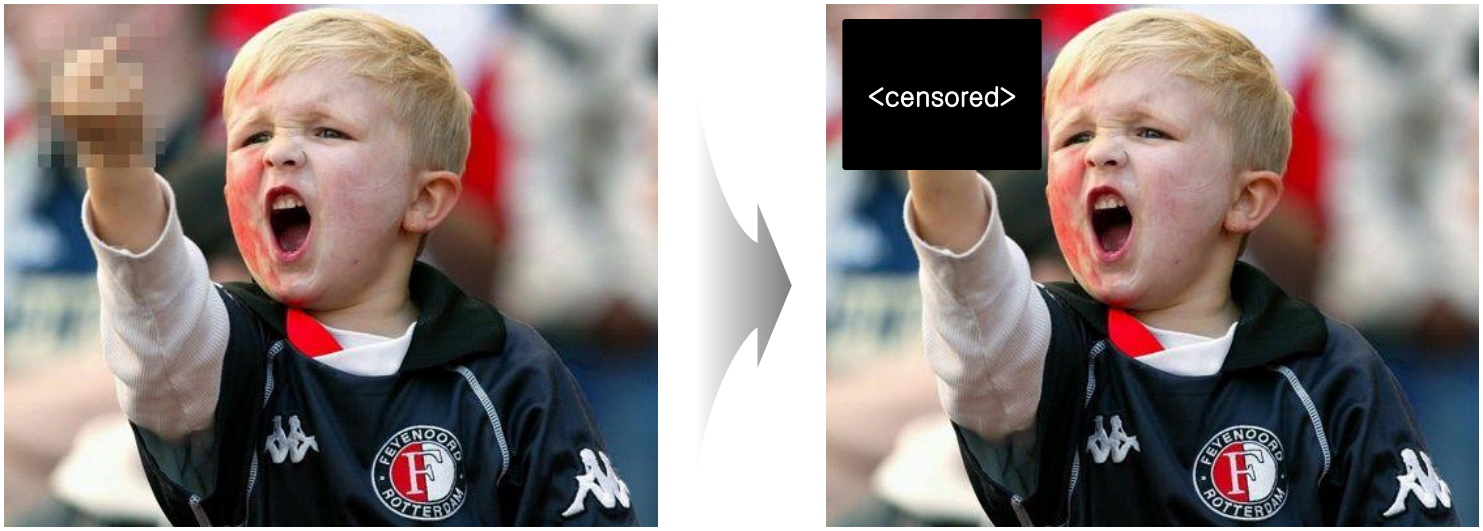
3. Applications

Applications

Image classification

❖ **Content moderation**

- 유해 사항을 포함한 이미지를 탐색하여 수정해주는 알고리즘

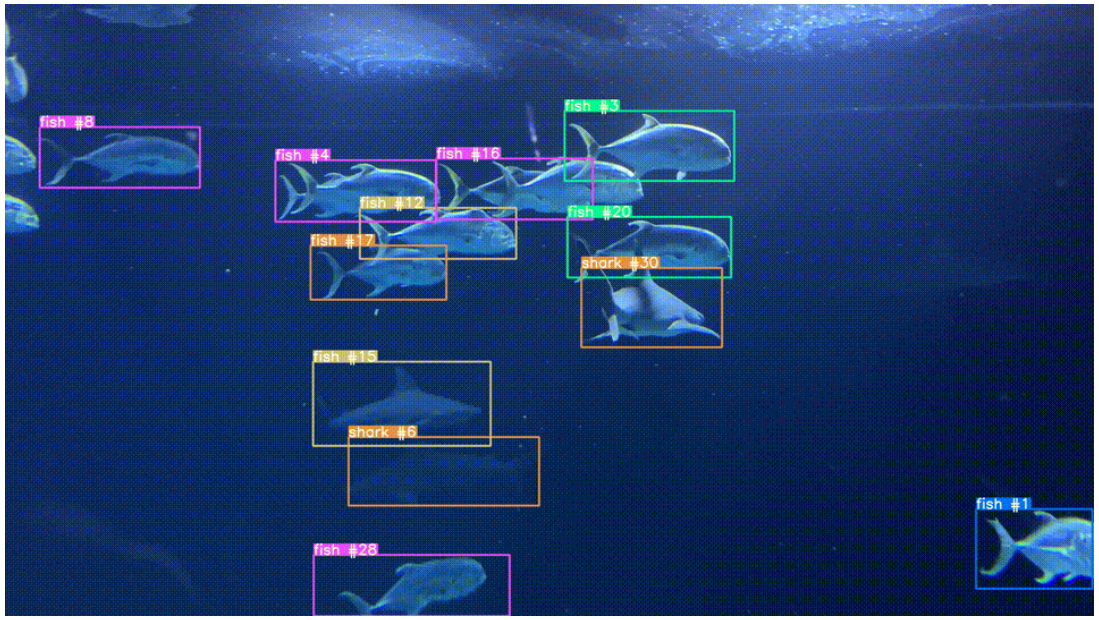


Applications

Image classification

❖ Object tracking

- 프레임간 동일한 객체를 탐지하여 추적하는 알고리즘

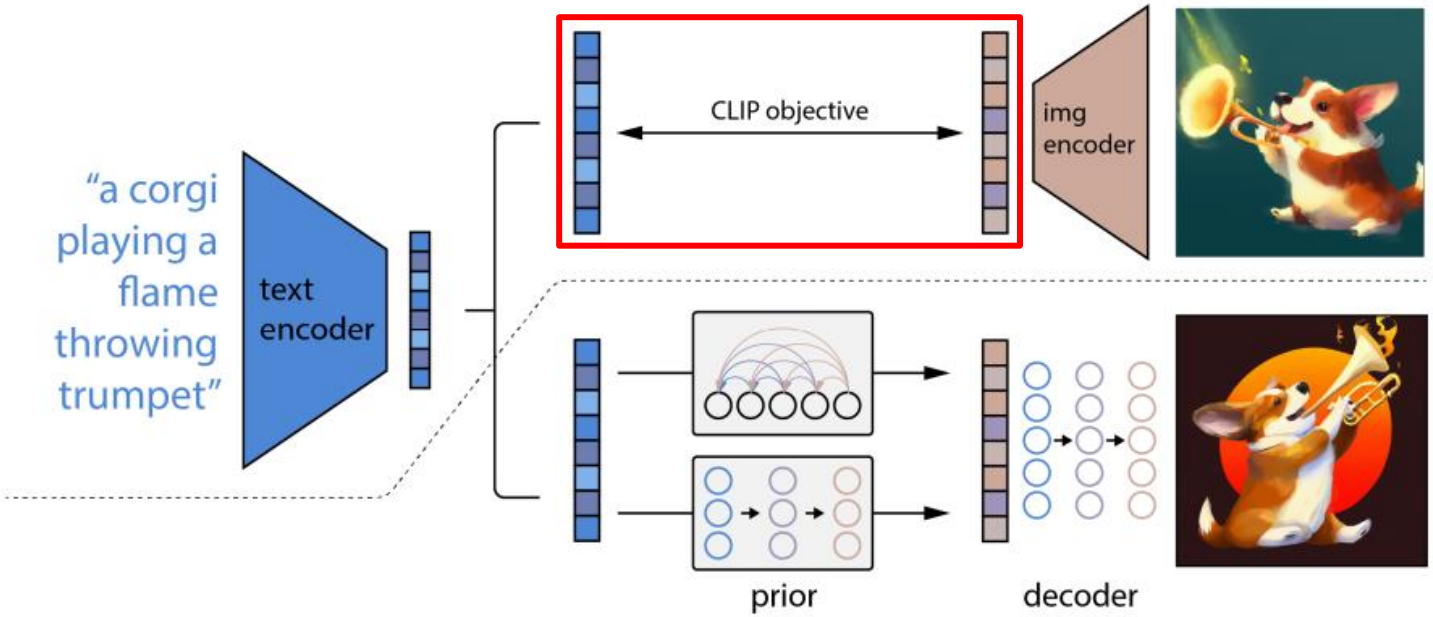


Applications

Image generation

❖ **DALL-E 2**

- 텍스트 설명을 기반으로 이미지로 생성해주는 이미지 생성 알고리즘



Applications

Image generation

❖ **DALL-E 2**

- 텍스트 설명을 기반으로 이미지로 생성해주는 이미지 생성 알고리즘

“An astronaut riding a horse in a photorealistic style”

Applications

Image generation

❖ **DALL-E 2**

- 텍스트 설명을 기반으로 이미지로 생성해주는 이미지 생성 알고리즘

“An astronaut riding a horse as a pencil drawing”

4. Conclusion

Conclusion

Summary

❖ **CLIP: Connecting Text and Images**

[Why?]

- Fine-tuning이 필요 없는 일반화된 모델
- 이미지 수집 및 정답 레이블 생성에 적은 노력이 드는 모델
- 벤치마크 데이터셋 외 여러 현실 데이터셋에서도 좋은 성능을 보이는 강건한 모델

[How?]

- Web-based image-text pair를 기반으로 visual representation을 학습
- 대량의 레이블링이 필요없는 데이터셋을 기반으로 다양한 분야에 대해 학습 가능
- Contrastive learning 기반 pre-training을 통해 효율적이면서 domain shift에 강건한 학습 가능

References

1. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748-8763). PMLR.
2. Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., ... & Olah, C. (2021). Multimodal neurons in artificial neural networks. Distill, 6(3), e30.
3. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125.
4. https://openai.com/blog/clip/
5. https://blog.roboflow.com/clip-model-eli5-beginner-guide/
6. https://blog.roboflow.com/openai-clip/
7. https://jiho-ml.com/weekly-nlp-42/
8. https://dealicious-inc.github.io/2021/03/22/learning-transferable-visual-models.html
9. https://inforience.net/2021/02/09/clip_visual-model_pre_training/
10. http://dmqm.korea.ac.kr/activity/seminar/308

Thank you