Kakaobrain © All rights Reserved.

Recent Trends in Machine Learning: A Large-scale Perspective

A Short Introduction to Multi-modal AI Models (Part 1)

Saehoon Kim @ Kakaobrain

Confidential
상기 문서는 영업비밀보호 및 관련 법률에 따라 보호의 대상이 되는 내용을 포함하고 있습니다. 본 문서에 포함된 정보의 전부 또는 일부를 무단으로 사용하거나 공개 및 배포가 금지됩니다.
This document, including any attachments, is the property of kakaobrain., and may contain confidential information, privileged material or constitute non-public information. Any use, retention, dissemination, distribution or reproduction of this transmission by unintended recipients is strictly prohibited and may be unlawful.

Outline of This Course

CLIP
Encoder-only

05/04

DALL-E
Decoder-only

05/11

DALL-E 2
Enc-Dec

05/18

Kakaobrain © All rights Reserved.

Kakaobrain © All rights Reserved.

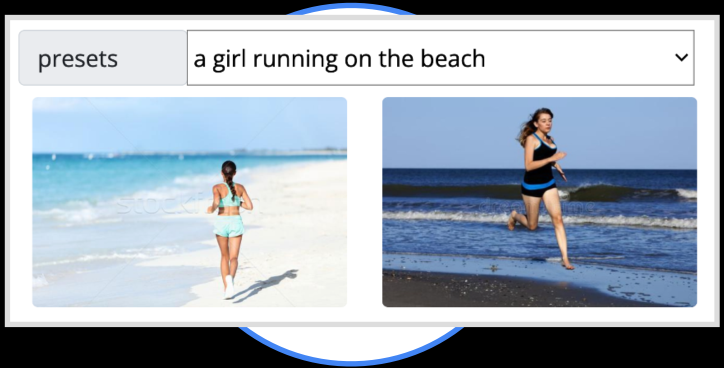
Outline of This Course

Contrastive Learning

DALL-E
Decoder-only

DALL-E 2
Enc-Dec

Outline of This Course



DALL-E
Decoder-only

DALL-E 2
Enc-Dec

Kakaobrain © All rights Reserved.

Outline of This Course

Contrastive Learning

Autoregressive Model

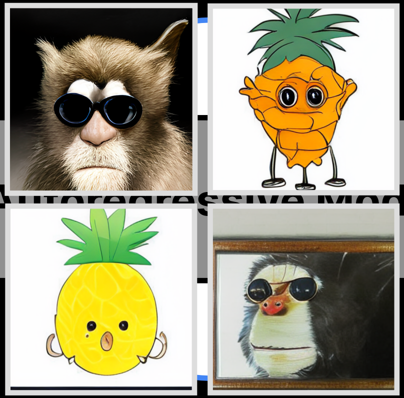
DALL-E 2
Enc-Dec

Kakaobrain © All rights Reserved.

Outline of This Course

Contrastive Learning

Autoregressive Model



DALL-E 2
Enc-Dec

Kakaobrain © All rights Reserved.

Outline of This Course

Contrastive Learning

Autoregressive Model

Diffusion Model

Kakaobrain © All rights Reserved.

Outline of This Course

Contrastive Learning

Autoregressive Model

Kakaobrain © All rights Reserved.
Kakaobrain

Kakaobrain © All rights Reserved.

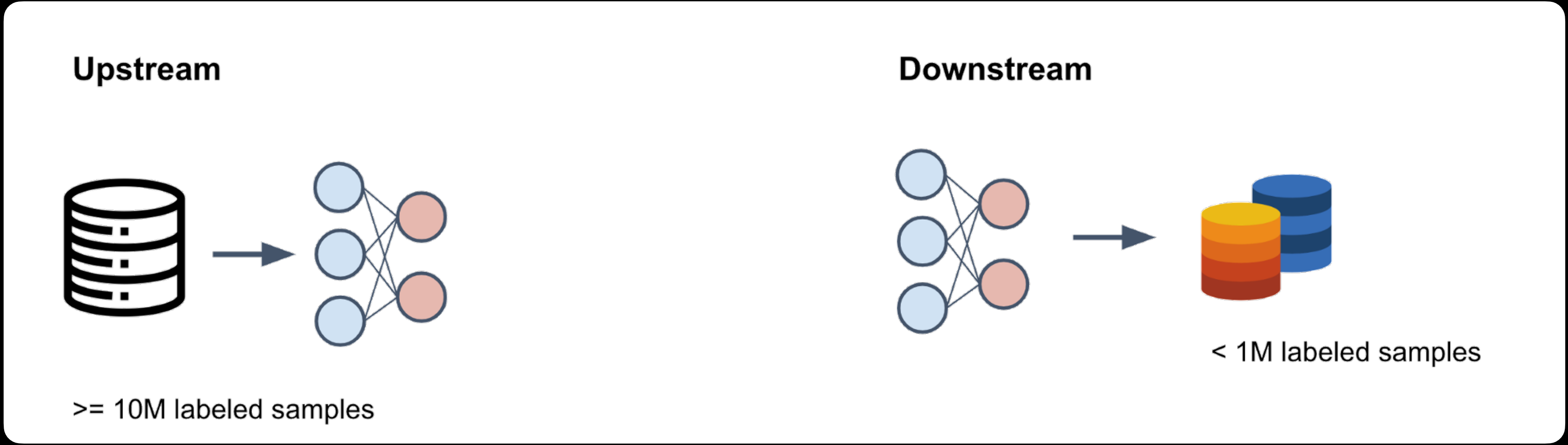
Background

Self-Supervised Representation Learning

Kakaobrain © All rights Reserved.

Transfer Learning

Transferring visual features learned from a large annotated set into small-scale downstream tasks has been significantly improved the performance!

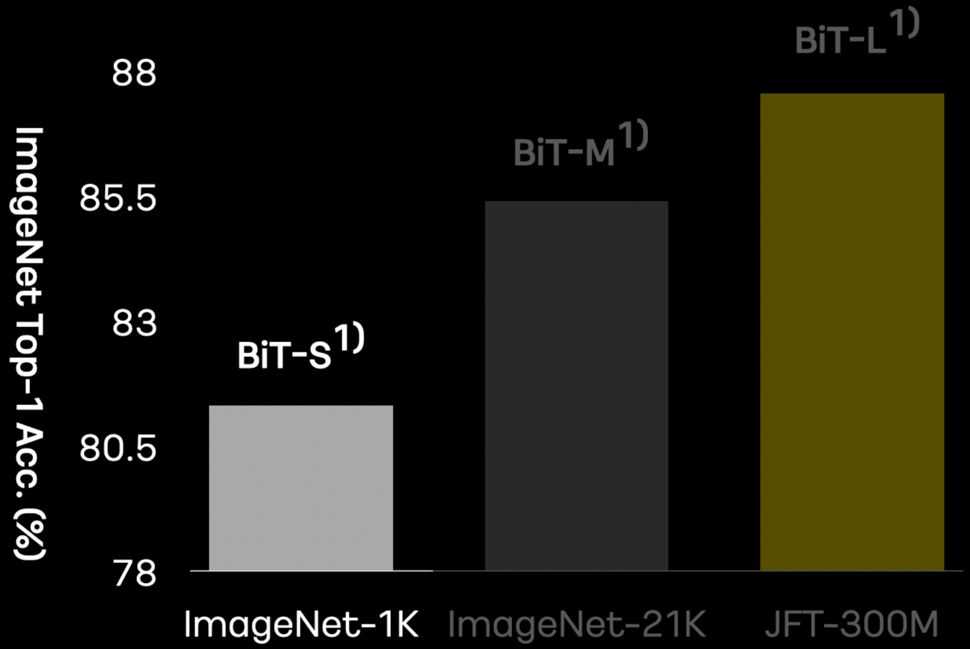


Kakaobrain

Kolesnikov et al., “Big Transfer (BiT): General Visual Representation Learning”, ECCV’20.

Kakaobrain © All rights Reserved.

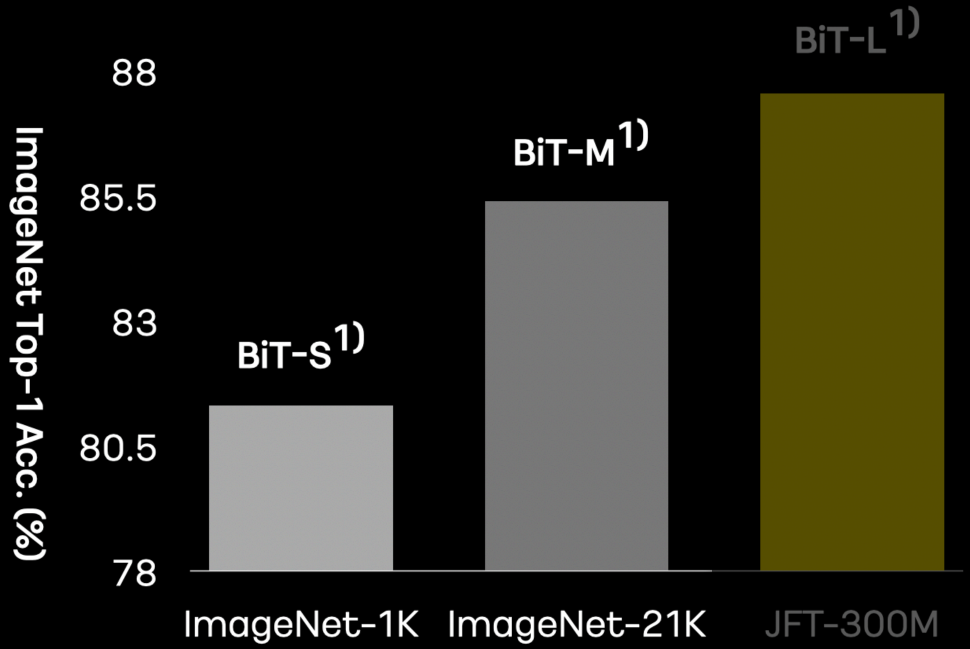
Transfer Learning



Kolesnikov et al., “Big Transfer (BiT): General Visual Representation Learning”, ECCV’20.

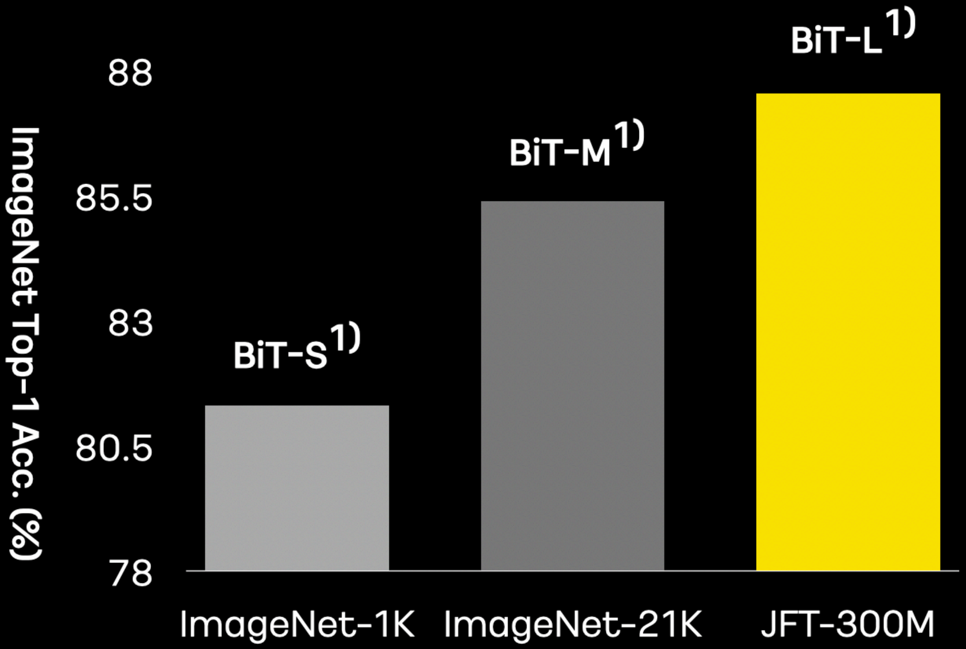
Kakaobrain © All rights Reserved.

Transfer Learning



Kolesnikov et al., "Big Transfer (BiT): General Visual Representation Learning", ECCV'20.

Transfer Learning



Kolesnikov et al., “Big Transfer (BiT): General Visual Representation Learning”, ECCV’20.

Kakaobrain © All rights Reserved.
Kakaobrain

Transfer Learning

Can we learn visual features without labeled samples in the upstream pre-training?

Kolesnikov et al., "Big Transfer (BiT): General Visual Representation Learning", ECCV'20.

Kakaobrain © All rights Reserved.

Contrastive Learning

Learning the global representations by comparing the semantically similar and dissimilar images without human annotations

Kakaobrain © All rights Reserved.

Contrastive Learning

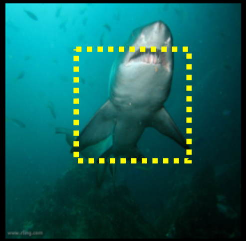
How to automatically obtain similar and dissimilar pairs without labels?

Kakaobrain © All rights Reserved.

Kakaobrain © All rights Reserved.

Contrastive Learning

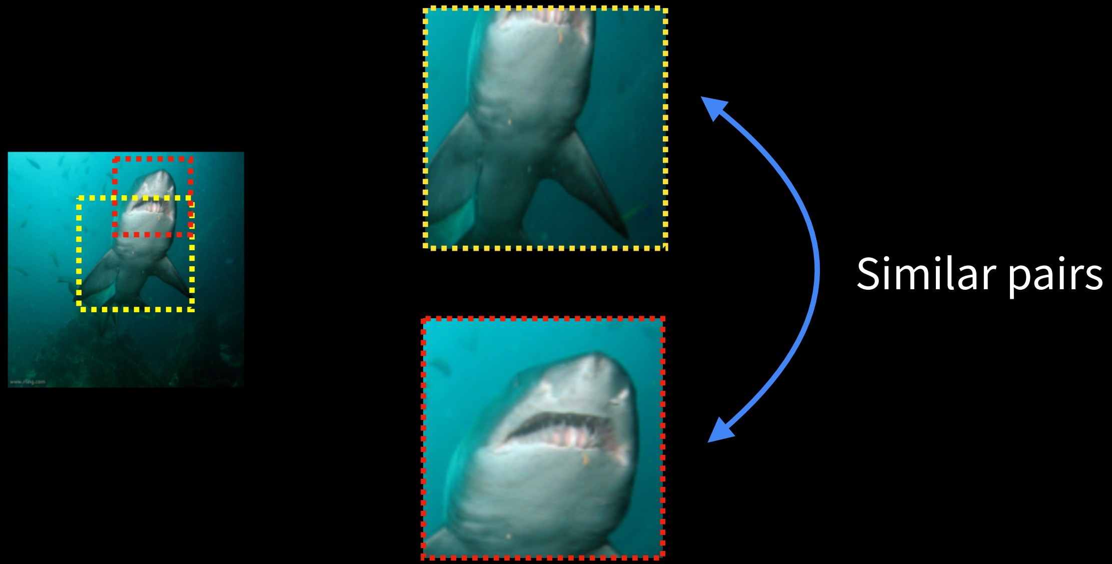
How to automatically obtain similar and dissimilar pairs without labels?



Kakaobrain © All rights Reserved.

Contrastive Learning

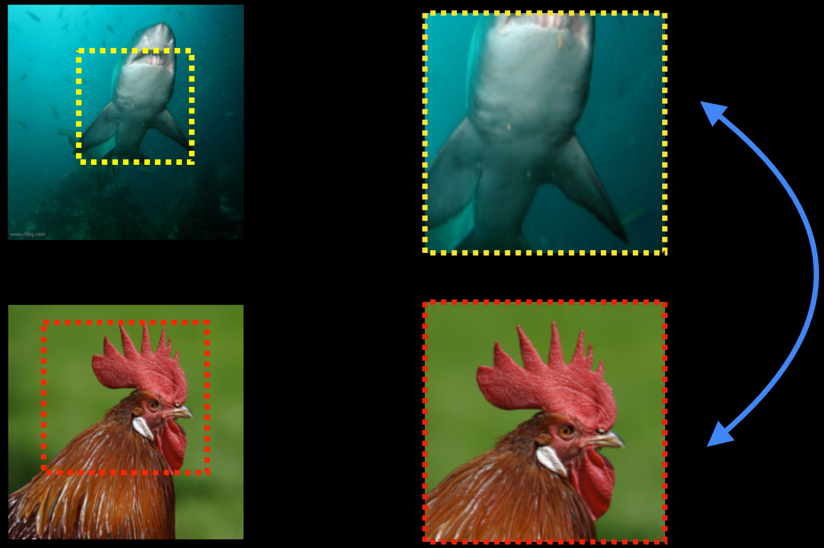
How to automatically obtain similar and dissimilar pairs without labels?



Kakaobrain © All rights Reserved.

Contrastive Learning

How to automatically obtain similar and dissimilar pairs without labels?

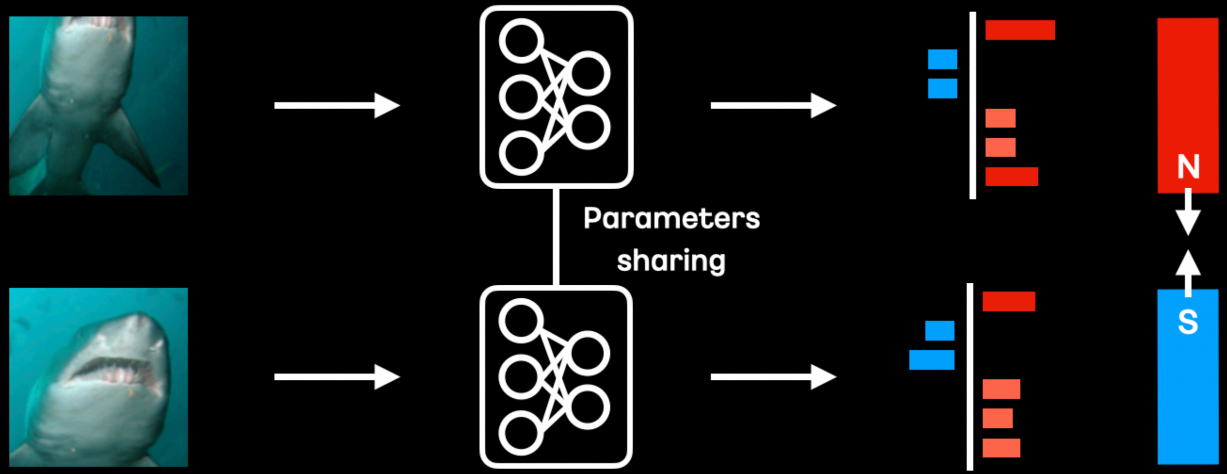


Dissimilar pairs

Kakaobrain © All rights Reserved.

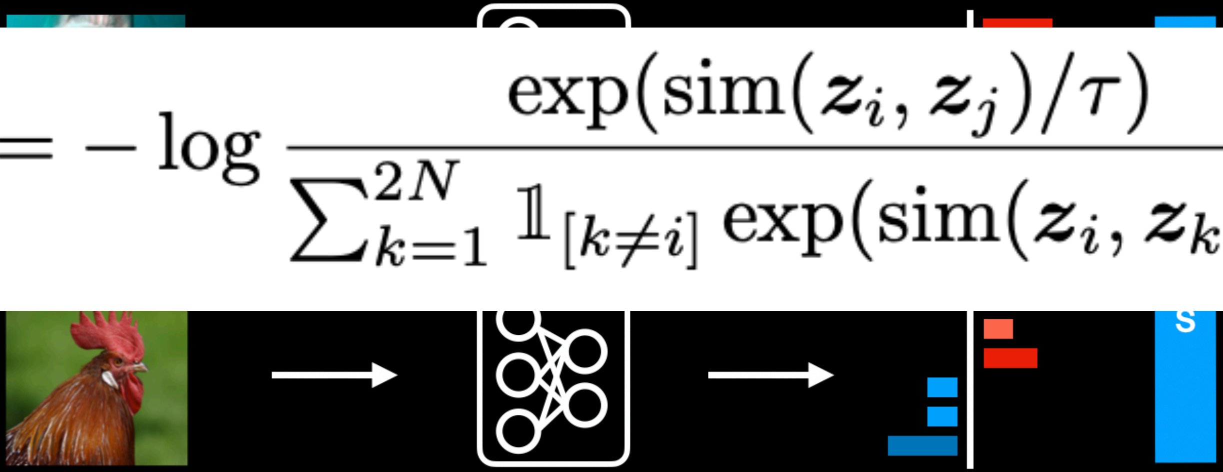
Contrastive Learning

How to automatically obtain similar and dissimilar pairs without labels?



Contrastive Learning

Using a simple contrastive objective to learn global representations

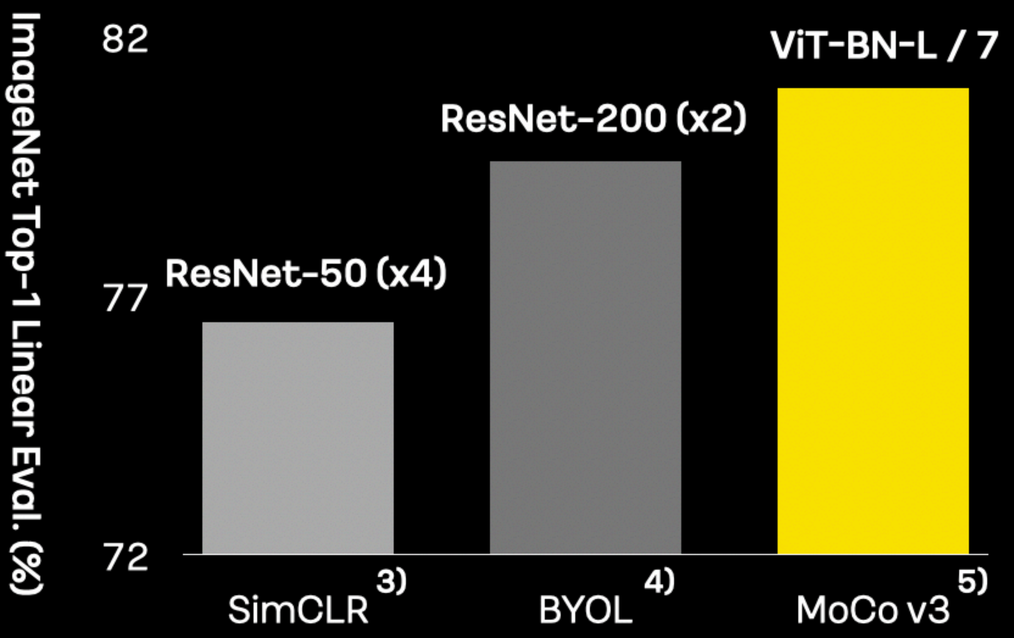


$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

Kakaobrain © All rights Reserved.

Kakaobrain © All rights Reserved.

This simple approach really works well!



Kakaobrain © All rights Reserved.

What's the Next Step?

Self-Supervised Multi-modal Representation Learning

Kakaobrain © All rights Reserved.

CLIP: Connecting Text and Images

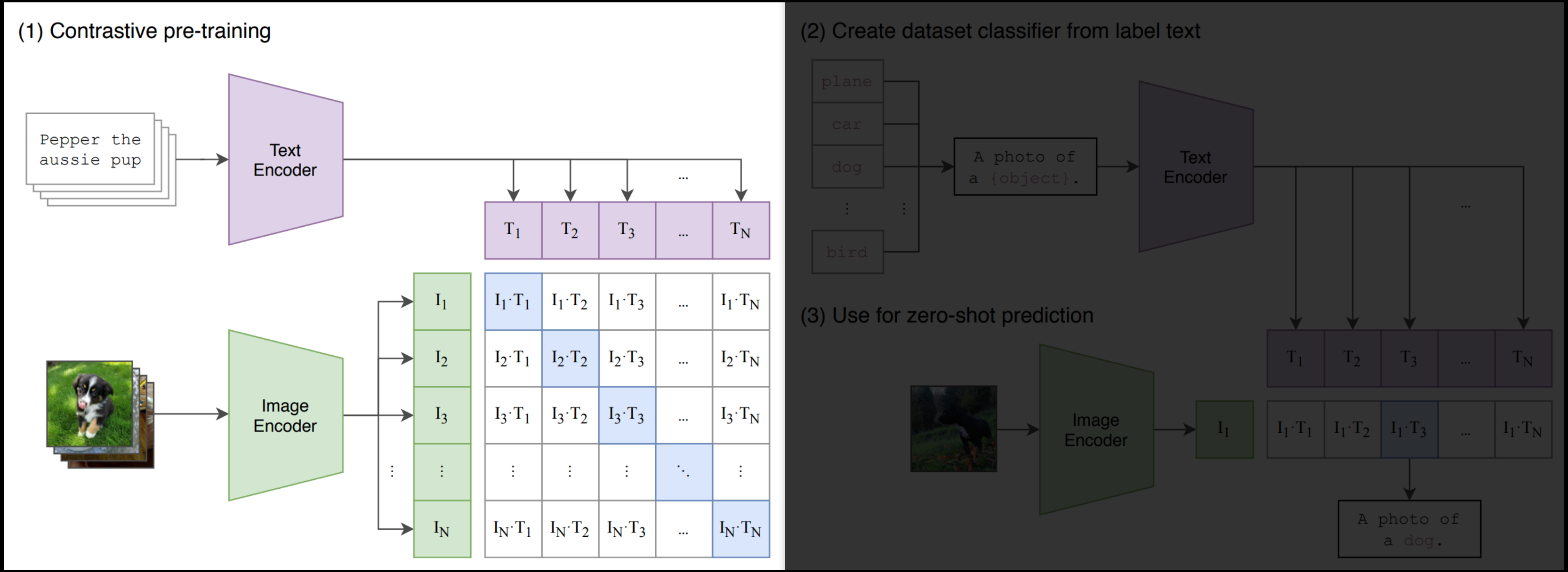
Learning the shared global representations from images and texts!

Kakaobrain

Radford and Kim et al. “Learning Transferrable Visual Models from Natural Language Supervision”, ICML’21.

Kakaobrain © All rights Reserved.

CLIP: Connecting Text and Images



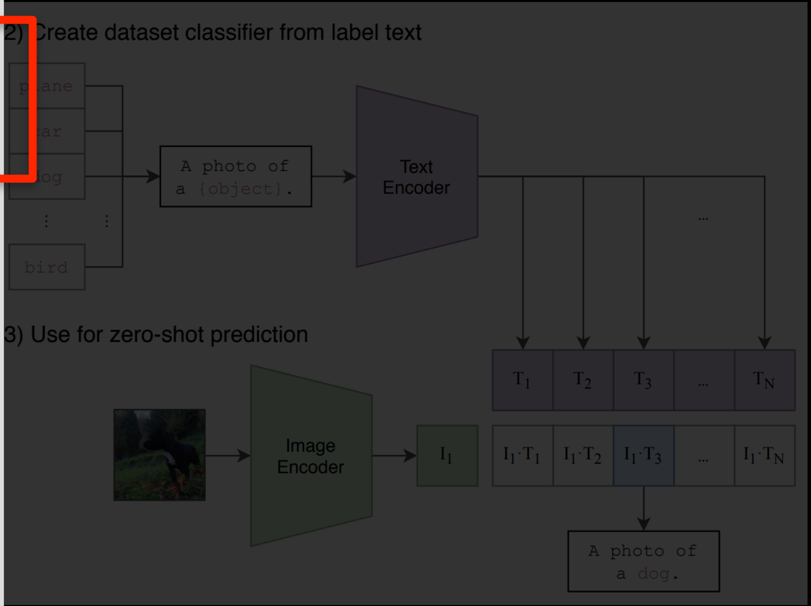
Radford and Kim et al. “Learning Transferrable Visual Models from Natural Language Supervision”, ICML’21.

Kakaobrain

and Images

```
# 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)

# 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
```

Figure 3. Numpy-like pseudocode for the core of an implementation of CLIP.



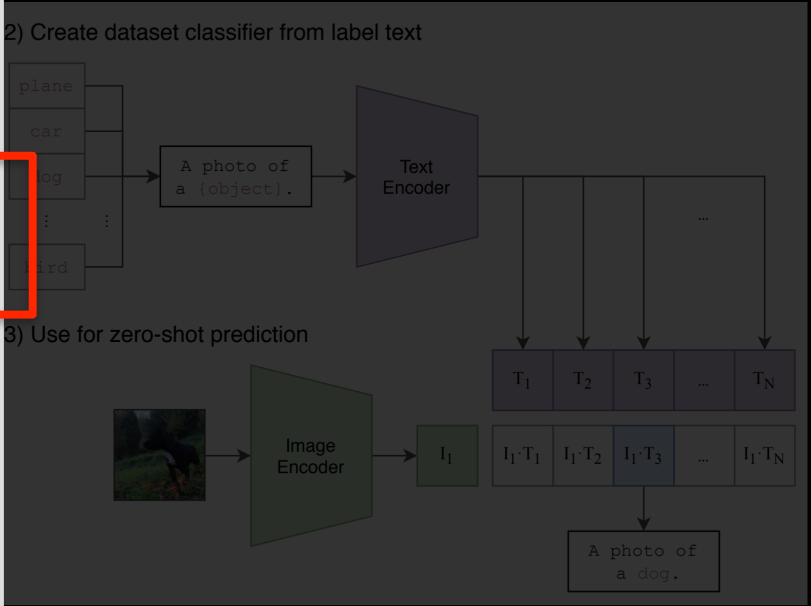
Kakaobrain © All rights Reserved.
Kakaobrain

Radford and Kim et al. "Learning Transferrable Visual Models from Natural Language Supervision", ICML'21.

and Images

```python
# 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)

# 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
```

Figure 3. Numpy-like pseudocode for the core of an implementation of CLIP.

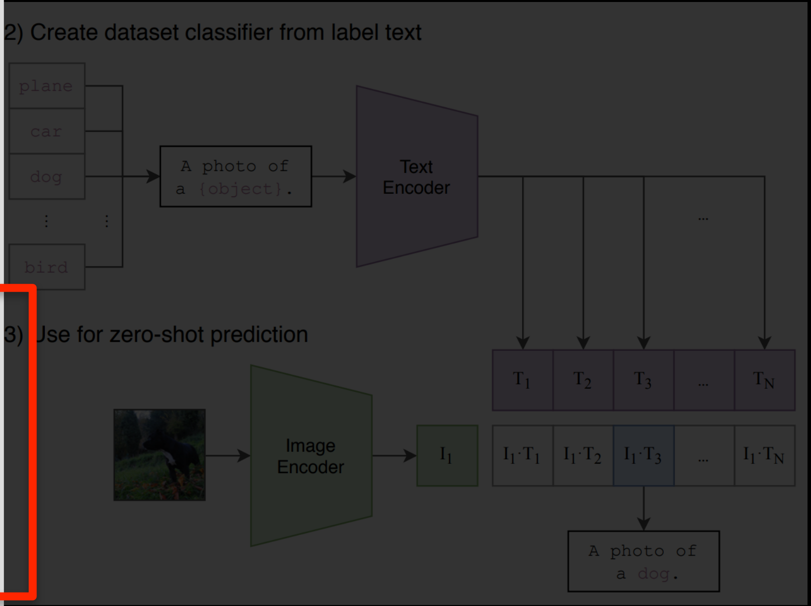


Kakaobrain © All rights Reserved.
Kakaobrain

and Images

```
# 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)

# 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
```

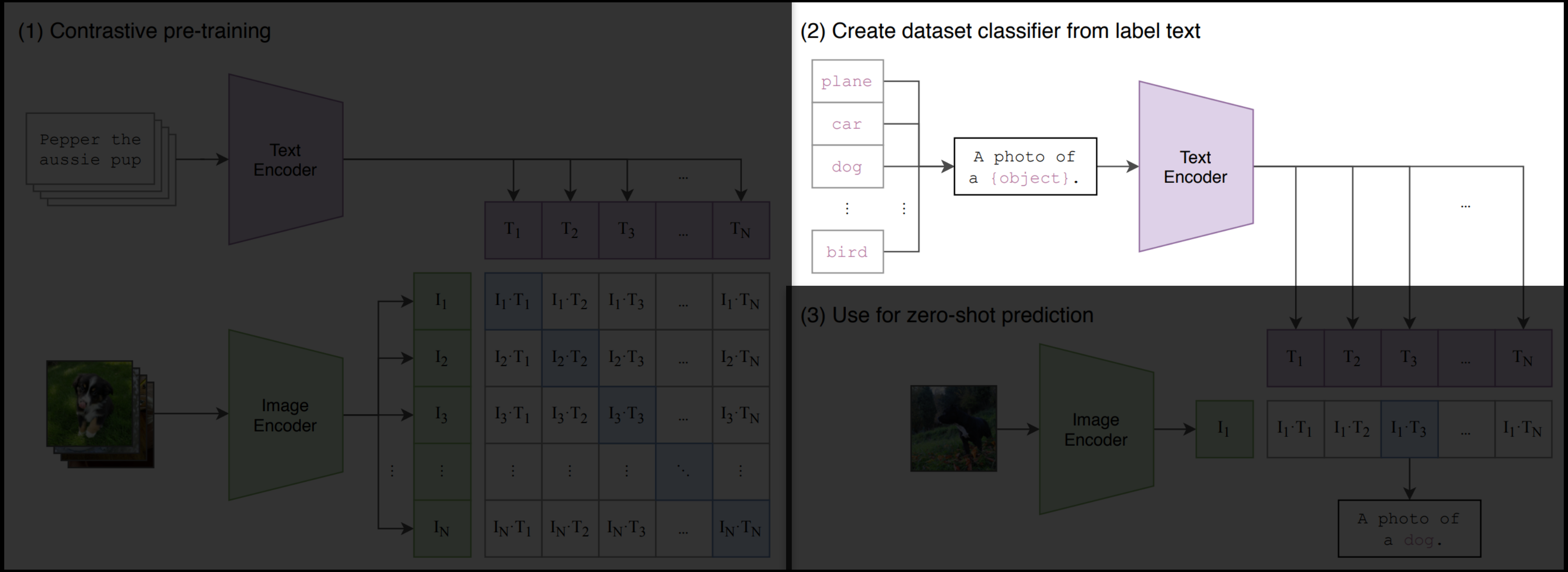
Figure 3. Numpy-like pseudocode for the core of an implementation of CLIP.



Radford and Kim et al. “Learning Transferrable Visual Models from Natural Language Supervision”, ICML’21.

Kakaobrain © All rights Reserved.
Kakaobrain

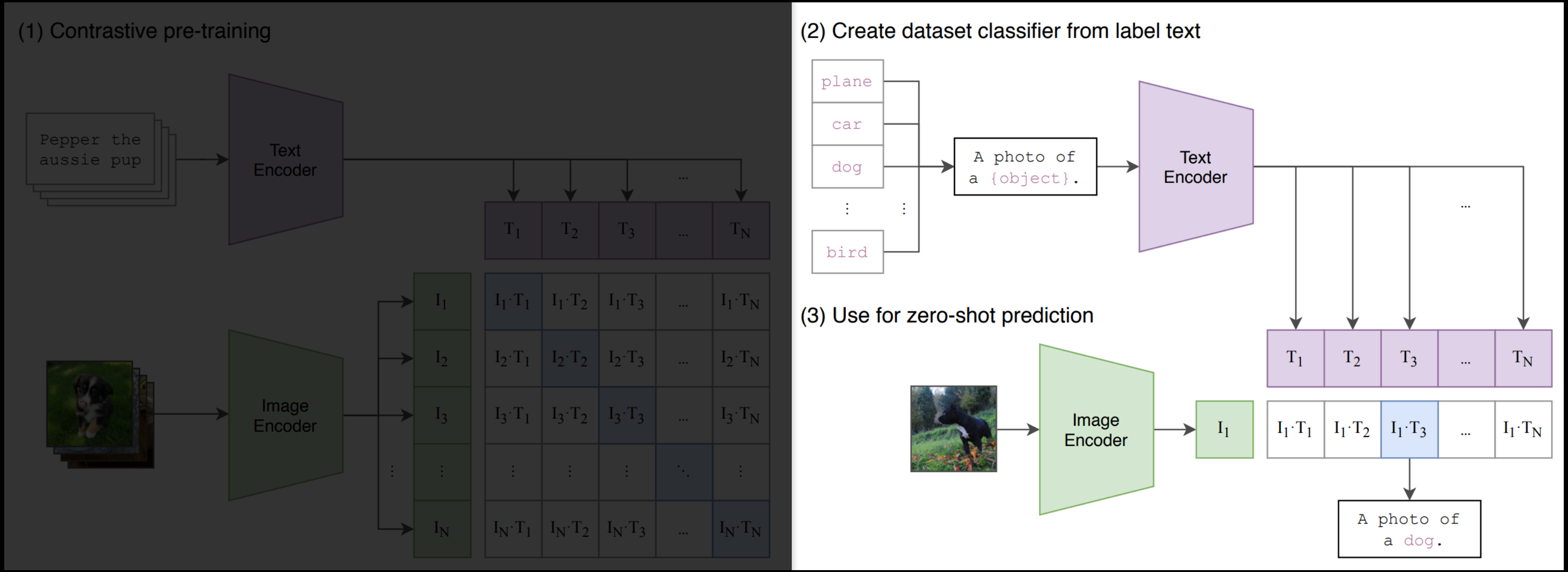
CLIP: Connecting Text and Images



Radford and Kim et al. “Learning Transferrable Visual Models from Natural Language Supervision”, ICML’21.

Kakaobrain © All rights Reserved.

CLIP: Connecting Text and Images



Radford and Kim et al. “Learning Transferrable Visual Models from Natural Language Supervision”, ICML’21.

Kakaobrain © All rights Reserved.
Kakaobrain

Kakaobrain © All rights Reserved.

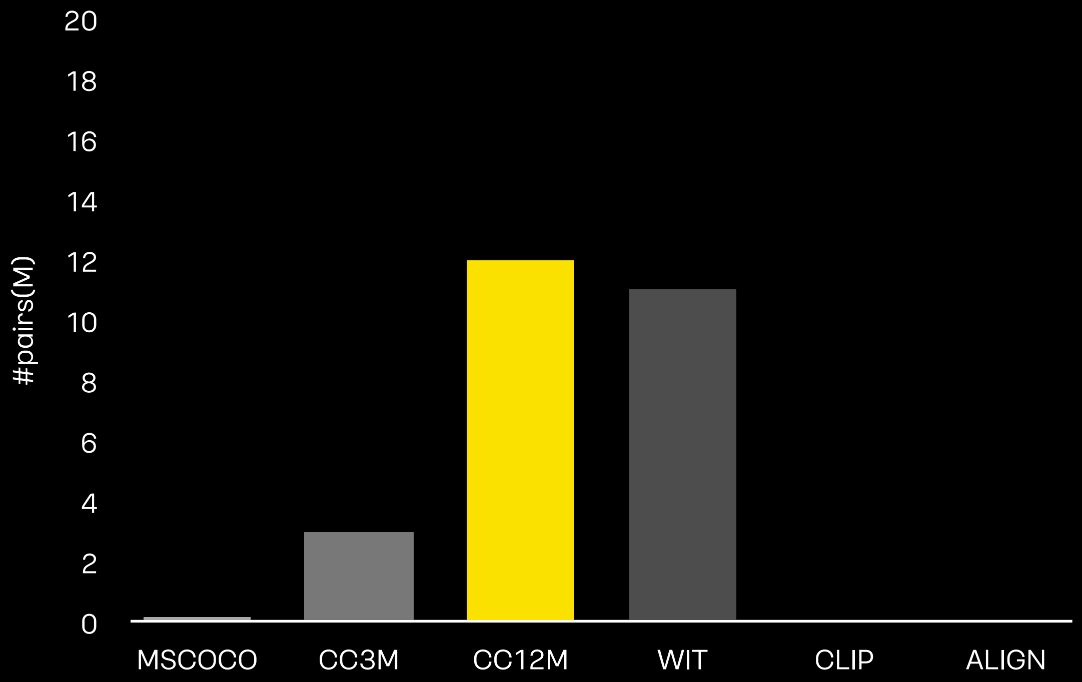
Large-scale Image-Text Pairs

MSCOCO sample

“A shoe rack with some shoes and a dog sleeping on them”

Kakaobrain © All rights Reserved.

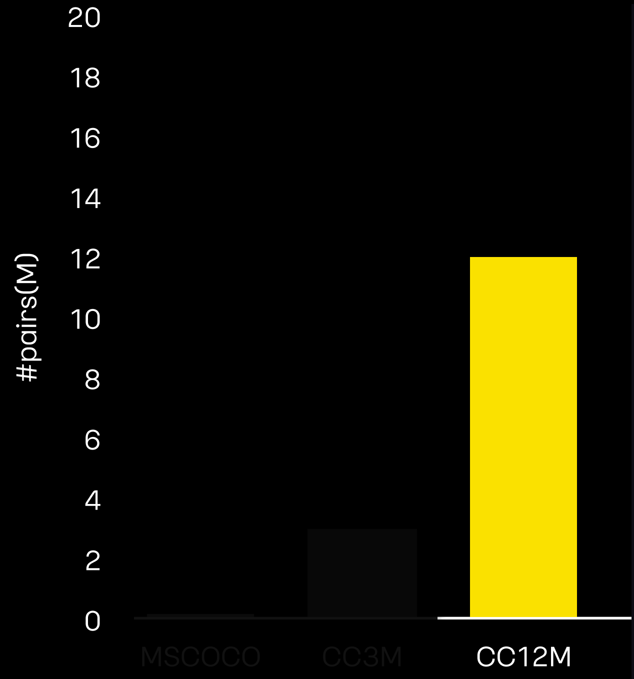
Large-scale Image-Text Pairs

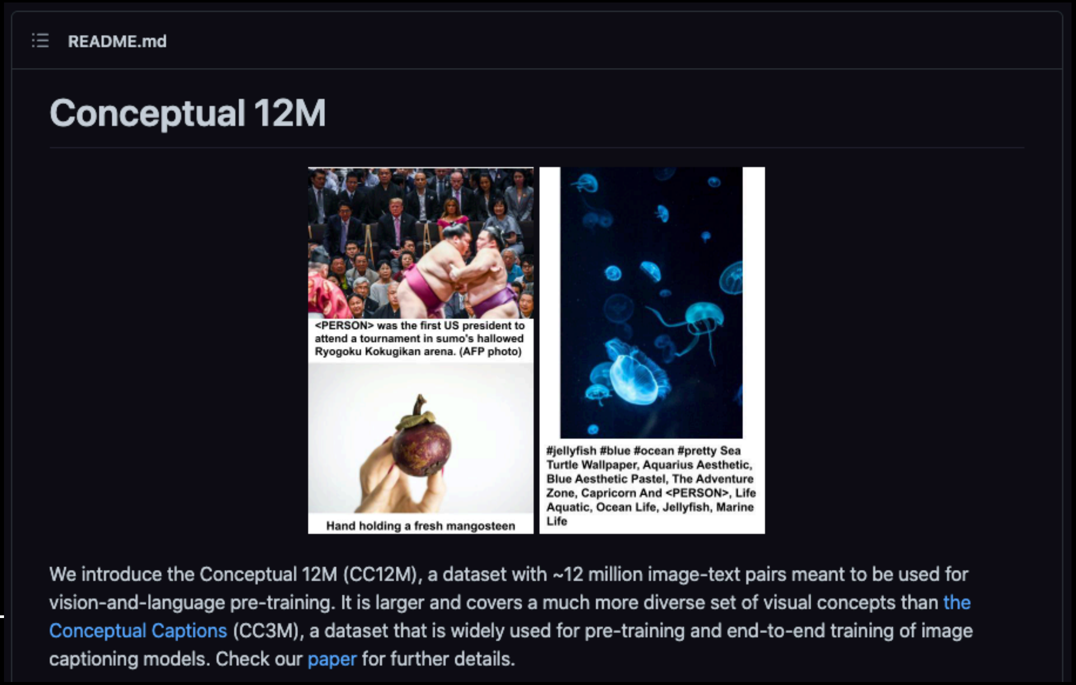


Kakaobrain © All rights Reserved.

Large-scale Image-Text Pairs

https://github.com/google-research-datasets/conceptual-12m (CVPR'21)



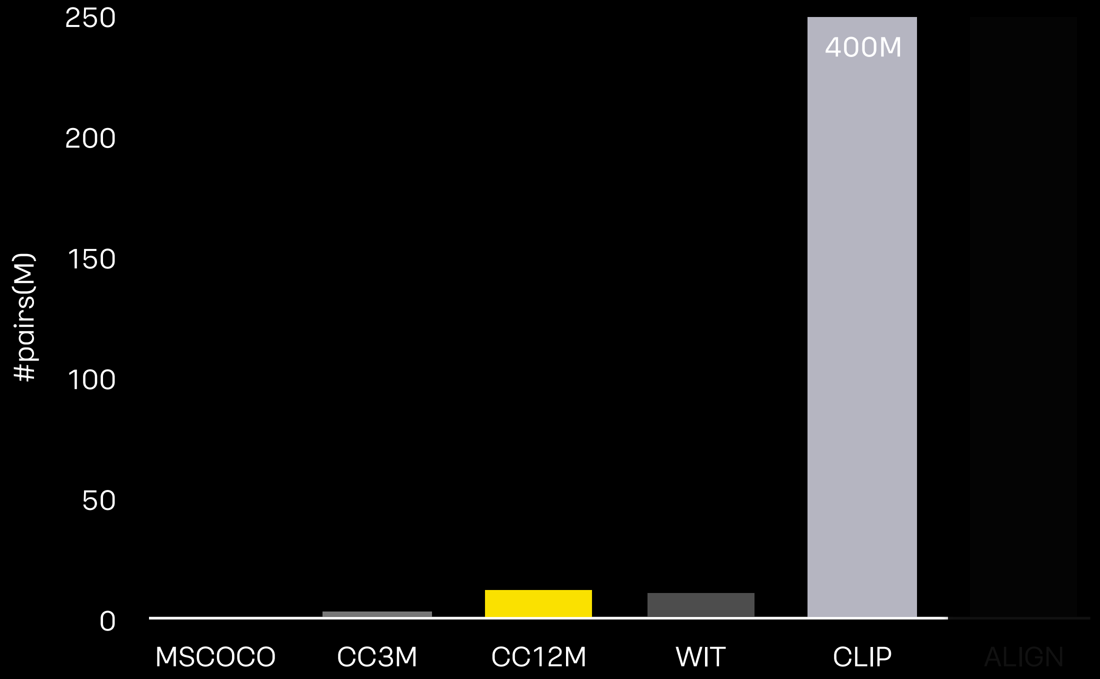

README.md

Conceptual 12M

We introduce the Conceptual 12M (CC12M), a dataset with ~12 million image-text pairs meant to be used for vision-and-language pre-training. It is larger and covers a much more diverse set of visual concepts than the Conceptual Captions (CC3M), a dataset that is widely used for pre-training and end-to-end training of image captioning models. Check our paper for further details.

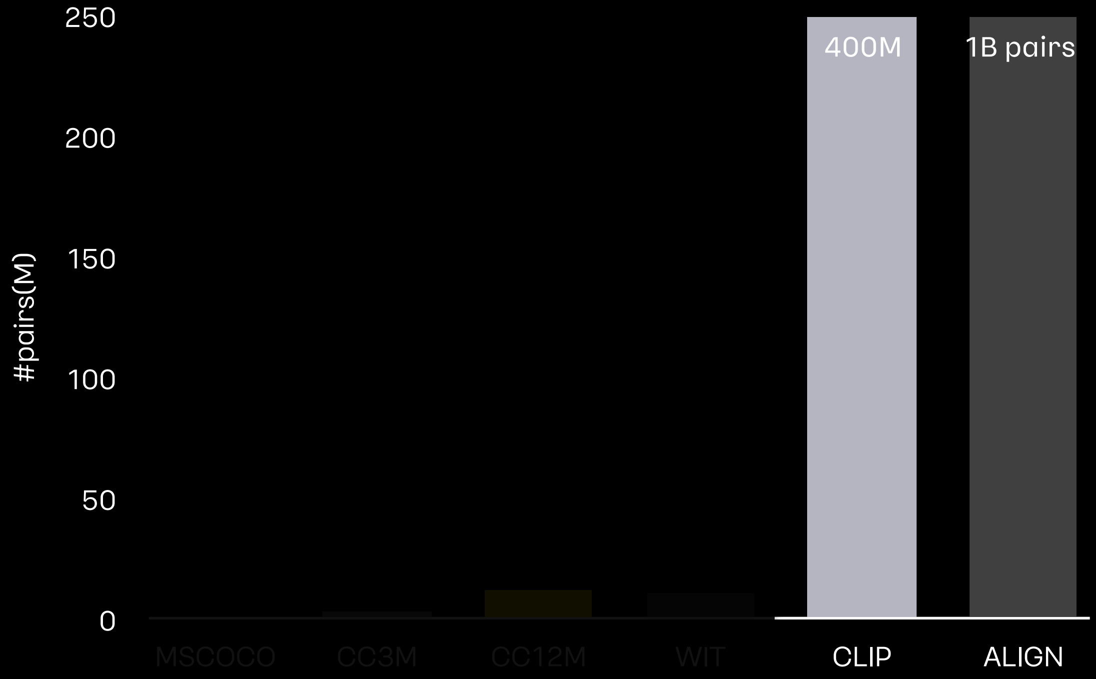
Kakaobrain © All rights Reserved.

Large-scale Image-Text Pairs



Kakaobrain © All rights Reserved.

Large-scale Image-Text Pairs



Google Research (ICML'21)

arXiv:2102.05918v2 [cs.CV] 11 Jun 2021

Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision

Chao Jia[1] Yinfei Yang[1] Ye Xia[1] Yi-Ting Chen[1] Zarana Parekh[1] Hieu Pham[1] Quoc V. Le[1] Yunhsuan Sung[1] Zhen Li[1] Tom Duerig[1]

Abstract

Pre-trained representations are becoming crucial for many NLP and perception tasks. While representation learning in NLP has transitioned to training on raw text without human annotations, visual and vision-language representations still rely heavily on curated training datasets that are expensive or require expert knowledge. For vision applications, representations are mostly learned using datasets with explicit class labels such as ImageNet or OpenImages. For vision-language, popular datasets like Conceptual Captions, MSCOCO, or CLIP all involve a non-trivial data collection (and cleaning) process. This costly curation process limits the size of datasets and hence hinders the scaling of trained models. In this paper, we leverage a noisy dataset of over one billion image alt-text pairs, obtained without expensive filtering or post-processing steps in the Conceptual Captions dataset. A simple dual-encoder architecture learns to align visual and language representations of the image and text pairs using a contrastive loss. We show that the scale of our corpus can make up for its noise and leads to state-of-the-art representations even with such a simple learning scheme. Our visual representation achieves strong performance when transferred to classification tasks such as ImageNet and VTAB. The aligned visual and language representations enables zero-shot image classification and also set new state-of-the-art results on Flickr30K and MSCOCO image-text retrieval benchmarks, even when compared with more sophisticated cross-attention models. The representations also enable cross-modality search with complex text and text + image queries.

1. Introduction

In the existing literature, visual and vision-language representation learning are mostly studied separately with different training data sources. In the vision domain, pre-training on large-scale supervised data such as ImageNet (Deng et al., 2009), OpenImages (Kuznetsova et al., 2020), and JFT-300M (Sun et al., 2017; Kolesnikov et al., 2020) has proven to be critical for improving performance on downstream tasks via transfer learning. Curation of such pre-training datasets requires heavy work on data gathering, sampling, and human annotation, and hence is difficult to scale.

Pre-training has also become the de-facto approach in vision-language modeling (Lu et al., 2019; Chen et al., 2020c; Li et al., 2020). However, vision-language pre-training datasets such as Conceptual Captions (Sharma et al., 2018), Visual Genome Dense Captions (Krishna et al., 2016), and ImageBERT (Qi et al., 2020) require even heavier work on human annotation, semantic parsing, cleaning and balancing. As a result, the scales of these datasets are only in the realm of ~10M examples. This is at least an order of magnitude smaller than their counterparts in the vision domain, and much smaller than large corpora of text from the internet for NLP pre-training (e.g., Devlin et al. (2019); Radford et al. (2019); Yang et al. (2019); Liu et al. (2019b); Raffel et al. (2020)).

In this work, we leverage a dataset of over one billion noisy image alt-text pairs to scale visual and vision-language representation learning. We follow the procedures described in the Conceptual Captions dataset (Sharma et al., 2018) to have a large noisy dataset. But instead of applying the complex filtering and post-processing steps as proposed by (Sharma et al., 2018) to clean the dataset, we only apply simple frequency-based filtering. The resulting dataset is noisy, but is two orders of magnitude larger than the Conceptual Captions dataset. We show that visual and vision-language representations pre-trained on our exascale dataset achieve very strong performance on a wide range of tasks.

To train our model, we use an objective that aligns the visual and language representations in a shared latent embedding space using a simple dual-encoder architecture. Similar

[1] Google Research. Correspondence to: Chao Jia <chaojia@google.com>, Yinfei Yang <yinfeiy@google.com>.

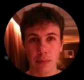
Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

Where the Source?

Kakaobrain © All rights Reserved.

Kakaobrain © All rights Reserved.

LAION Projects

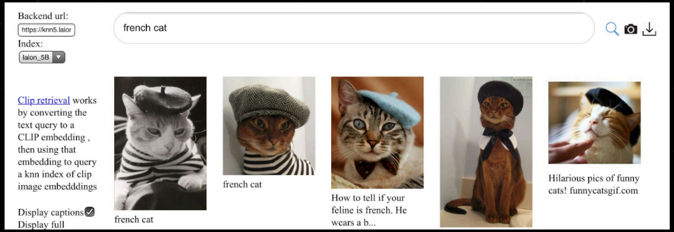


Romain Beaumont
31. March 2022
1 Comment
Uncategorized

LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI-MODAL DATASETS

We present a dataset of 5,85 billion CLIP-filtered image-text pairs, 14x bigger than LAION-400M, previously the biggest openly accessible image-text dataset in the world.

Authors: Christoph Schuhmann, Richard Vencu, Romain Beaumont, Theo Coombes, Cade Gordon, Aarush Katta, Robert Kaczmarczyk, Jenia Jitsev



Kakaobrain

Kakaobrain © All rights Reserved.

Experiments — Zero-shot Classification

+

A photo of {label}

A photo of a cat

A photo of a dog

⋮

Kakaobrain © All rights Reserved.

Experiments — Zero-shot Classification

+

A photo of {label}, a type of flower

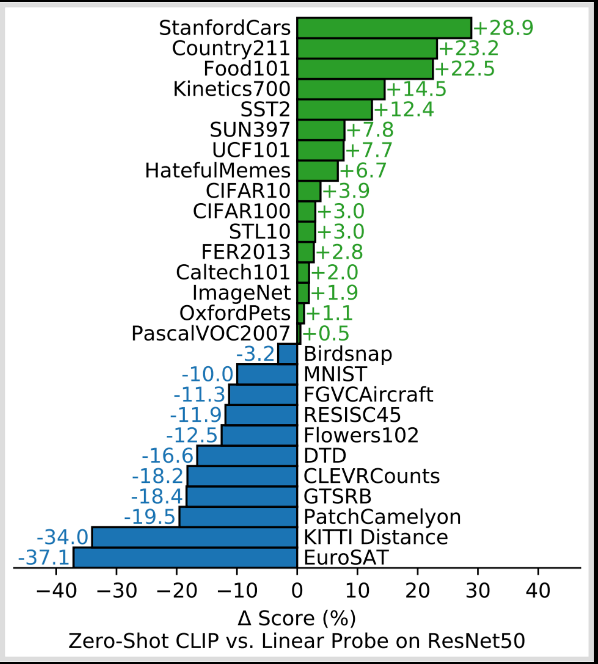
A photo of a cat

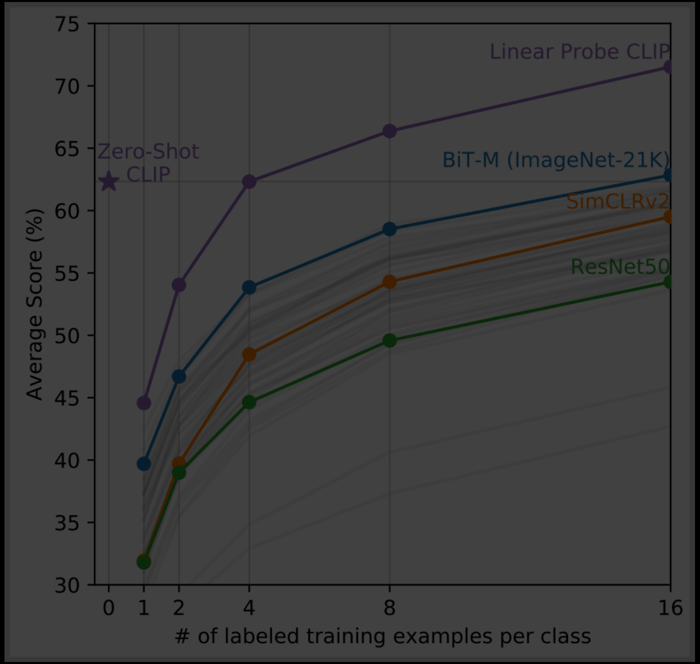
A photo of a dog

⋮

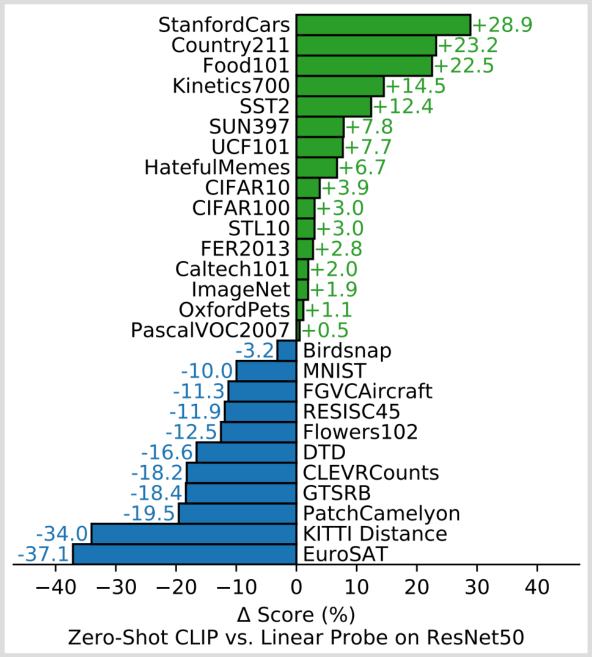
Kakaobrain © All rights Reserved.

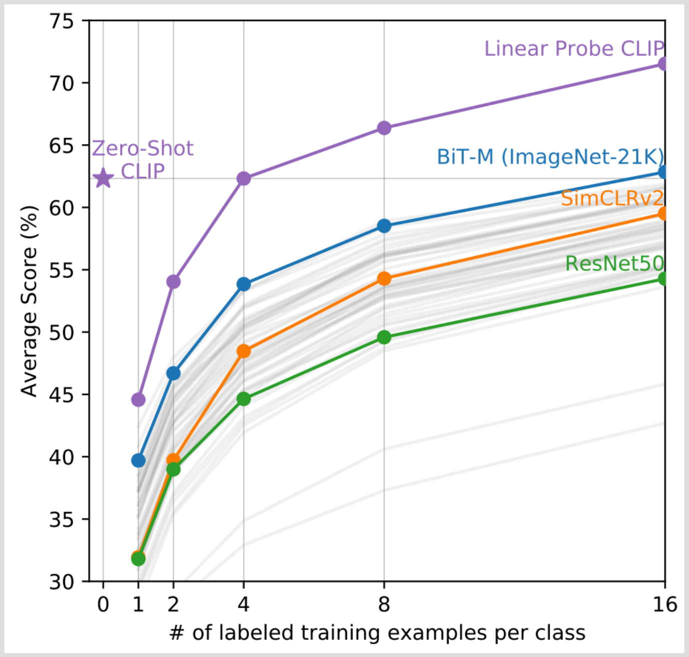
Experiments — Zero-shot Classification





Experiments — Zero-shot Classification

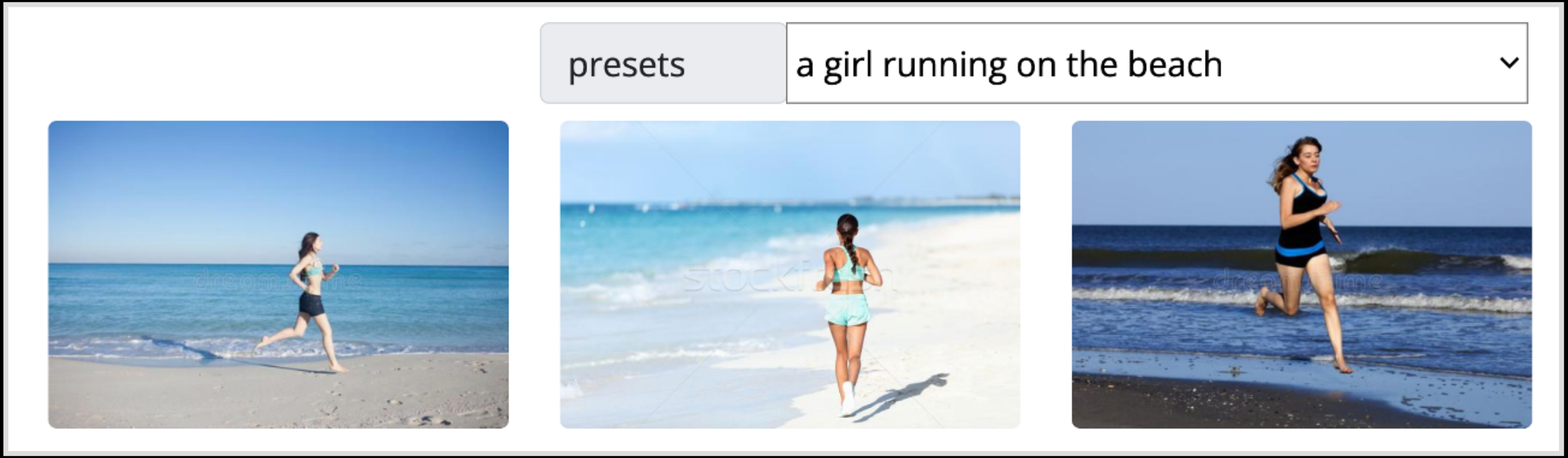




Kakaobrain © All rights Reserved.

Kakaobrain © All rights Reserved.

Application of CLIP — Search Engine



Kakaobrain © All rights Reserved.

Application of CLIP

"A drawing of a cat". "Horse eating a cupcake". "A 3D rendering of a temple". "Family vacation to Walt Disney World". "Self".

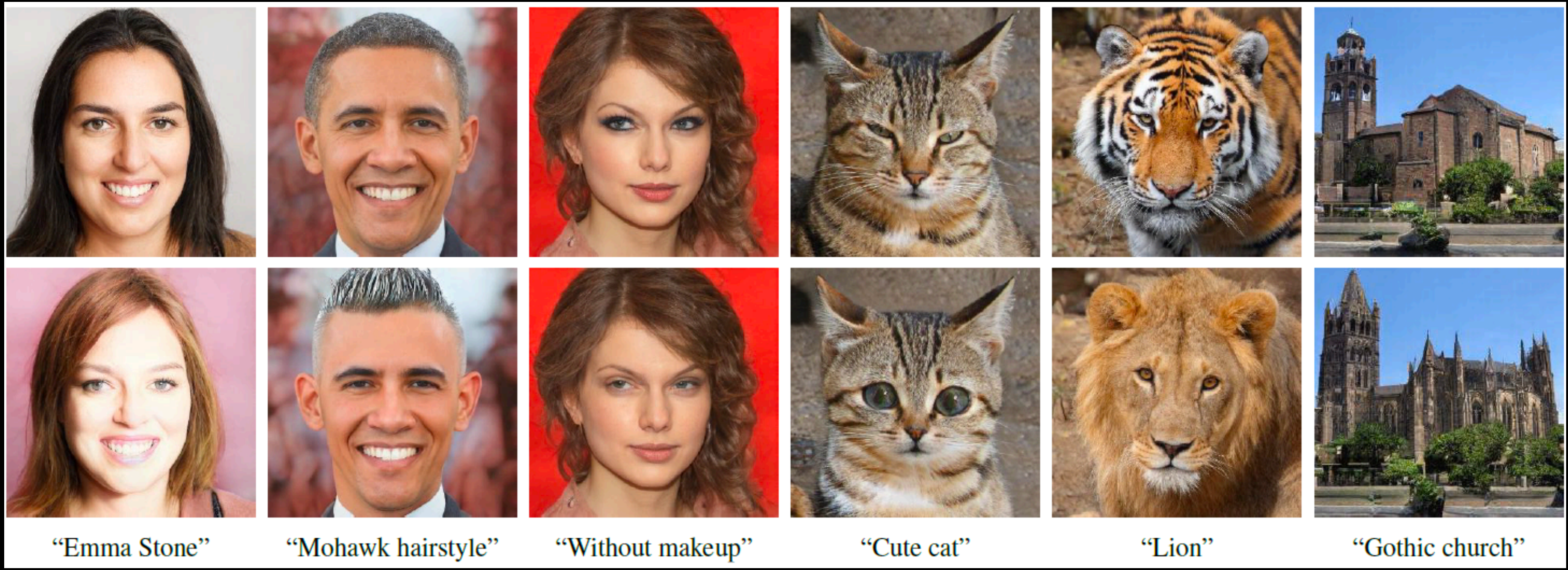
Various drawings synthesized by CLIPDraw, along with the corresponding description prompts used. CLIPDraw synthesizes images from text by performing gradient descent over a set of RGBA Bézier curves, with the goal of minimizing cosine distance between the CLIP encodings of generated images and description prompts. CLIPDraw does not require learning a new model, and can generally synthesize images within a minute on a typical GPU.

Kakaobrain

Frans et al. "CLIPDraw: Exploring Text-to-Drawing Synthesis through Language-Image Encoders", arXiv'21

Kakaobrain © All rights Reserved.

Application of CLIP



Kakaobrain

Patashnik et al. "StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery", ICCV'21 (oral)

Kakaobrain © All rights Reserved.

Conclusion

Now, it’s possible to learn a shared representation from text-image pairs

Kakaobrain