Representation Learning: How Should Feature Be Learned in Vision?

Introducing *CLIP* and an *Unsupervised Semantic Segmentation* Approach

Presenter: Sungguk Cha

MINDs Lab

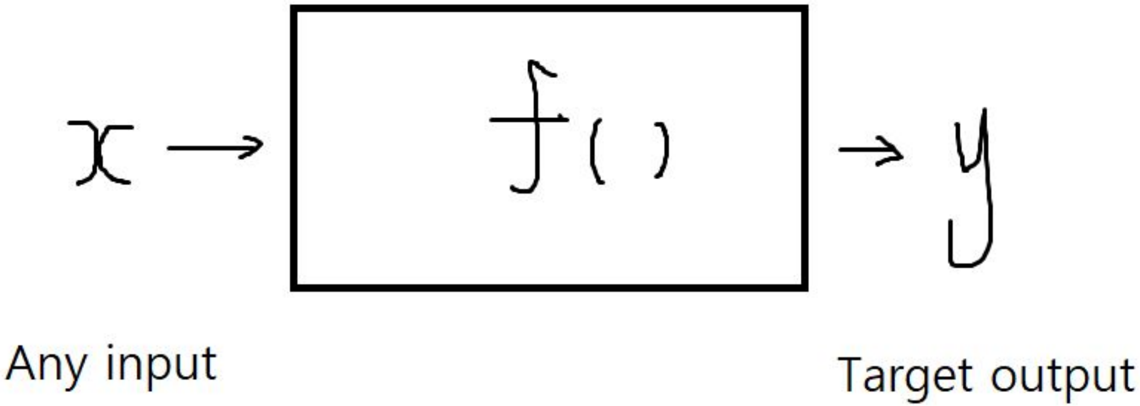
Contents

- Representation Learning?
- (Review) CLIP: Learning Transferable Visual Models From Natural Language Supervision
- (Review) Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals
- Conclusion

MINDs Lab

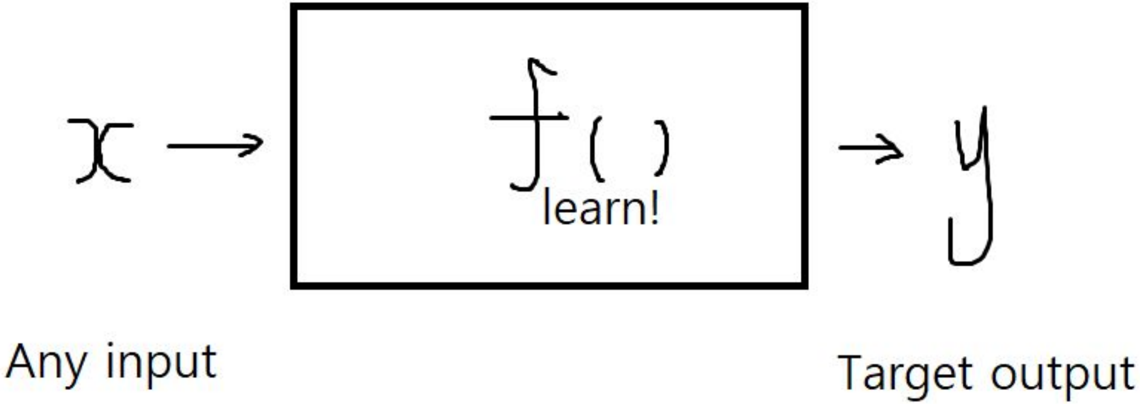
Representation Learning?

Function



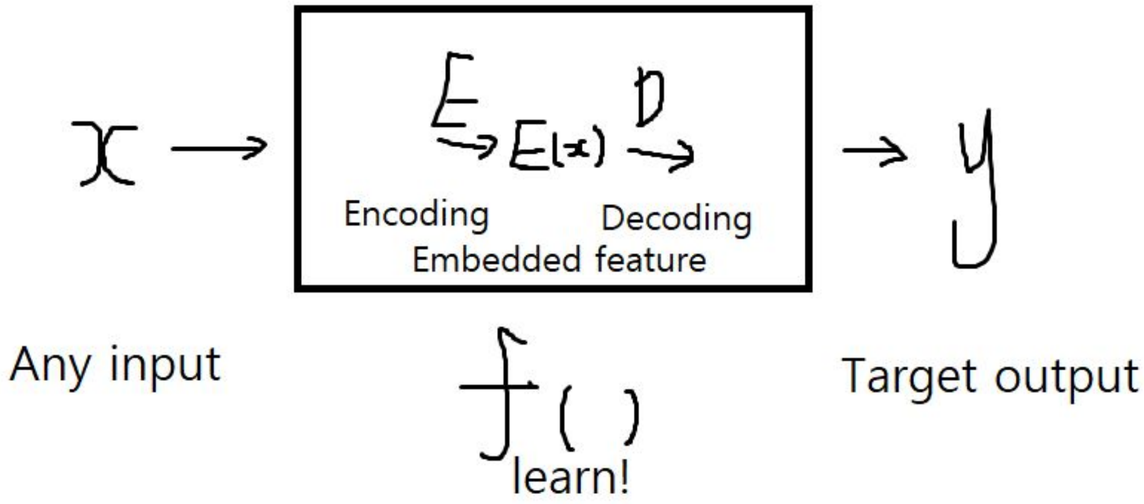
Representation Learning?

Function that learns == Machine Learning



Representation Learning?

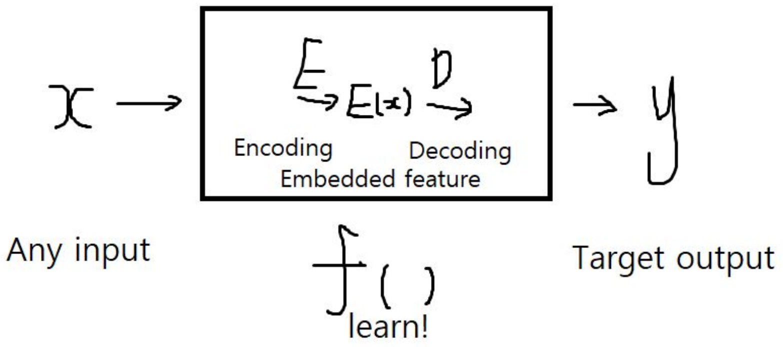
Machine Learning



Representation Learning?

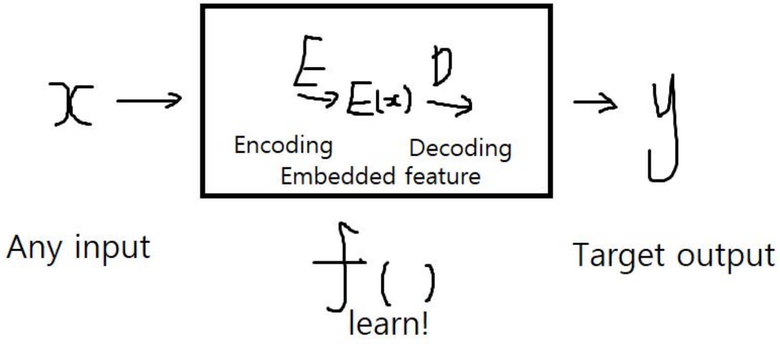
Machine Learning

We need stronger representation (feature).



Representation Learning?

Machine Learning



We need stronger representation (feature).

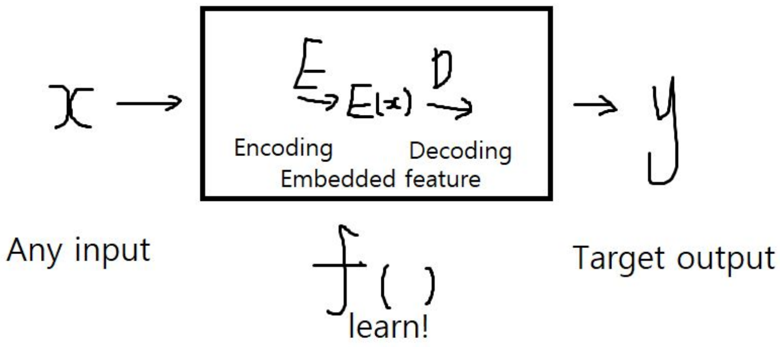
We cannot use

- raw digital signal in audio,
- raw text in NLP,
- raw image in vision

to solve a problem.

Representation Learning?

Machine Learning



We need stronger representation (feature).

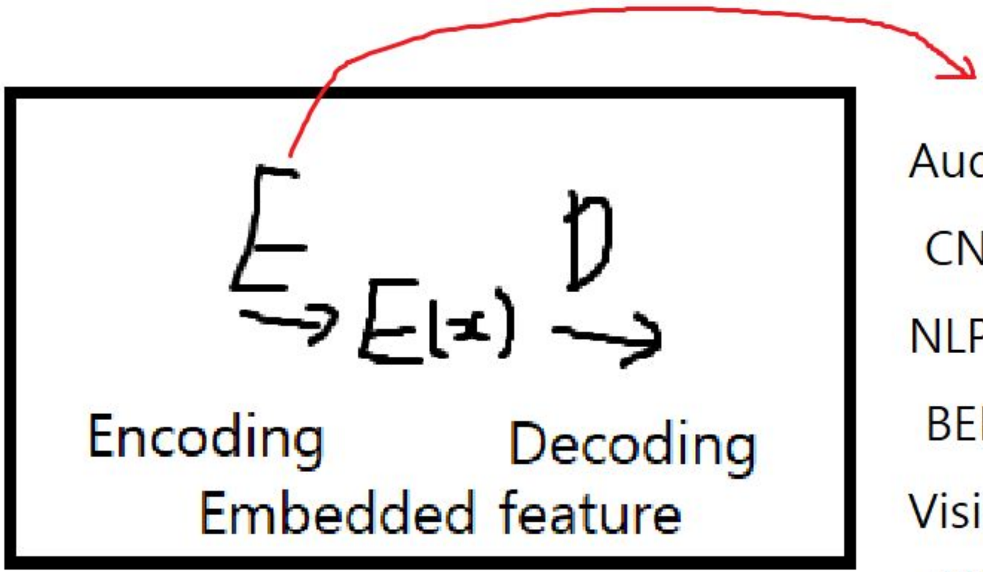
We cannot use
- raw digital signal in audio,
- raw text in NLP,
- raw image in vision

to solve a problem.

So, we encode the input to have stronger representation.

MINDs Lab

Representation Learning?



Lately

Audio:

CNN, ...

NLP:

BERT, GPT, ...

Vision:

CNN, ...

How the Encoder Learns in Vision?

Supervised Learning

- e.g., predictive learning (image classification)

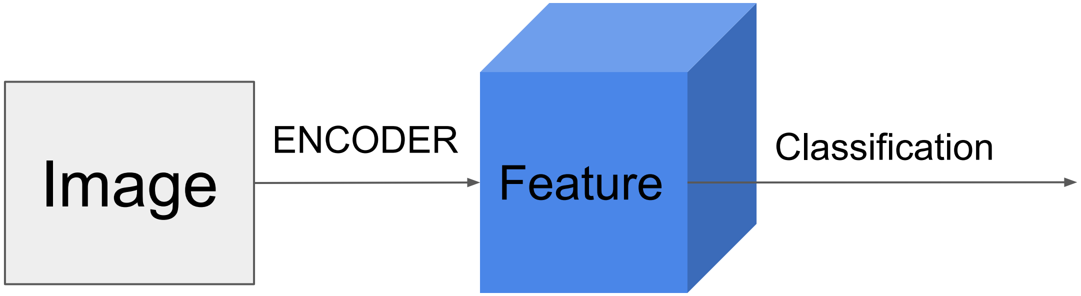
Self-supervised Learning

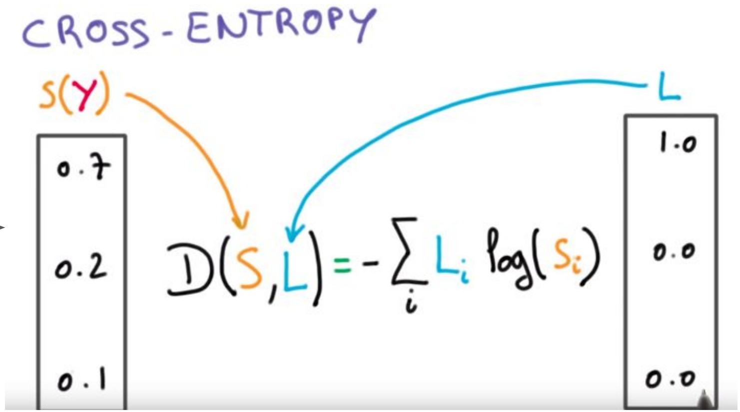
- e.g., contrastive learning

MINDs Lab

Supervised Learning

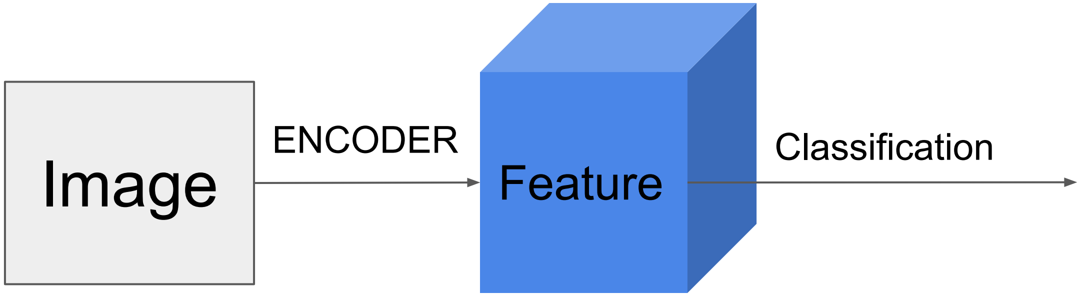
example: predictive learning (image classification)

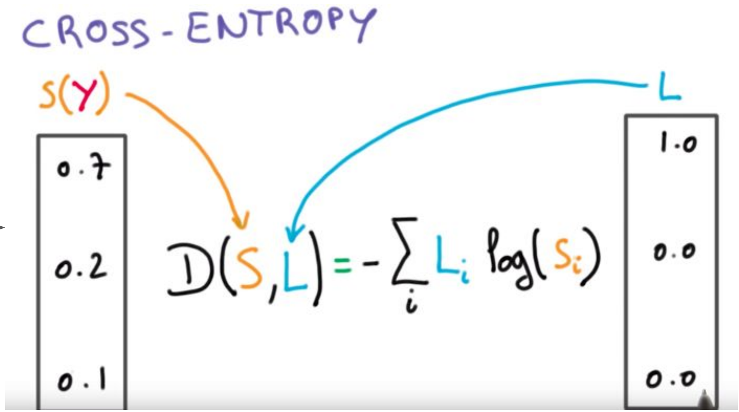




Cross entropy image source:
https://medium.com/data-science-bootcamp/understand-cross-entropy-loss-in-minutes-9fb263caee9a

Supervised Learning
example: predictive learning (image classification)





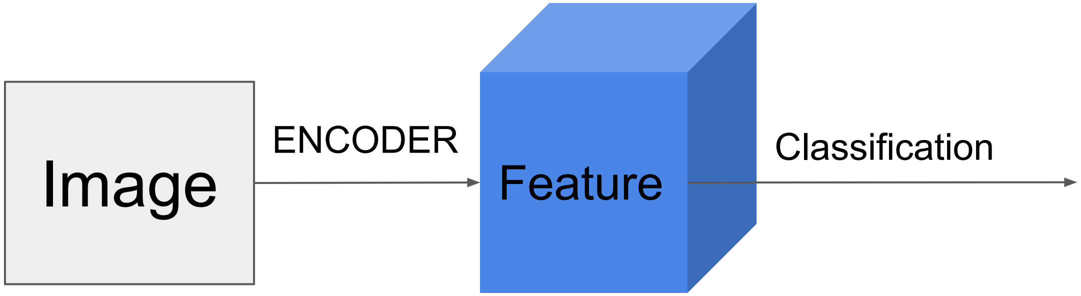
How encoder learns?

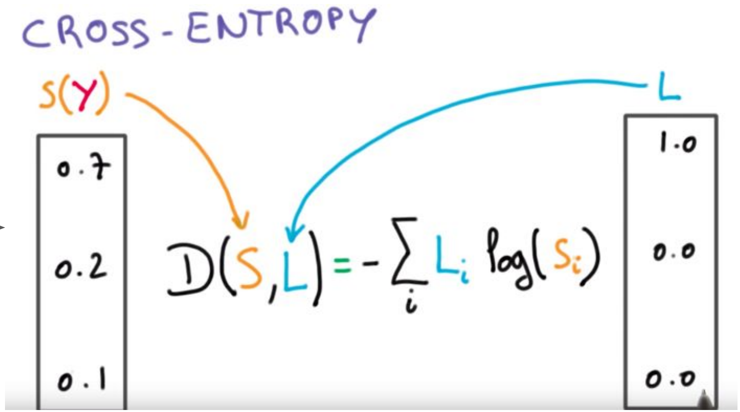
Cross entropy image source:
https://medium.com/data-science-bootcamp/understand-cross-entropy-loss-in-minutes-9fb263caee9a

MINDs Lab

Supervised Learning

example: predictive learning (image classification)





***Anyway*, the king-god-backpropagation will teach the encoder.**

MINDs Lab

Cross entropy image source:
https://medium.com/data-science-bootcamp/understand-cross-entropy-loss-in-minutes-9fb263caee9a

Self-Supervised Learning
example: contrastive learning

Intuition:
Same category, **similarly** encoded.
Different category, **differently** encoded.

MINDs Lab

Self-Supervised Learning

example: contrastive learning

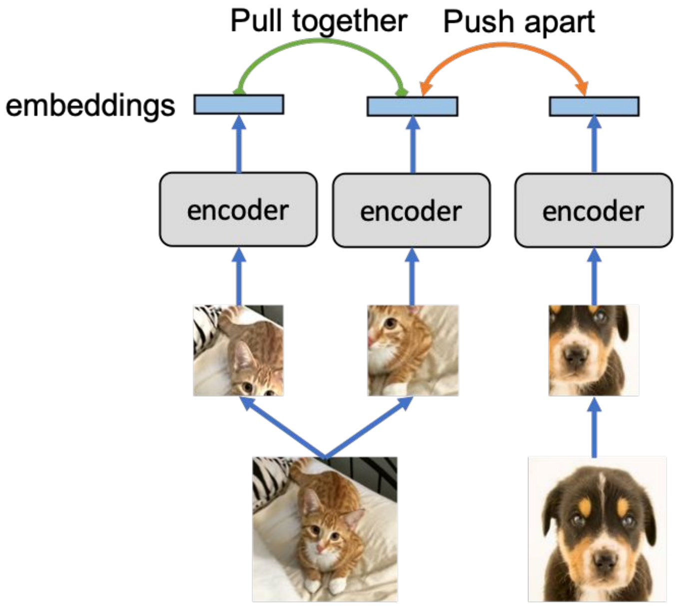


Figure 1 (a)

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

Self-Supervised Learning
example: contrastive learning

Positive Pair

$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i / \tau)}$$

NLL with softmax

K Negative pairs and
one Positive pair

Representation Learning

is a research field where

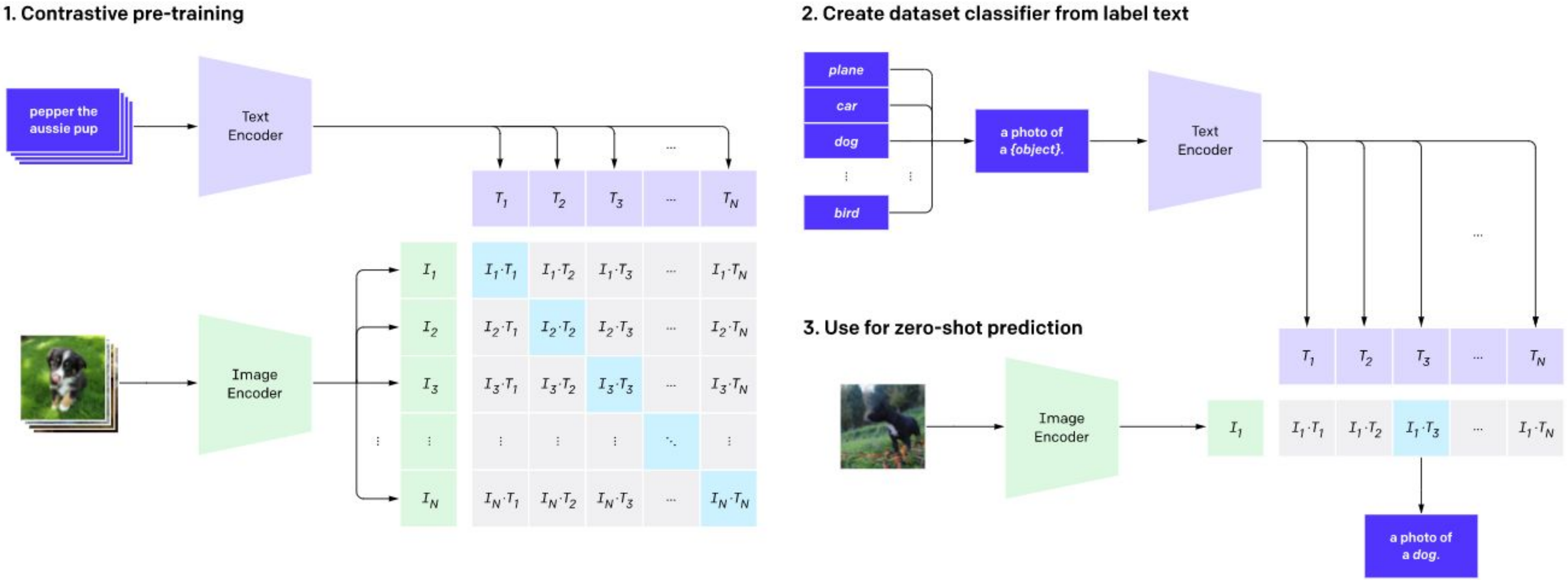
*"how to give some **prior** to representation"*

is researched.

Contrastive learning is one example.

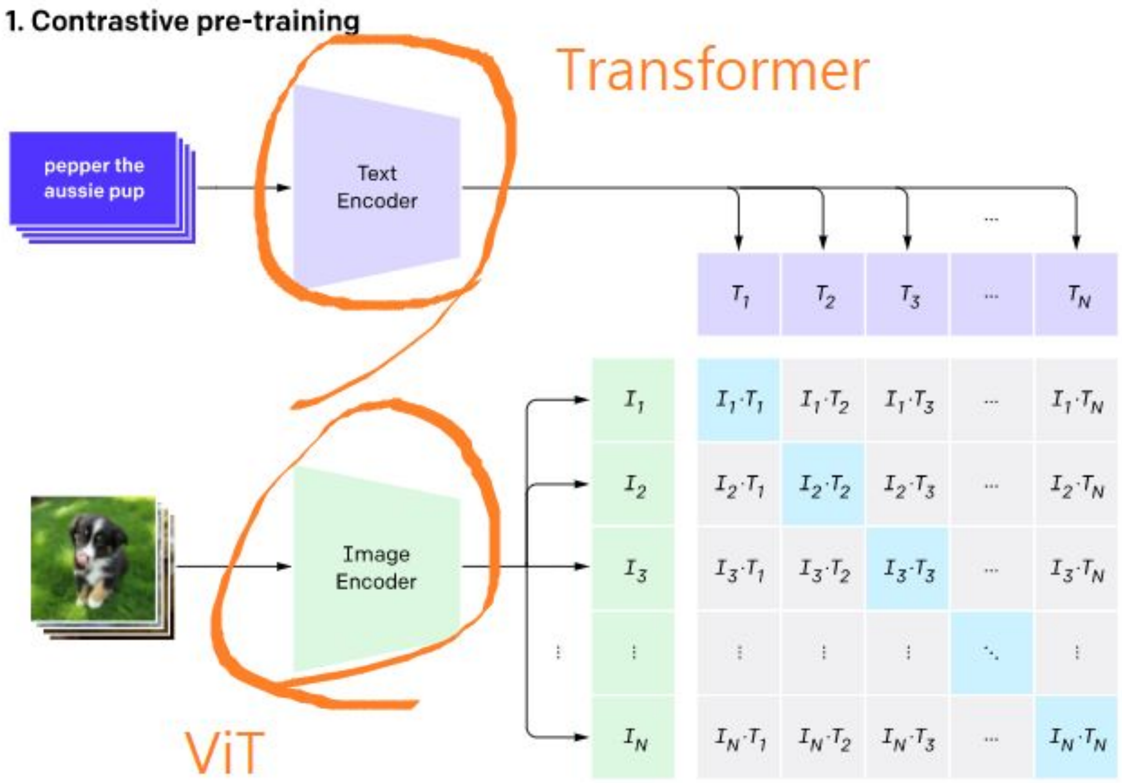
(Review)
CLIP: Connecting Text and Image
*Learning Transferable Visual Models From **Natural Language Supervision***

(Review)
CLIP: Connecting Text and Image
*Learning Transferable Visual Models From **Natural Language Supervision***



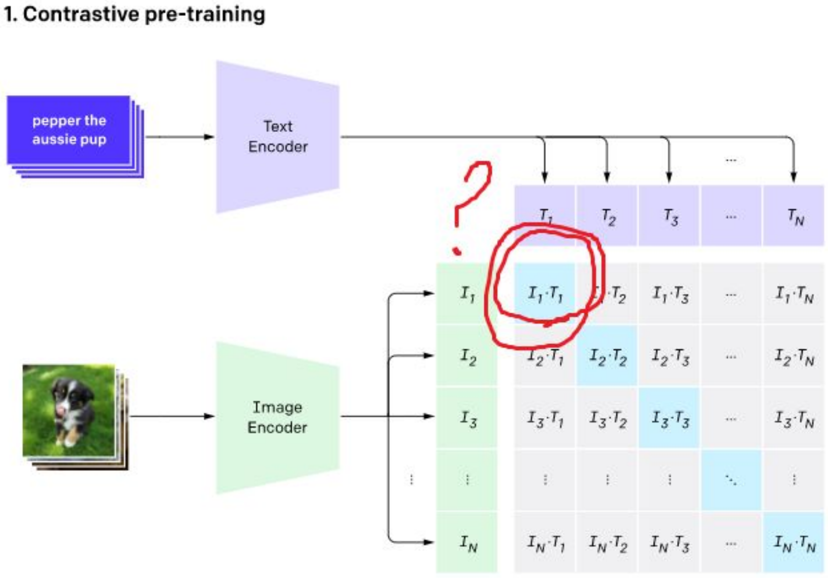
MINDs Lab

CLIP



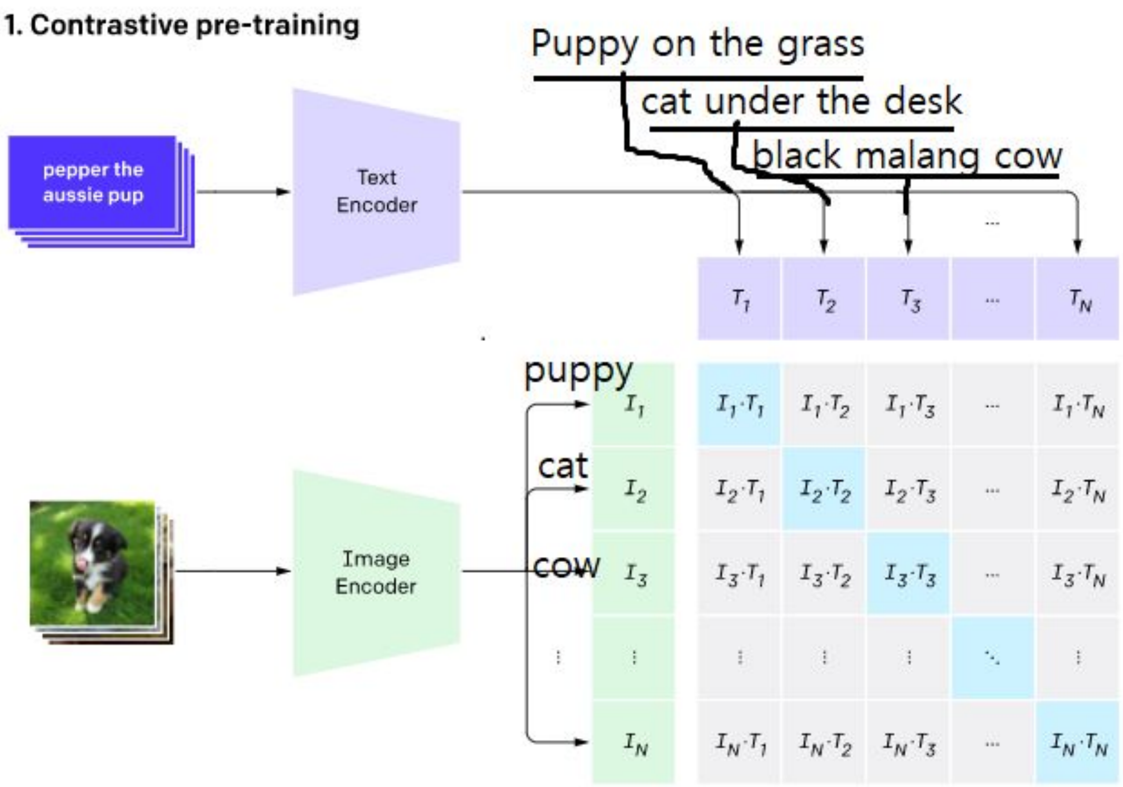
MINDs Lab

CLIP



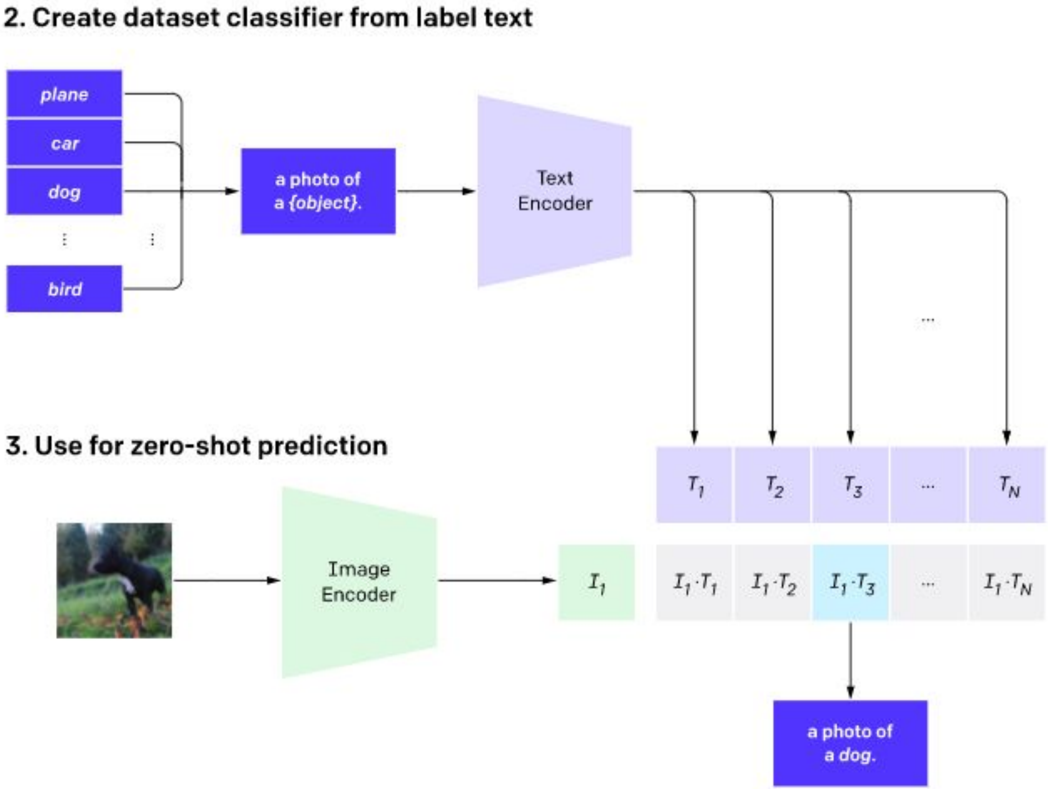

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned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

CLIP



MINDs Lab

CLIP as a Zero-shot Classifier



MINDs Lab

CLIP

- representation learning with **natural language supervision**
- multi-modal learning
- zero-shot capability
- ...

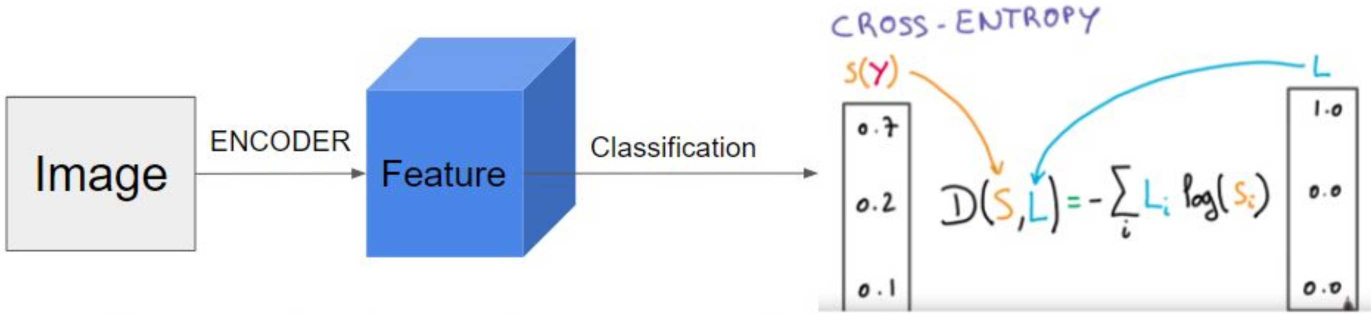
(Review)
Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals

MINDs Lab

(Review)
Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals

Motivation:

Representation in <u>anyway backpropagation learning</u> is weird.
(아무튼 역전파 학습법)



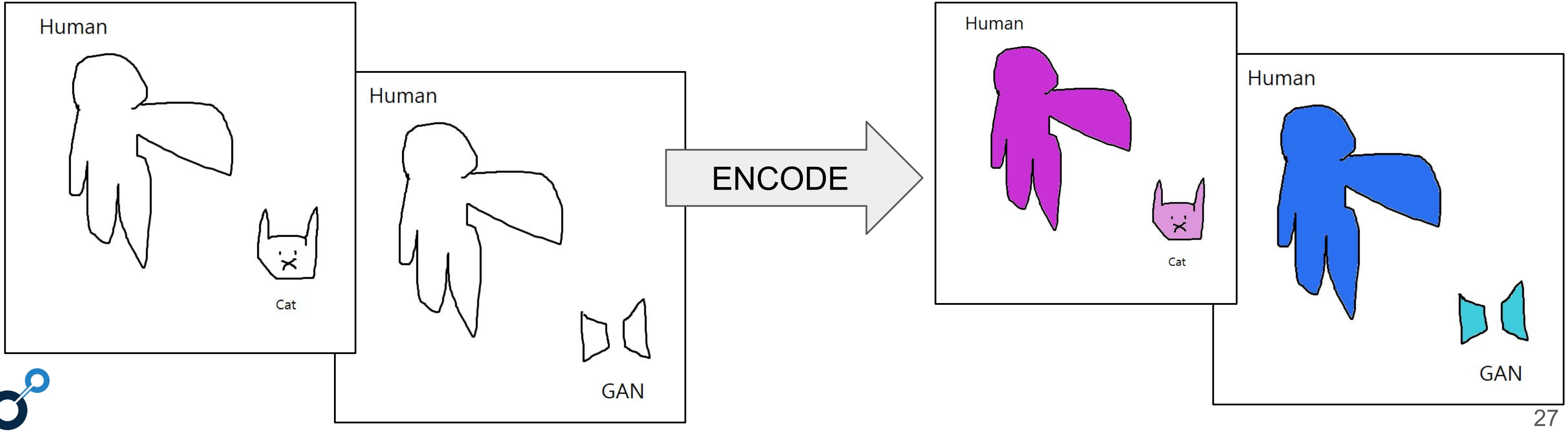
***Anyway*, the king-god-backpropagation will teach the encoder.**

MINDs Lab

(Review)
Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals

Motivation:

Representation is **affected by co-occur objects**.

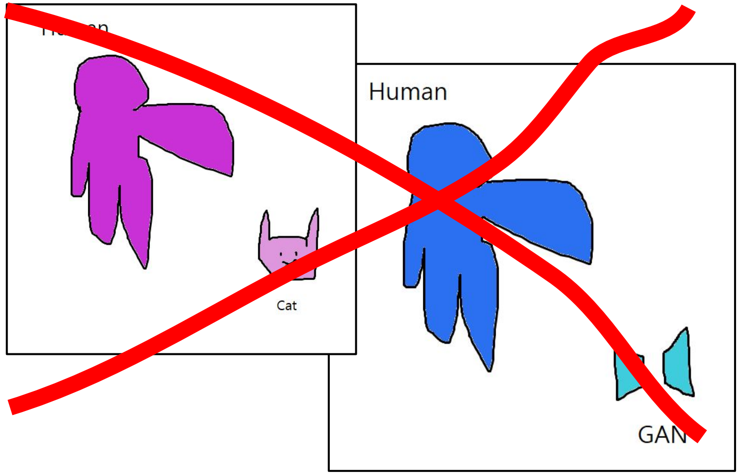


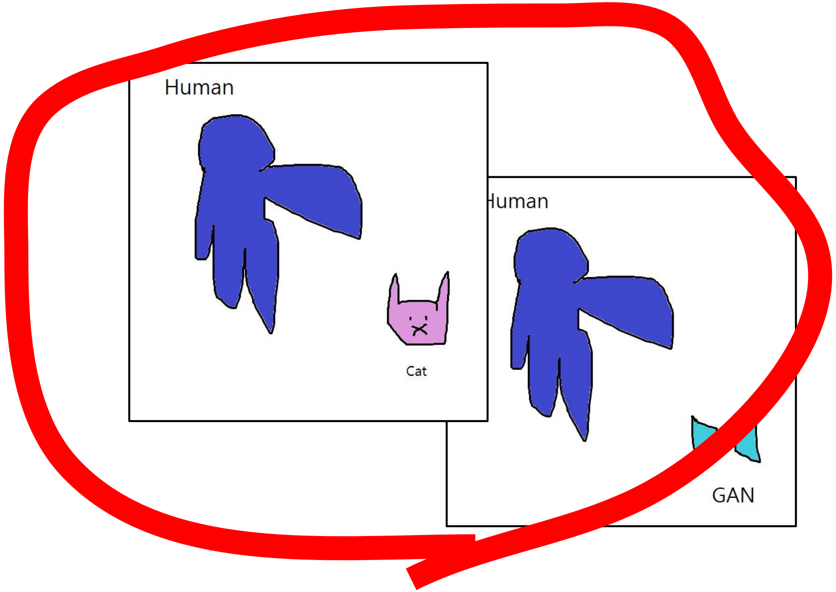
MINDs Lab

(Review)
Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals

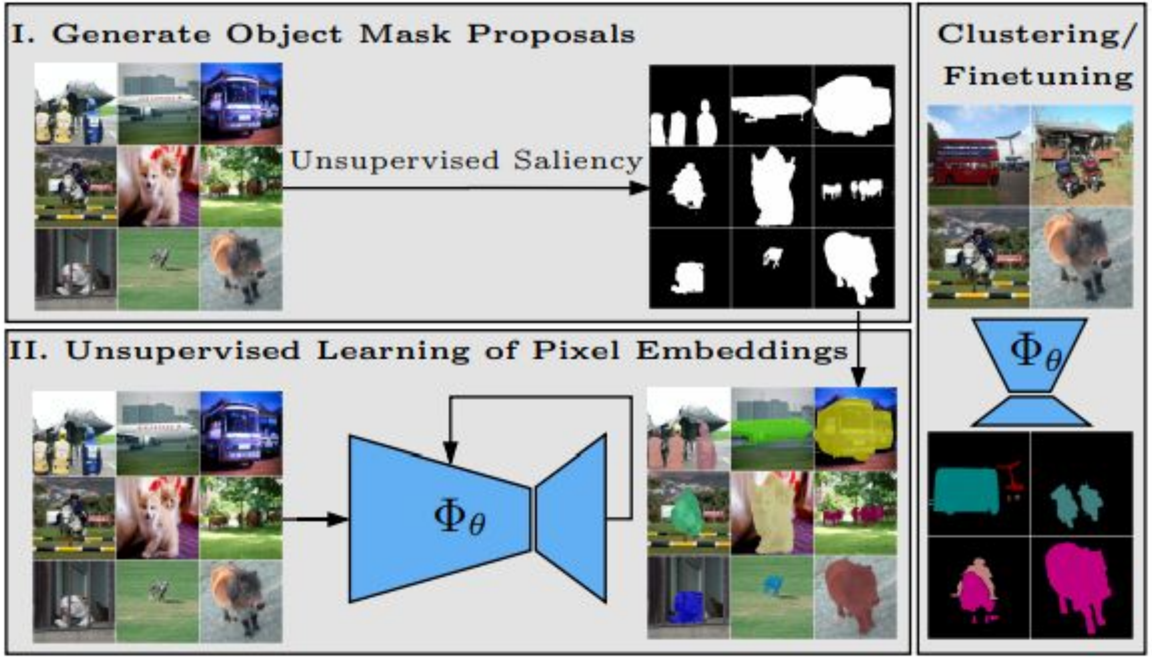
Intuition:

We want pixel representations of the same category to be similar.





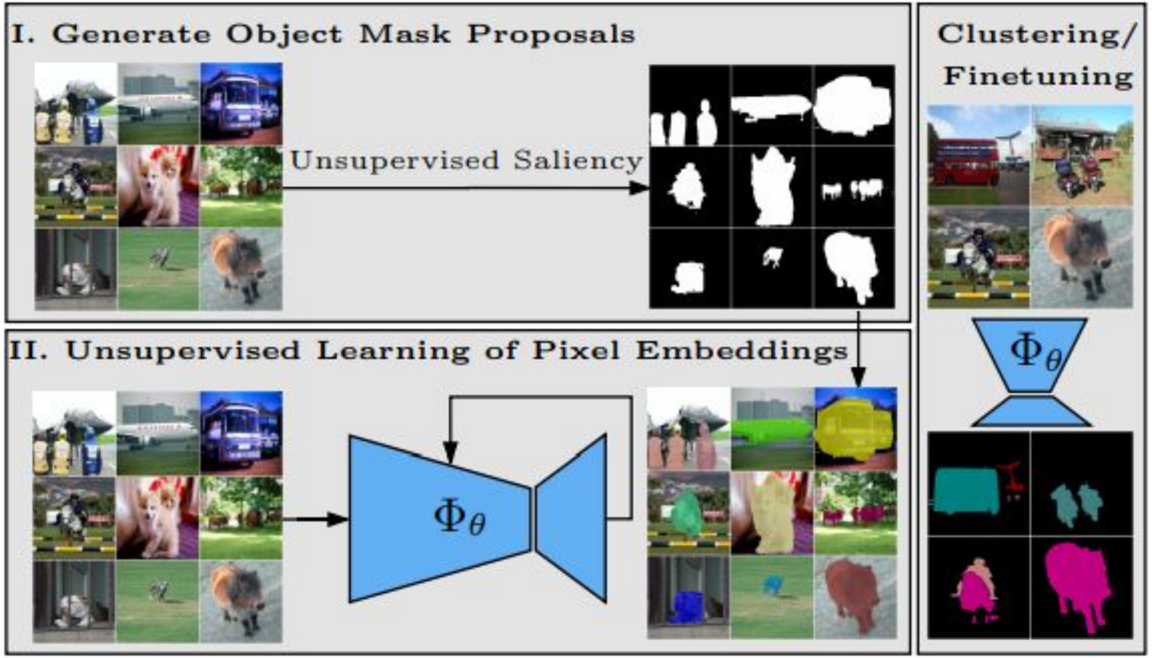
Approach



MINDs Lab

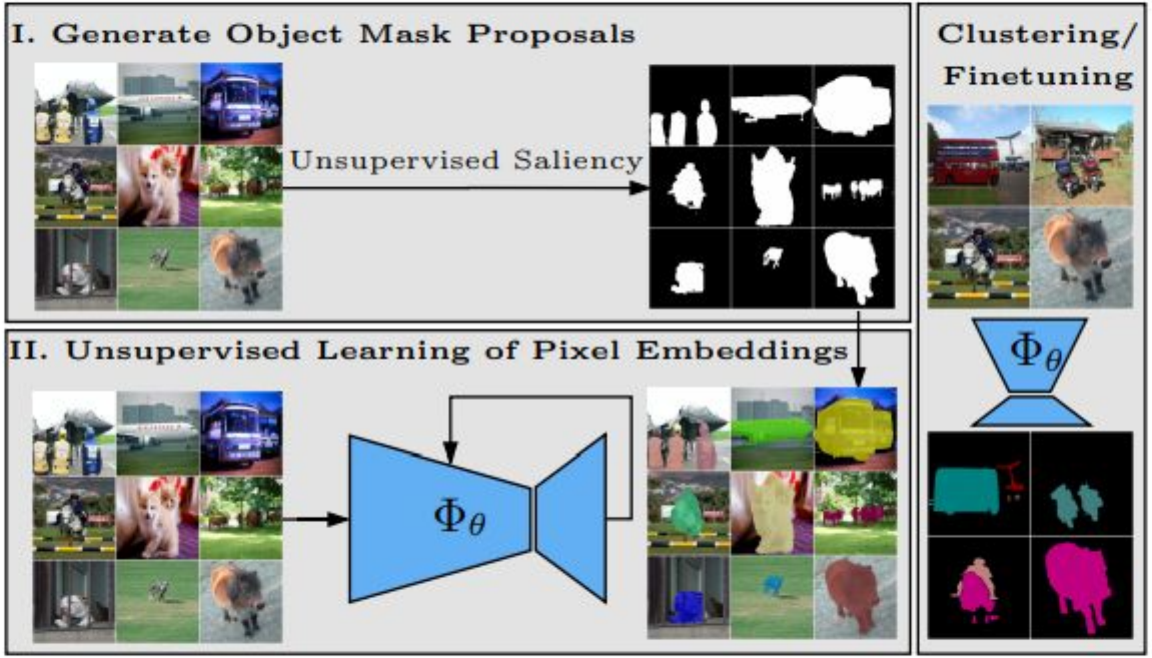
Approach

Assuming every image contains **only one category**,
using **saliency detection model** it generates mask proposal.



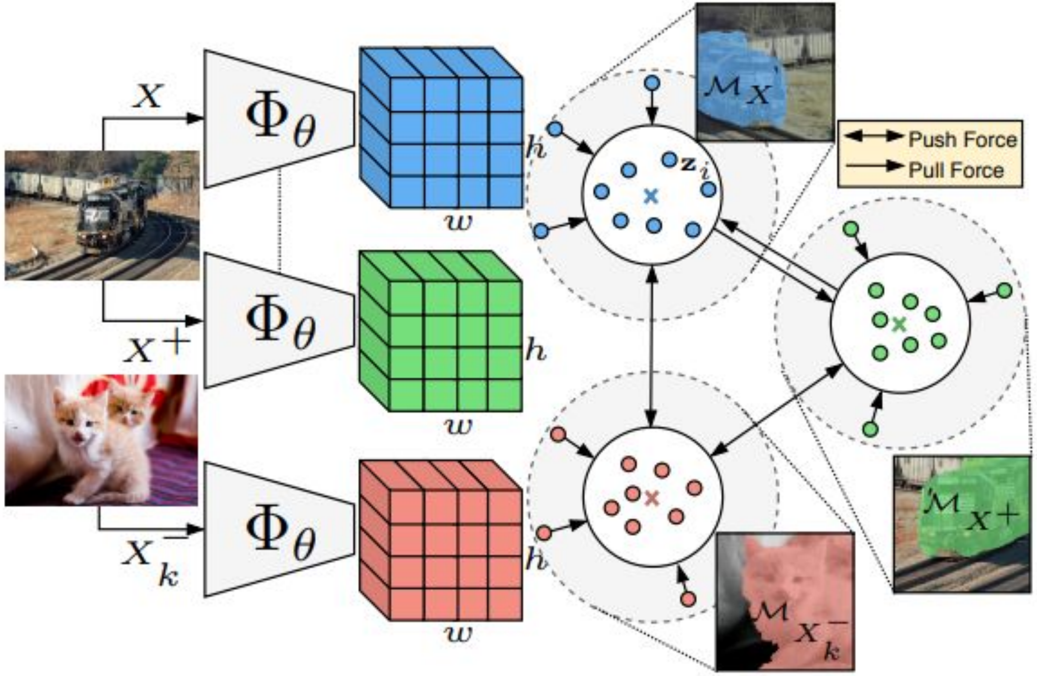
MINDs Lab

Approach



Pixel-level contrastive learning using the proposed mask

(Review)
Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals



(Review)
Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals

$$\mathbf{z}_{\mathcal{M}_n} = \frac{1}{|\mathcal{M}_n|} \sum_{i \in \mathcal{M}_n} \mathbf{z}_i.$$

E.g.) train representation :=
mean representation of the train pixels

for a pixel $i \in \mathcal{M}_X$

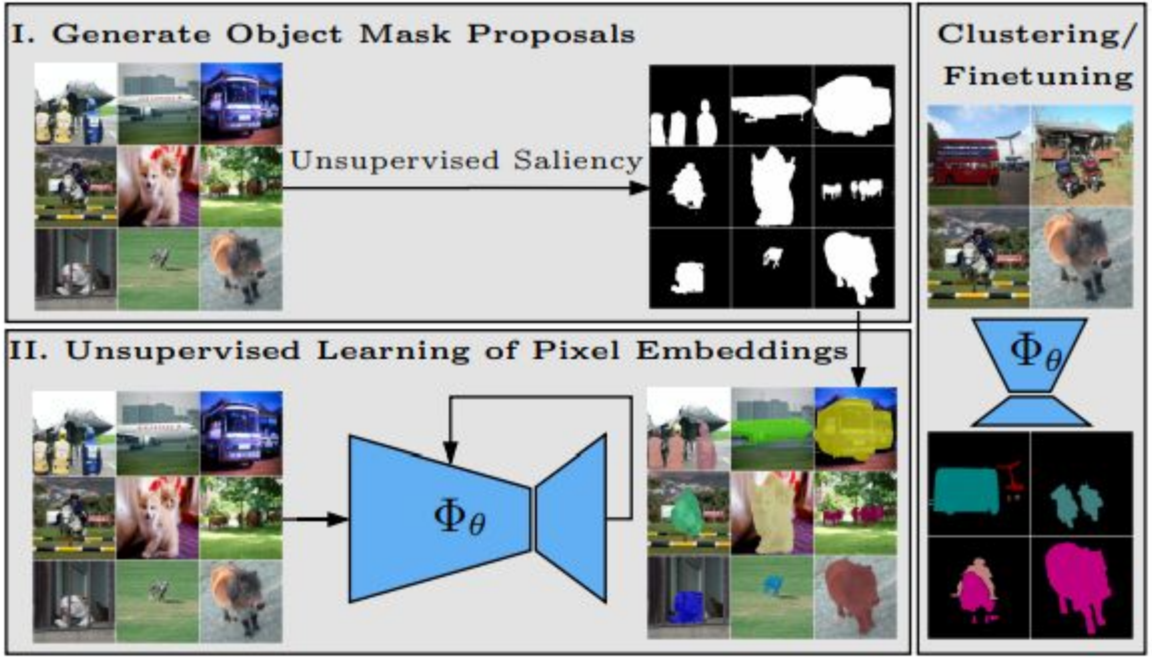
$$\mathcal{L}_i = -\log \frac{\exp\left(\mathbf{z}_i \cdot \mathbf{z}_{\mathcal{M}_{X^+}} / \tau\right)}{\sum_{k=0}^{K} \exp\left(\mathbf{z}_i \cdot \mathbf{z}_{\mathcal{M}_{X_k^-}} / \tau\right)}$$

Pixel-level contrastive learning!

MINDs Lab

Approach

Finetune in practice

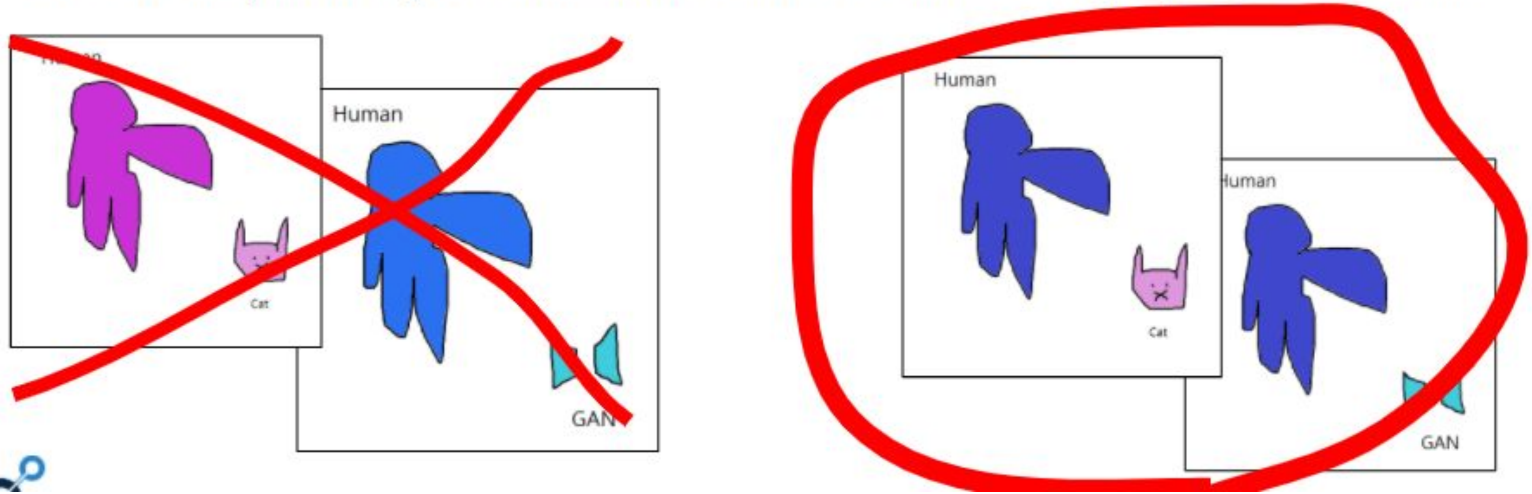


Comment

Isn’t it beautiful to think such representation learning?

Intuition:

We want pixel representations of the same category to be similar.



Conclusion

Thinking “how each feature should be” (**applying our prior knowledge**) is fun.

: the more data and parameter, the better.
human prior knowledge? No no.

Though some (above) stands against, I believe such trials are beautiful and worth.

MINDs Lab