김성수

[1]M. Caron *et al.*, *"Emerging properties in self-supervised vision transformers," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650-9660.*

[1] Facebook AI Research [2] Inria* [3] Sorbonne University

Emerging properties in self-supervised vision transformers

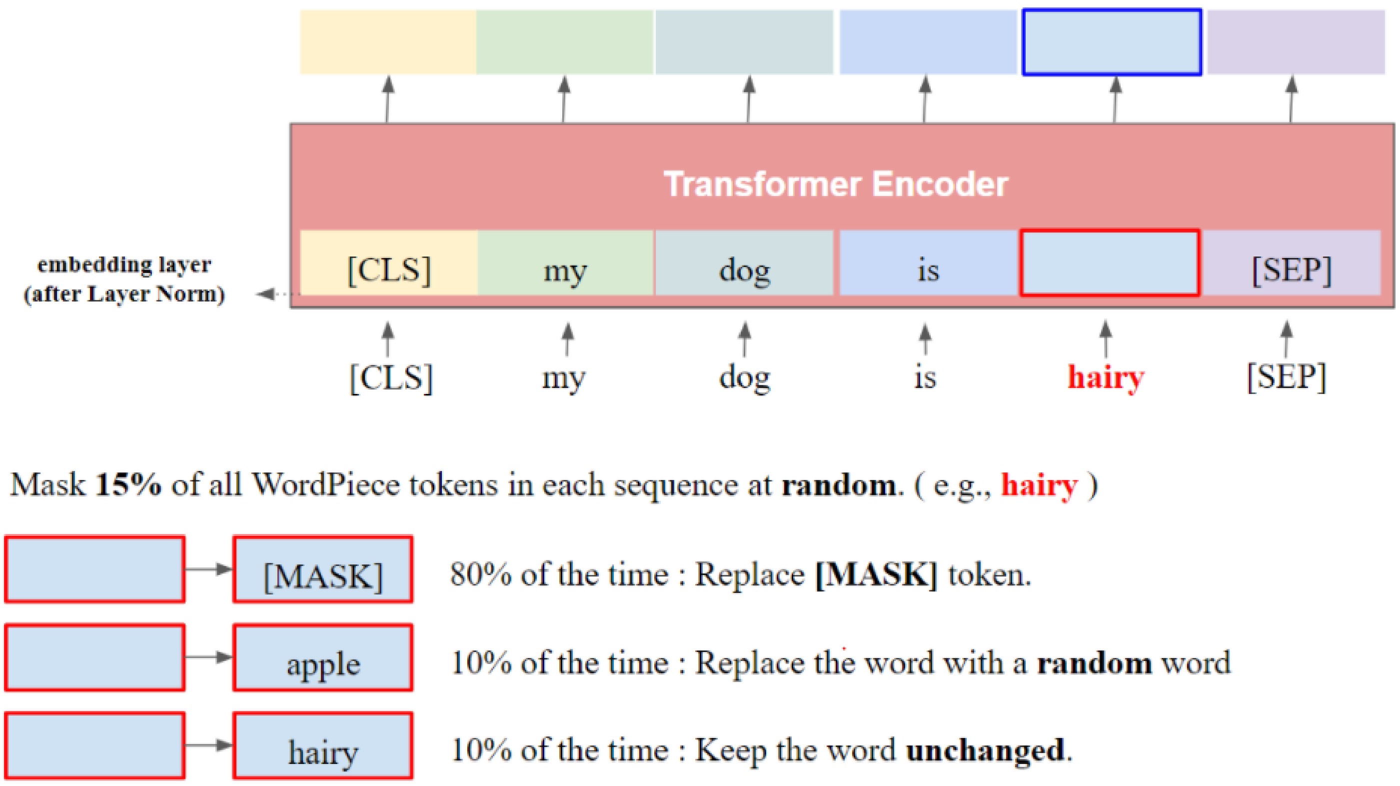
Seong Su Kim

[2]J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805, 2018.*

김성수

Self-supervised Learning

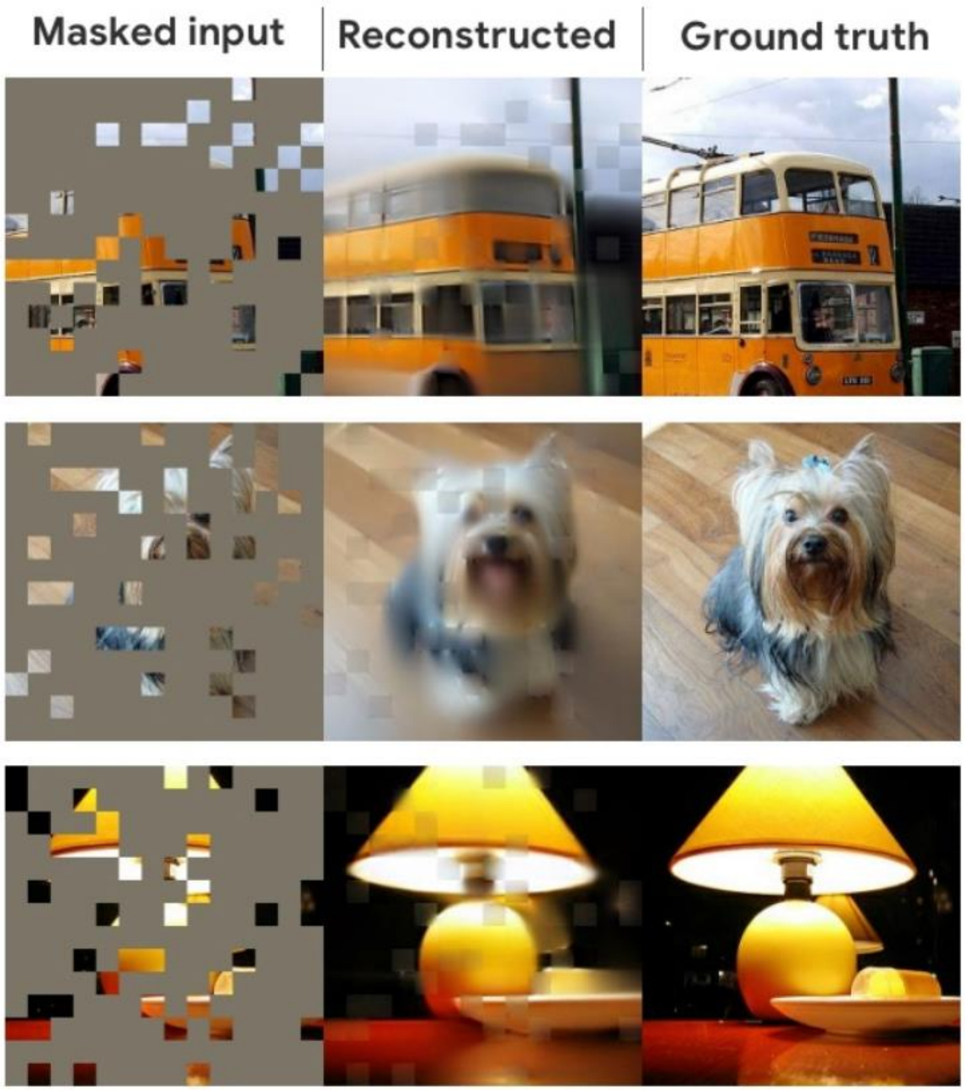
- *BERT[2] : Masked Language Modeling*



[3]K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," *arXiv preprint arXiv:2111.06377, 2021.*

Self-supervised Learning

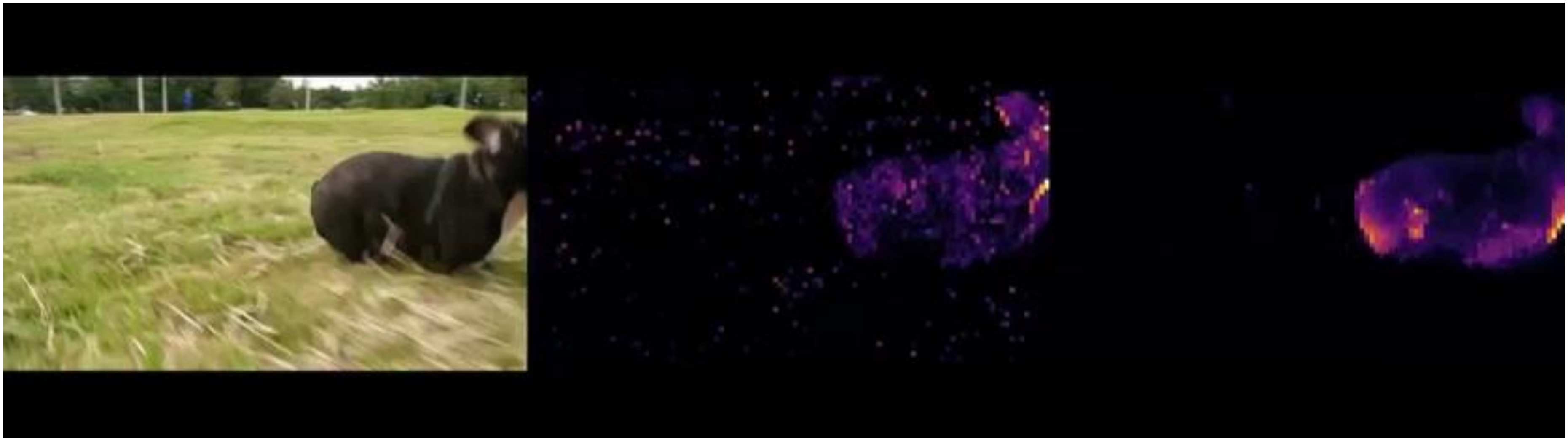
- *MAE : Masked Auto Encoder[3]*



DINO[1]

[1]M. Caron *et al., "Emerging properties in self-supervised vision transformers," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650-9660.*

김성수



Supervised

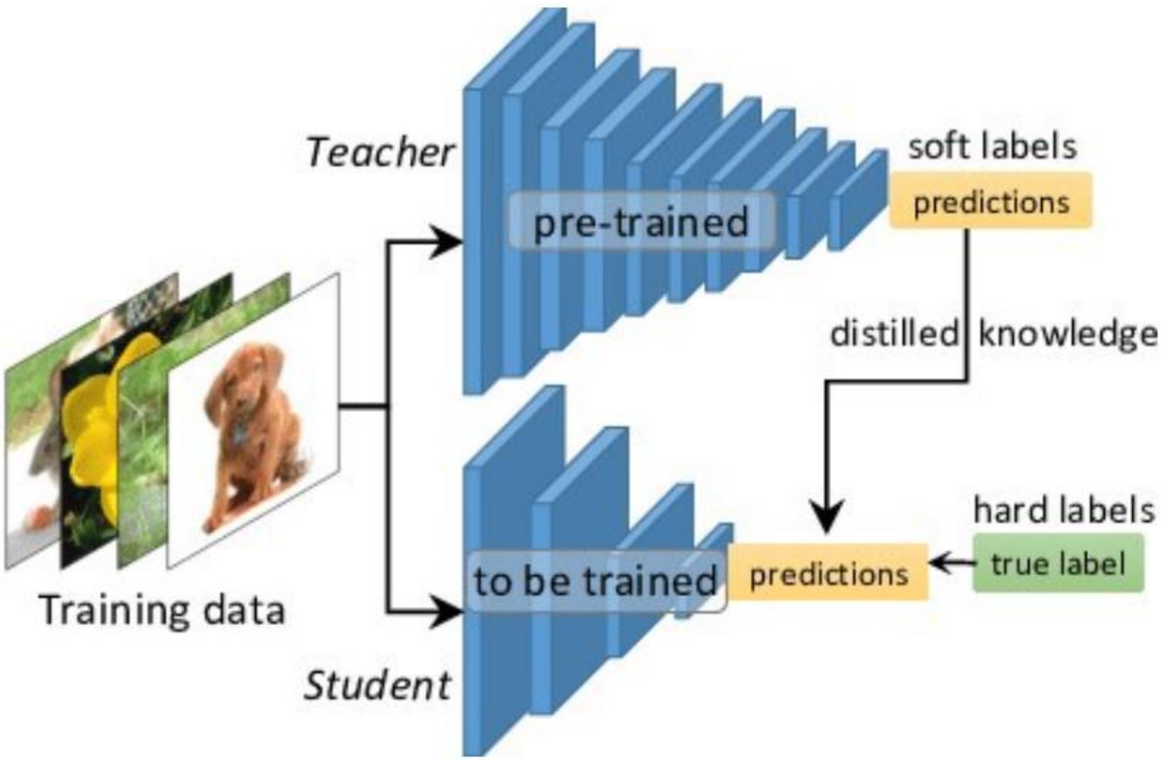
Self-supervised(DINO[1])

DINO[1]

[1]M. Caron *et al.*, *"Emerging properties in self-supervised vision transformers," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650-9660.*

김성수

- *Self-distillation with no labels*
  - *Knowledge Distillation*
  - *No labels : self-supervised learning*
- *Knowledge distillation(Supervised)*



DINO[1]

[1]M. Caron *et al.*, *"Emerging properties in self-supervised vision transformers," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650-9660.*

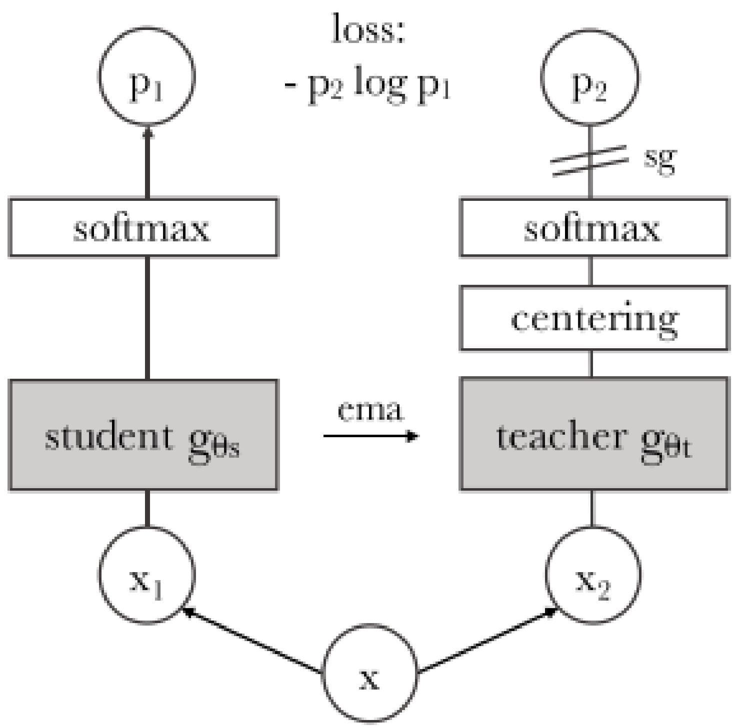
김성수

- *Self-distillation with no labels*

Output probability

Model

Input Image



- *Softmax, Student, teacher?*

DINO[1]

[1]M. Caron *et al.*, *"Emerging properties in self-supervised vision transformers," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650-9660.*

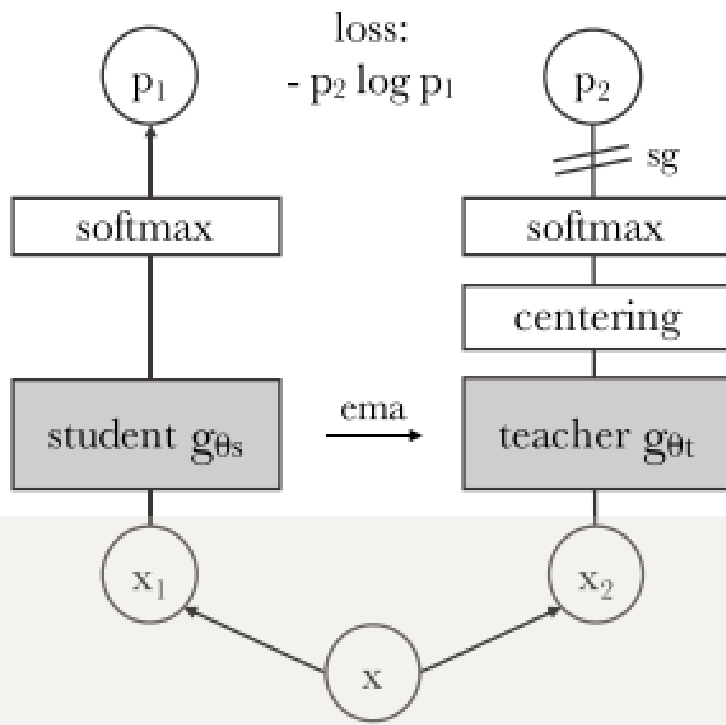
김성수

- *Self-distillation with no labels*
  - *기본 Idea : 같은 image에 대해 augmentation에 상관 없이 동일한 feature를 뽑을 수 있도록 하자!*

Output probability

Model

Input Image



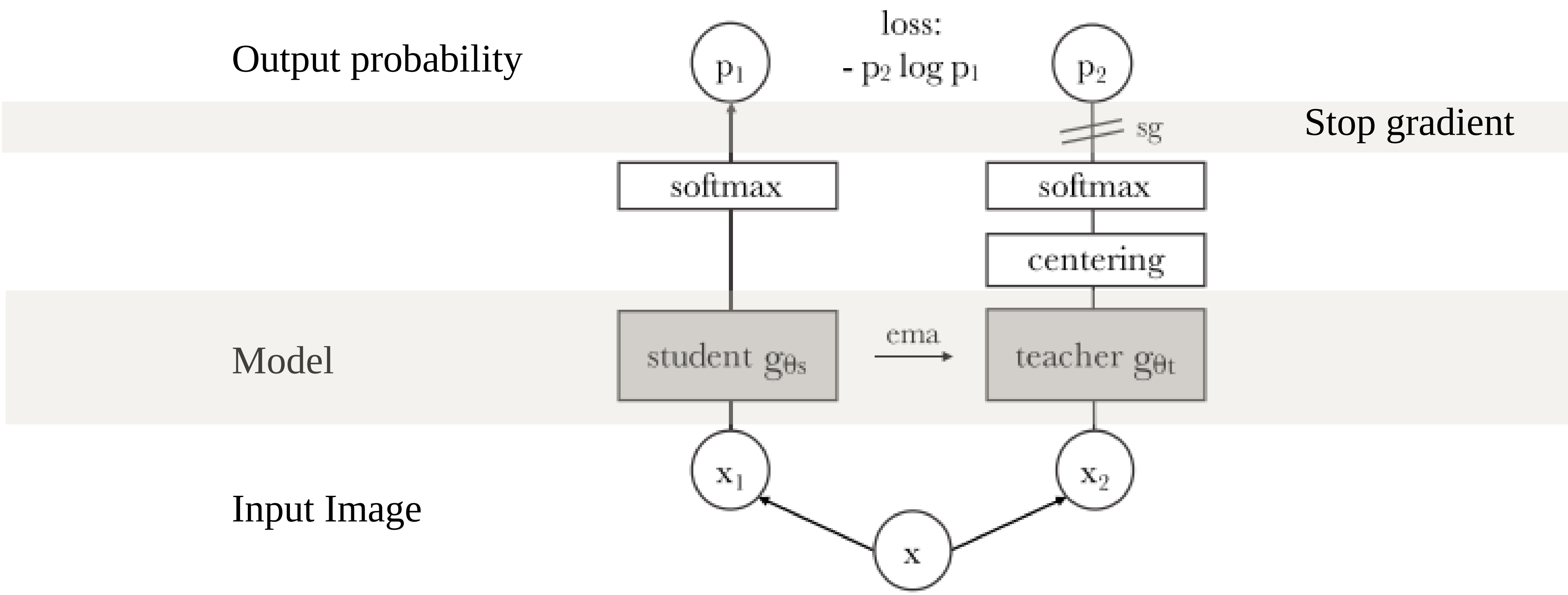
x1 = Augmentation 1
x2 = Augmentation 2

DINO[1]

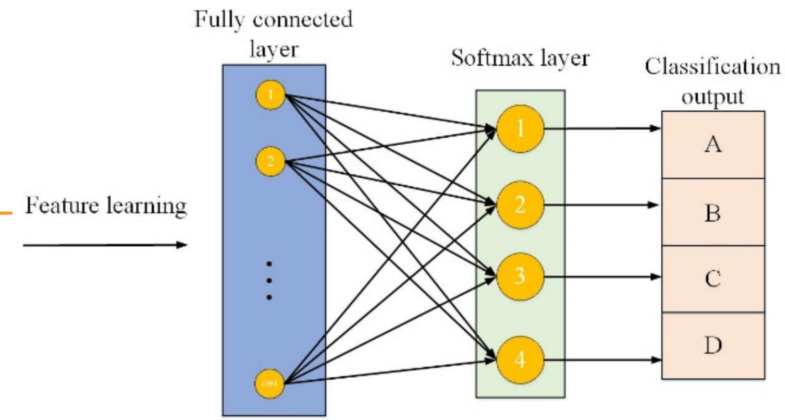
[1]M. Caron *et al.*, *"Emerging properties in self-supervised vision transformers," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650-9660.*

김성수

- *Self-distillation with no labels*
  - *Student model과 teacher model은 동일*
  - *Teacher는 Student의 parameter를 exponential moving average 하여 받음*



DINO[1]

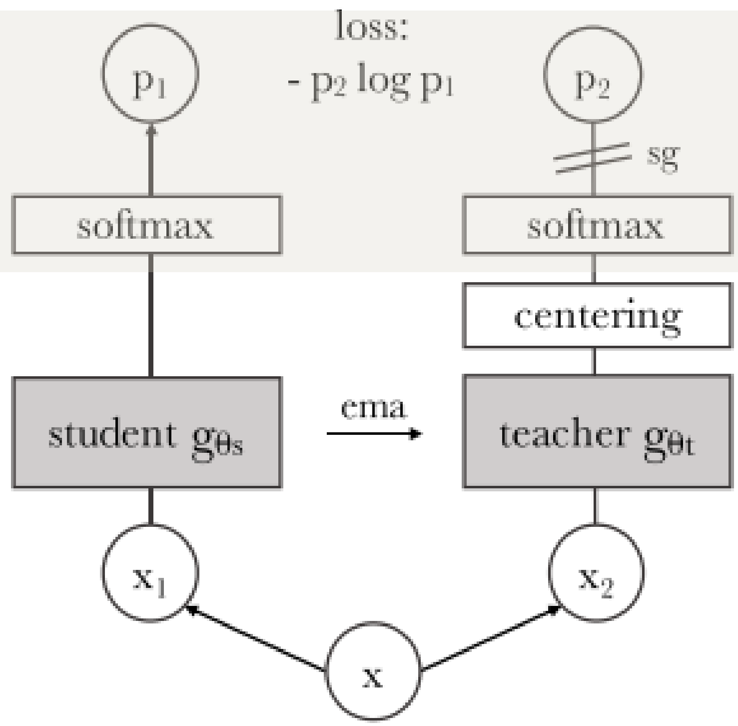


- *Self-distillation with no labels*
  - *Similarity metric : L2 distance, dot product..*
  - *Why they used Cross entropy with soft-max ?*
    - *Model이 집중해야할 feature를 찾는 것을마치 classification 문제 처럼 유도!*

Output probability

Model

Input Image



DINO[1]

$$P_s(x)^{(i)} = \frac{\exp(g_{\theta_s}(x)^{(i)}/\tau_s)}{\sum_{k=1}^{K} \exp(g_{\theta_s}(x)^{(k)}/\tau_s)}$$

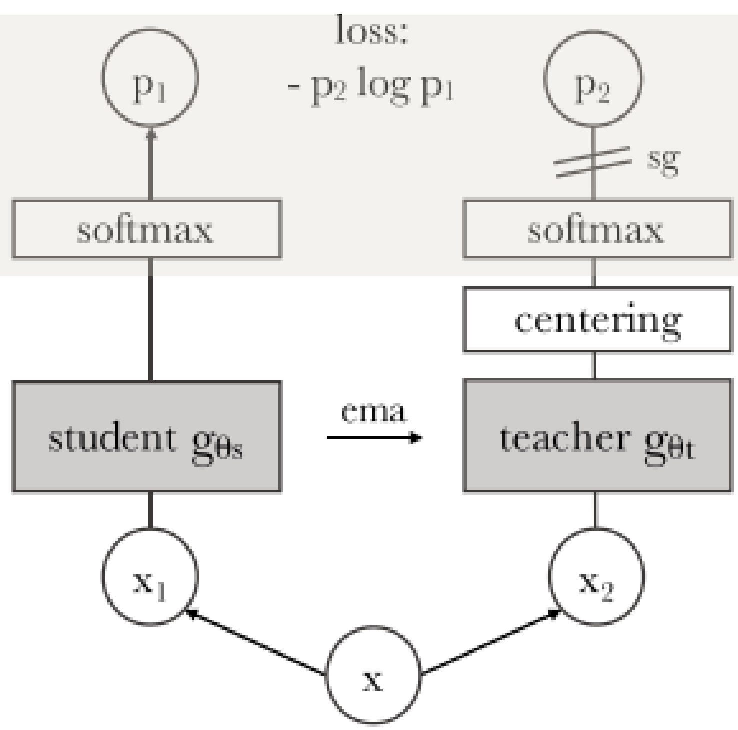
- *Self-distillation with no labels*
  - *Similarity metric : L2 distance, dot product..*
  - *Why they used Cross entropy with soft-max ?*
    - *Teacher : Low Temperature*
    - *Student : High Temperature*

Sharpening

Output probability

Model

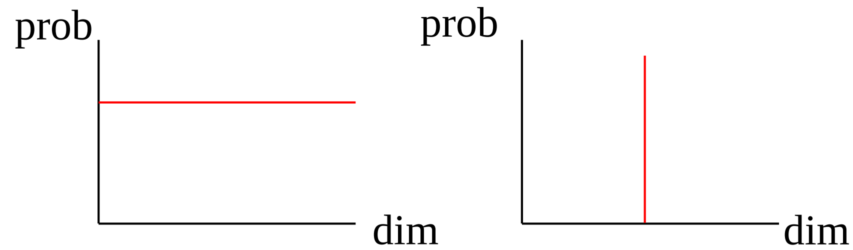
Input Image



김성수

DINO[1]

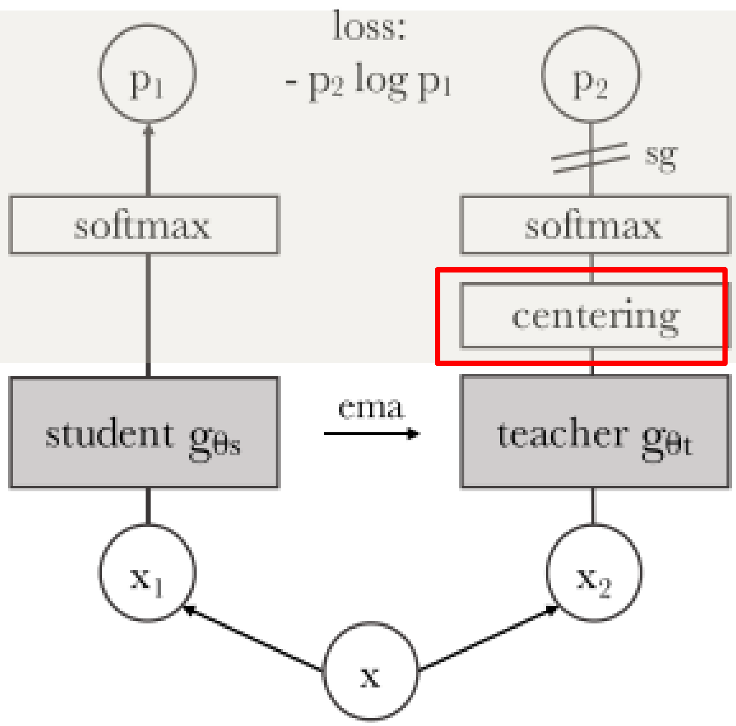
- *Self-distillation with no labels*
  - *Two reasons of collapse*
    - *Uniform distribution, Impulse distribution*



Output probability

Model

Input Image



$$c \leftarrow mc + (1-m)\frac{1}{B}\sum_{i=1}^{B} g_{\theta_t}(x_i)$$

DINO[1]

- *Self-distillation with no labels*
  - *Details*
    - *Collapse study*

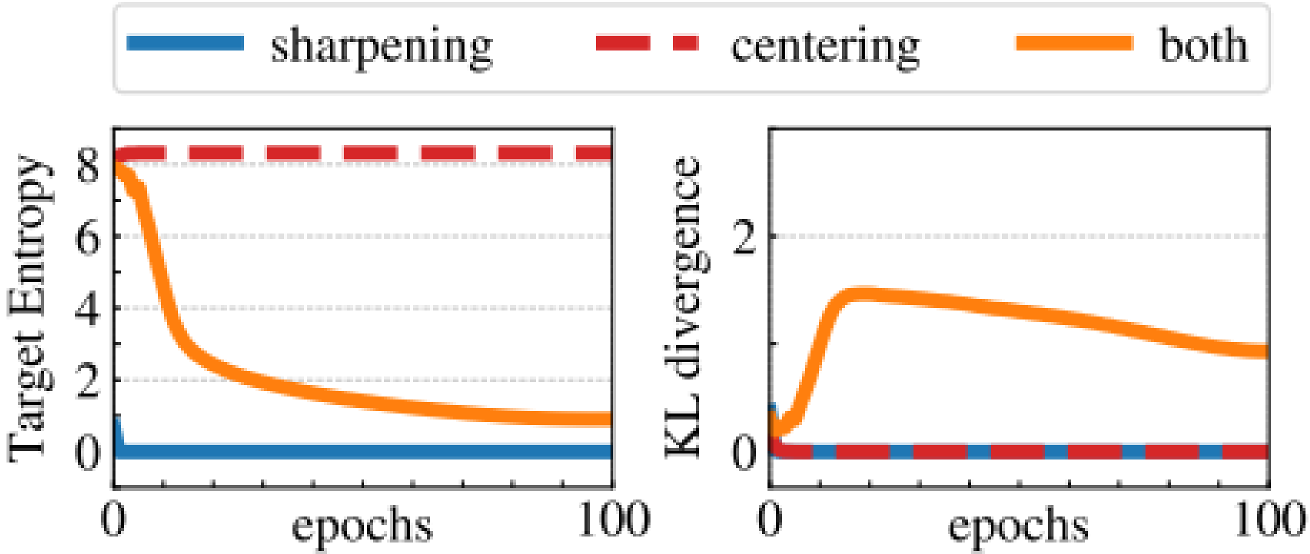


Figure 3: **Collapse study. (left)**: evolution of the teacher's target entropy along training epochs; **(right)**: evolution of KL divergence between teacher and student outputs.

김성수

DINO[1]

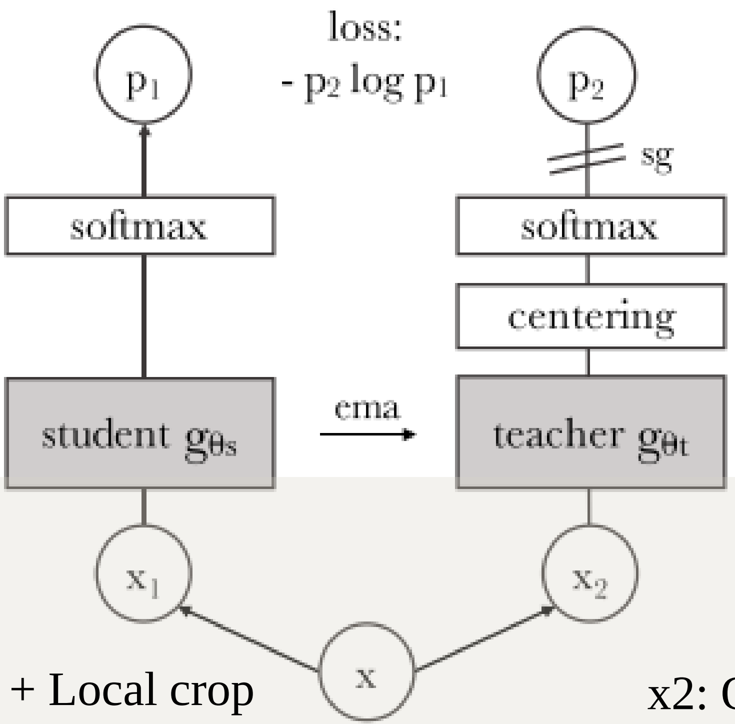
- *Self-distillation with no labels*
  - *Details*
    - *Teacher model : Only use Global crop*
    - *Student model : Global crop + Local crop*

Global crop Local crop

Output probability

Model

Input Image



x1 : Global crop + Local crop

x2: Only Global crop

DINO[1]

- *Self-distillation with no labels*
  - *Pseudo Code*

Algorithm 1 DINO PyTorch pseudocode w/o multi-crop.

```
# gs, gt: student and teacher networks
# C: center (K)
# tps, tpt: student and teacher temperatures
# l, m: network and center momentum rates
gt.params = gs.params
for x in loader: # load a minibatch x with n samples
    x1, x2 = augment(x), augment(x) # random views

    s1, s2 = gs(x1), gs(x2) # student output n-by-K
    t1, t2 = gt(x1), gt(x2) # teacher output n-by-K

    loss = H(t1, s2)/2 + H(t2, s1)/2
    loss.backward() # back-propagate

    # student, teacher and center updates
    update(gs) # SGD
    gt.params = l*gt.params + (1-l)*gs.params
    C = m*C + (1-m)*cat([t1, t2]).mean(dim=0)

def H(t, s):
    t = t.detach() # stop gradient
    s = softmax(s / tps, dim=1)
    t = softmax((t - C) / tpt, dim=1) # center + sharpen
    return - (t * log(s)).sum(dim=1).mean()
```

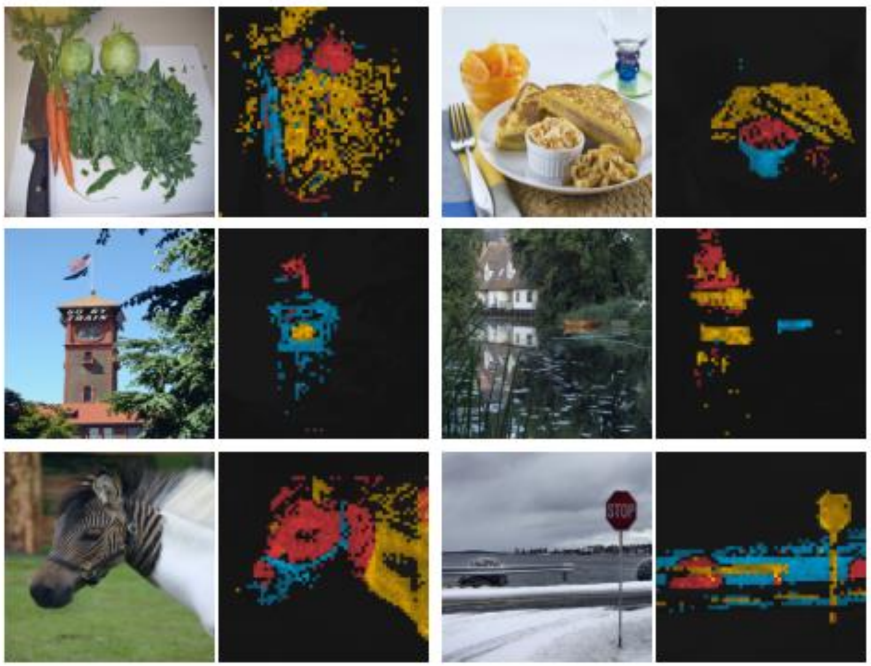
DINO[1]

- *Self-distillation with no labels*
  - *Pseudo Code*

김성수

DINO[1]

- *Self-**di**stillation with **no** labels*
  - *Result*



DINO[1]

- *Self-distillation with no labels*
  - *1. Augmentation*
  - *2. Datasets*