Towards Contrastive Learning

Open DMQA Seminar

2021-01-29

곽민구

발표자 소개

- 이름: 곽민구
- 학력
  - ✓ 2010.03 – 2016.02 | 학사 | 고려대학교 산업경영공학과
  - ✓ 2016.03 – 현재 | 석박사통합과정 | 고려대학교 산업경영공학과 (지도교수: 김성범)
- 연구분야
  - ✓ Self-supervised learning
  - ✓ Out-of-distribution detection
  - ✓ Signal data analysis
- e-mail: min9kwak@korea.ac.kr

세미나 개요

1. Contrastive learning and NCE loss

2. How to improve the contrastive learning?
   - ✓ Develop network
   - ✓ Robust data distribution
   - ✓ Define example nicely

story

01 Supervised Learning

- Input data (X)에 대한 레이블 정보(y)가 쌍으로 주어지고, 이 사이의 관계를 학습
- 다양한 분야에서 뛰어난 성능을 보이는 모델들이 개발됨

images (X)

?

network

labels (y)

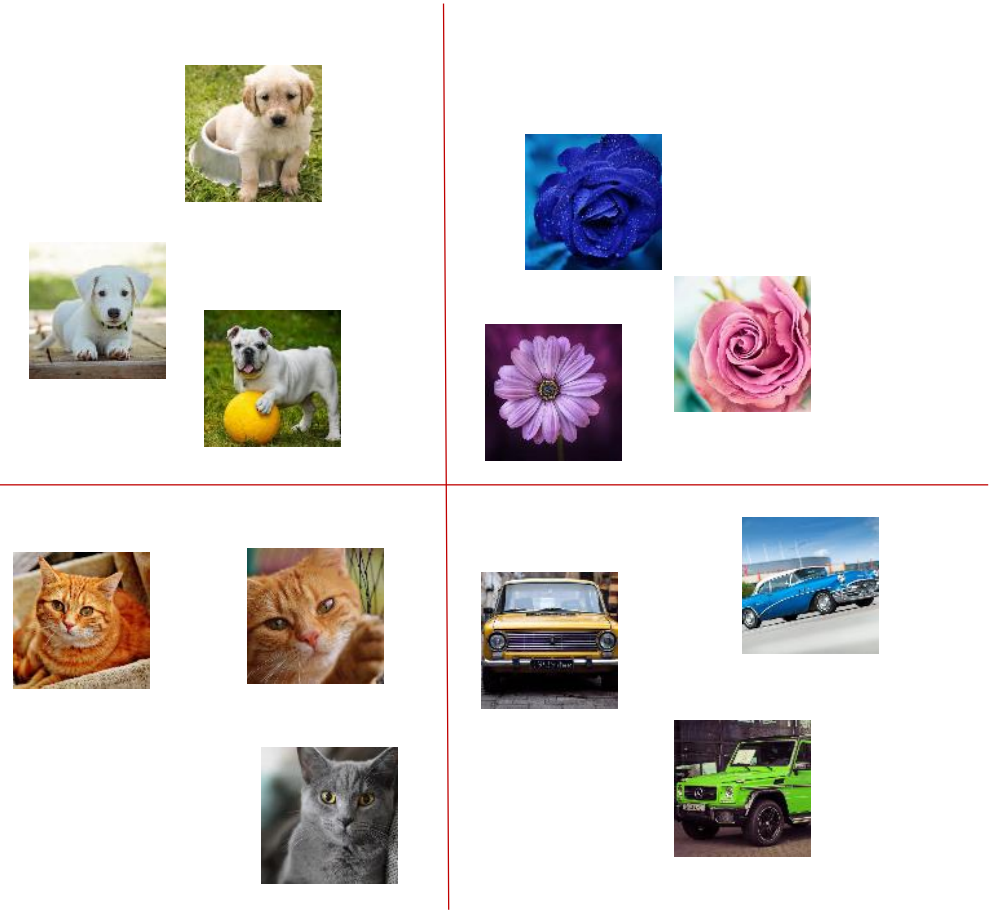
dog

cat

vehicle

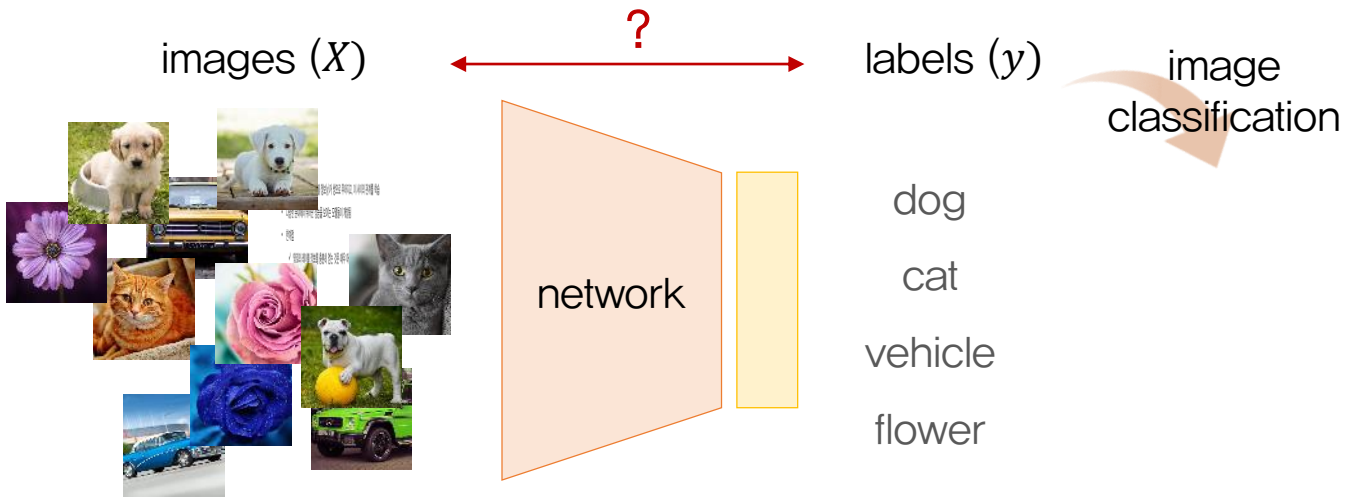
flower

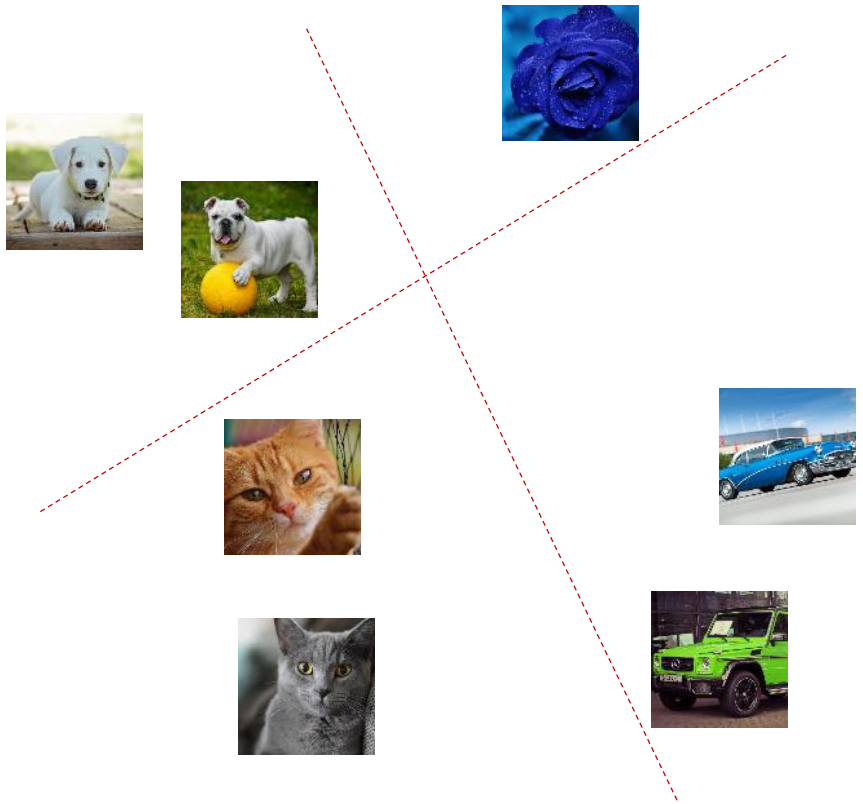
image classification



01 Supervised Learning

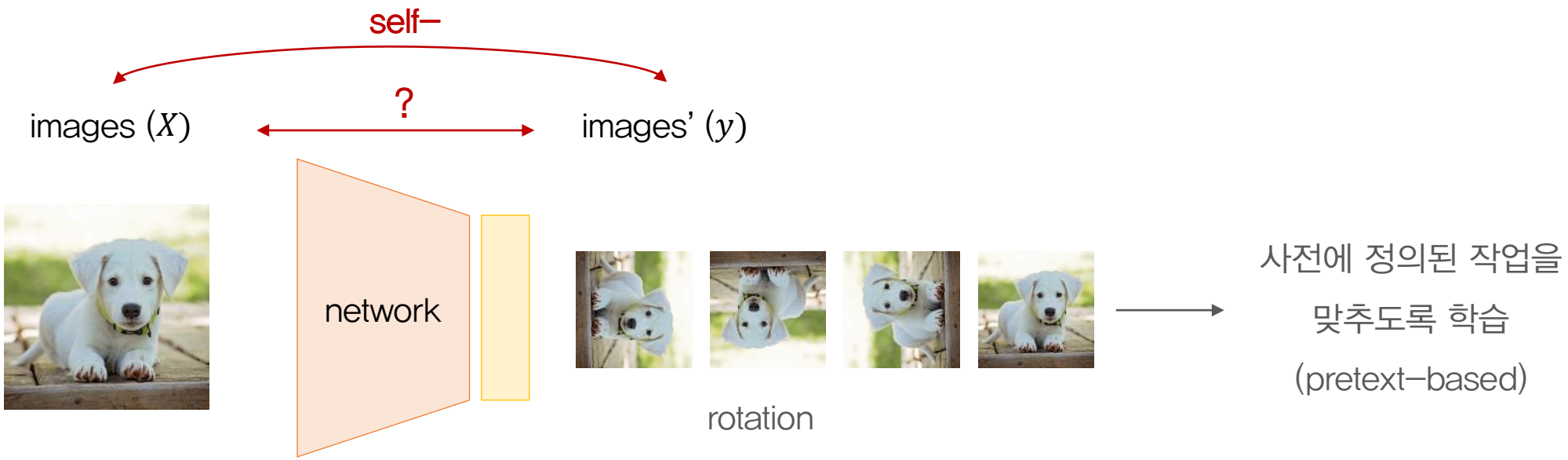
- Input data (X)에 대한 레이블 정보(y)가 쌍으로 주어지고, 이 사이의 관계를 학습
- 다양한 분야에서 뛰어난 성능을 보이는 모델들이 개발됨
- 한계점
  - ✓ 양질의 레이블 정보를 충분히 얻는 것은 매우 어려움





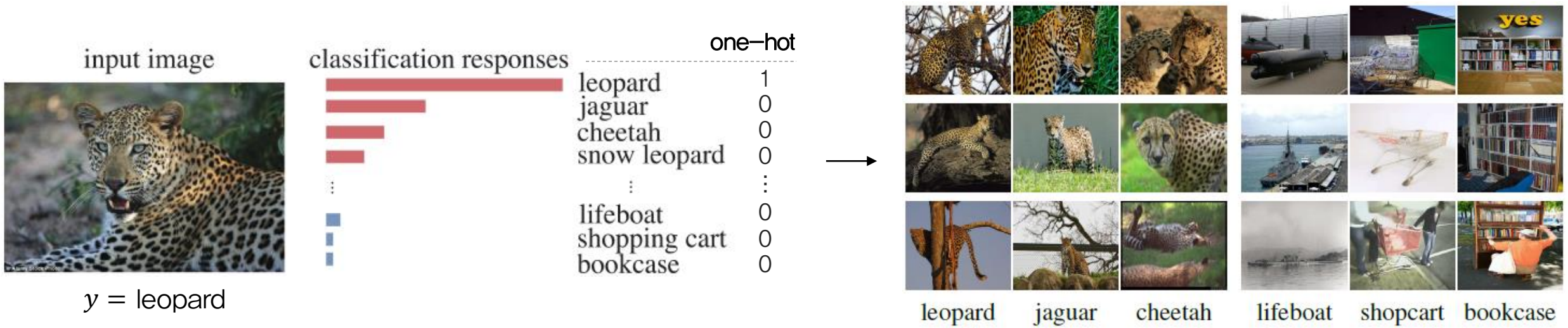
01 Self-Supervised Learning

- Input data (X) 자체에서 쉽게 얻을 수 있는 정보를 사용해서 좋은 representation을 학습
  - Pretext-based learning: jigsaw, colorization, rotation, etc
  - Contrastive learning: (without pretext) contrastive loss … ??



01 Instance Discrimination

- Supervised learning에서는 이미지가 특정 클래스에 속하는지 아닌지를 구분하는 1 vs. 0으로 학습 (one-hot)
- Semantic labeling이 아님에도 불구하고, 확률값이 높은 클래스는 타겟 클래스와 시각적으로 유사한 것을 발견



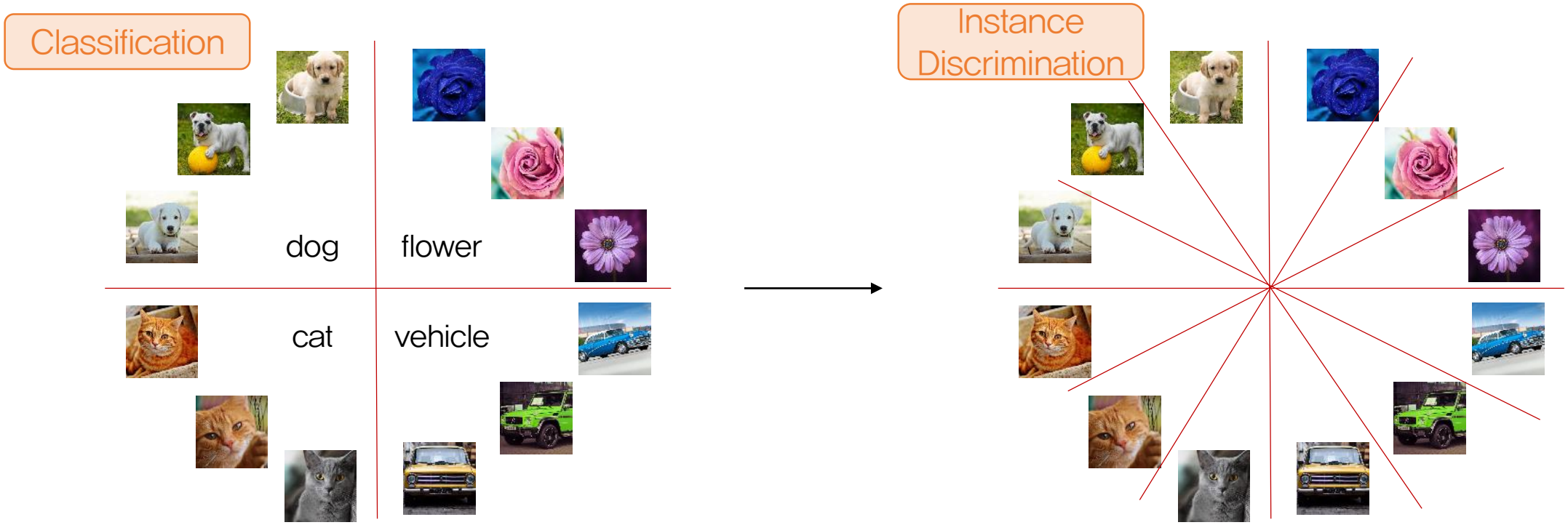
Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3733-3742).

01 Instance Discrimination

- 레이블 정보 없이 관측치 레벨에서 학습 · 공유하는 representation · semantic structure가 있을 것이다.

  → 관측치 사이의 유사성을 기반으로, 관측치끼리 구분하도록 학습을 한다면

  레이블 정보 없이도 좋은 representation을 얻을 수 있지 않을까?

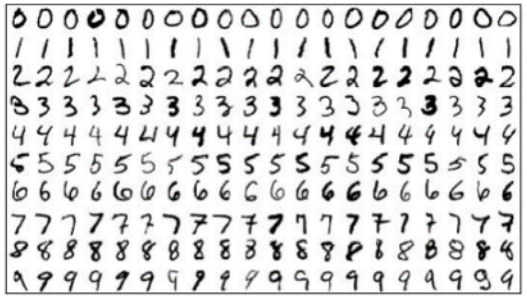


Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3733-3742).

01 Instance Discrimination

- 수많은 관측치들을 어떻게 구분하도록 만들 수 있을까?
  - ✓ ImageNet 데이터는 약 1200만개 이상의 관측치를 갖고 있어 softmax를 단순히 사용하기에는 어렵다
  - ✓ 분모가 커짐으로 인해 확률값이 너무 작아져서 학습이 어려움

[학습 데이터셋 개요]



- MNIST
  클래스 – 10개
  관측치 – 60,000개

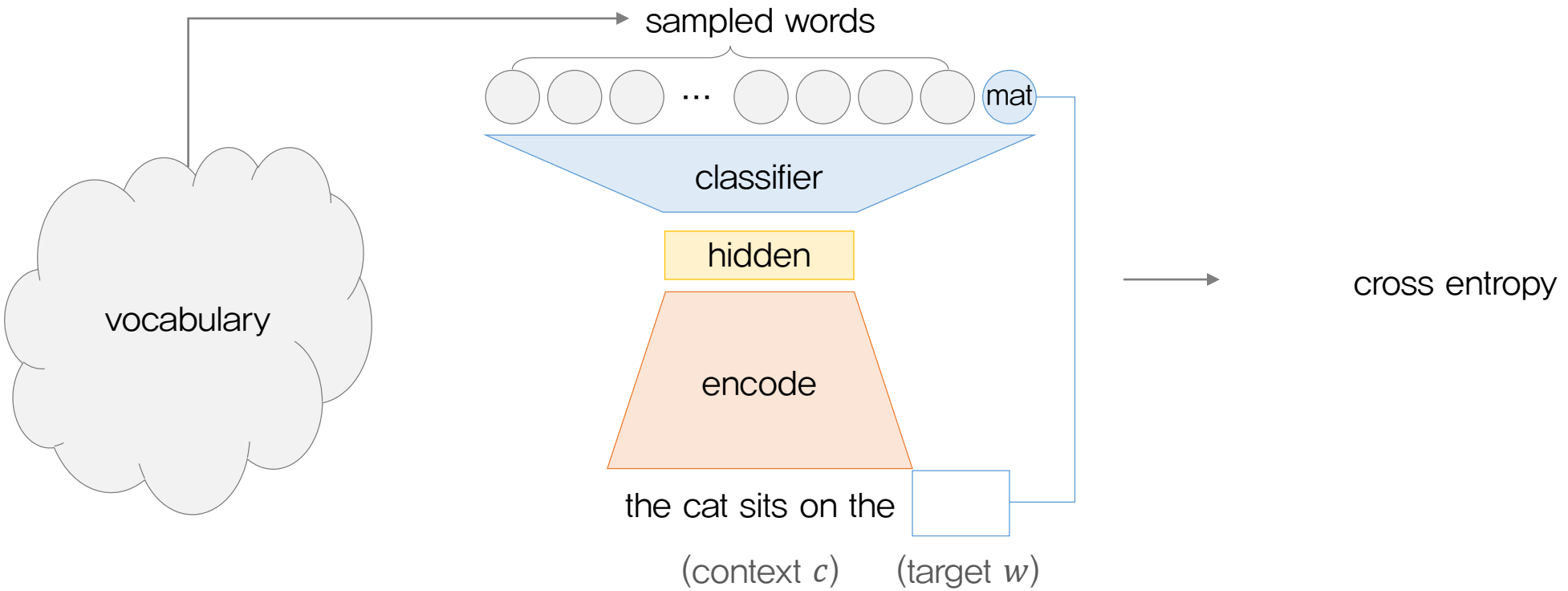
- CIFAR100
  클래스 – 100개
  관측치 – 50,000개

- ImageNet
  클래스 – 1,000개
  관측치 – 1,200,000개 이상
  → 1,200,000 클래스…?

Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3733-3742).

01 Noise Contrastive Estimation

- 자연어 처리 (natural language processing) Word Embedding Model에서도 같은 문제를 해결해야 함
  - ✓ 단어 개수 = 클래스 개수
- Softmax에서 너무 많은 단어들에 대해서 계산을 해야 하니, 일부만 샘플링하여 계산하자



Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3733-3742).

01 Intuitive Understanding

negative log-likelihood

non-parameteric softmax

memory bank

L2-normalization

Non-Parametric Classifier. The problem with the parametric softmax formulation in Eq. (1) is that the weight vector $\mathbf{w}$ serves as a class prototype, preventing explicit comparisons between instances.

We propose a *non-parametric* variant of Eq. (1) that replaces $\mathbf{w}_j^T\mathbf{v}$ with $\mathbf{v}_j^T\mathbf{v}$, and we enforce $\|\mathbf{v}\| = 1$ via a L2-normalization layer. Then the probability $P(i|\mathbf{v})$ becomes:

$$P(i|\mathbf{v}) = \frac{\exp\left(\mathbf{v}_i^T\mathbf{v}/\tau\right)}{\sum_{j=1}^{n}\exp\left(\mathbf{v}_j^T\mathbf{v}/\tau\right)}, \tag{2}$$

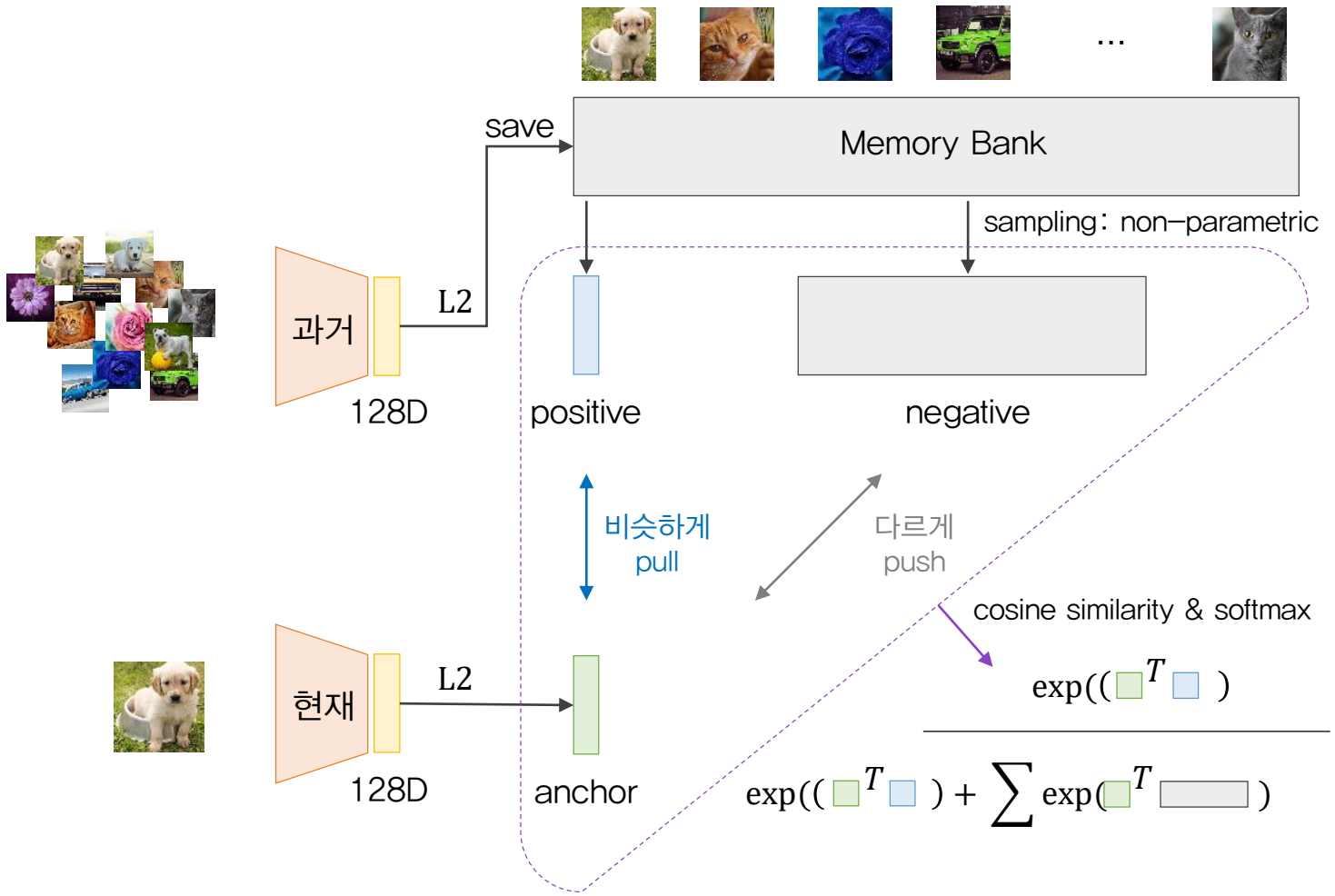
The learning objective is then to maximize the joint probability $\prod_{i=1}^{n} P_{\boldsymbol{\theta}}(i|f_{\boldsymbol{\theta}}(x_i))$, or equivalently to minimize the negative log-likelihood over the training set, as

$$J(\boldsymbol{\theta}) = -\sum_{i=1}^{n}\log P(i|f_{\boldsymbol{\theta}}(x_i)). \tag{3}$$

Learning with A Memory Bank. To compute the probability $P(i|\mathbf{v})$ in Eq. (2), $\{\mathbf{v}_j\}$ for all the images are needed. Instead of exhaustively computing these representations every time, we maintain a feature memory bank V for storing them [46]. In the following, we introduce separate notations for the memory bank and features forwarded from the network. Let $V = \{\mathbf{v}_j\}$ be the memory bank and $\mathbf{f}_i = f_{\boldsymbol{\theta}}(x_i)$ be the feature of x_i. During each learning iteration, the representation $\mathbf{f}_i$ as well as the network parameters θ are optimized via stochastic gradient descend. Then $\mathbf{f}_i$ is updated to V at the corresponding instance entry $\mathbf{f}_i \rightarrow \mathbf{v}_i$. We initialize all the representations in the memory bank V as unit random vectors.

Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3733-3742).

01 Intuitive Understanding



① Instance discrimination을 하기 위해 모든 데이터의 임베딩 정보를 저장한다. 이때 임베딩 크기를 맞추기 위해 L2-normalization을 수행한다.

② 현재 타겟 이미지의 과거 임베딩 벡터 & 나머지 이미지들의 일부 임베딩 벡터를 샘플링

③ 비슷한 정도를 측정하기 위해 cosine 유사도 측정 (L2 Euclidean – Cosine)

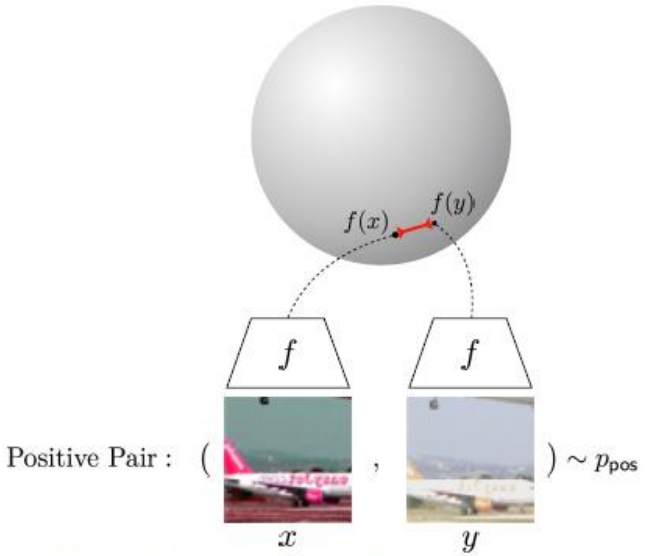
④ 나 자신에 대한 유사도는 높이며, 나머지는 내리고 싶다. Cosine 유사도를 logit으로 사용하고 확률로 표현하기 위해 softmax 적용 (NCE) → $P(i|\boldsymbol{v})$

⑤ Loss: 앞에서 얻은 cosine 유사도를 logit으로 사용하고, 타겟 클래스 레이블을 모두 0으로 cross entropy. (minimize negative log-likelihood)

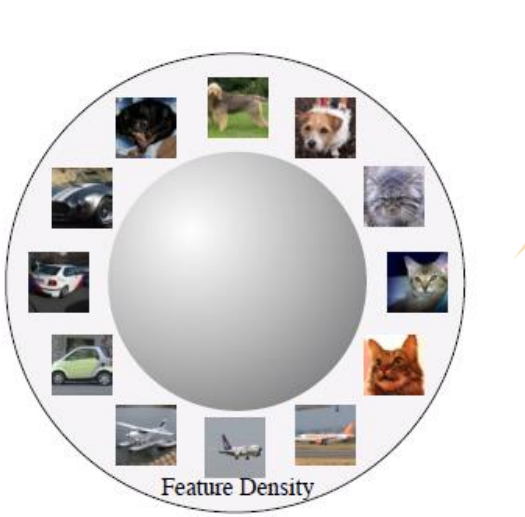
Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3733-3742).

01 Geometric Interpretation

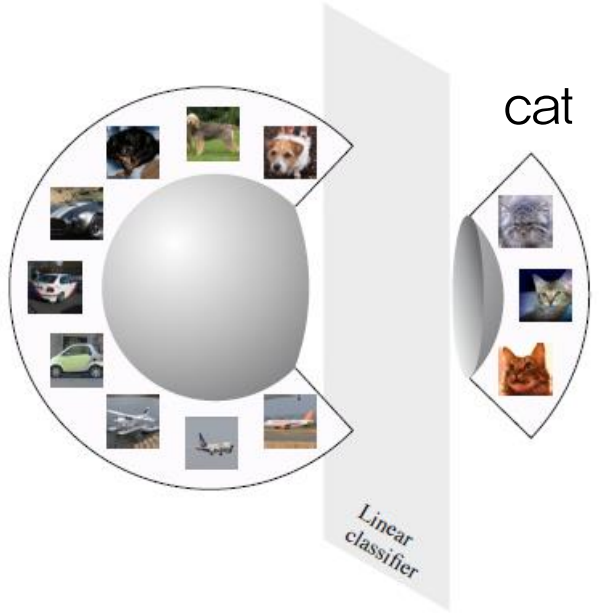
- Contrastive learning이 잘 되었을 때, hypersphere를 따라서 semantic 정보가 유사한 관측치들끼리 모인다
- 학습된 representation의 성능 · 품질을 확인하기 위해서 모델 가중치를 고정시키고 linear classifier를 사용하여 분류 정확도를 산출



Alignment: Similar samples have similar features.
(Figure inspired by Tian et al. (2019).)



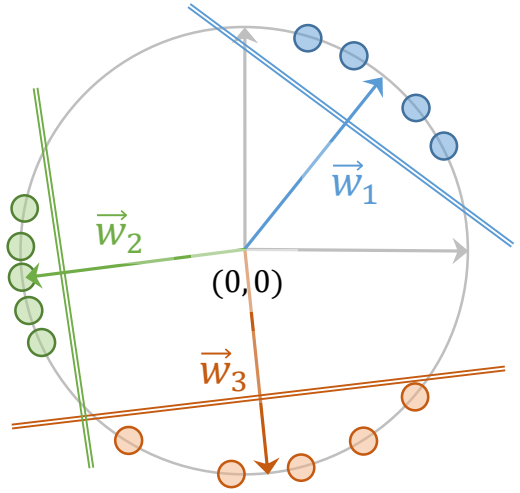
Uniformity: Preserve maximal information.



Wang, T., & Isola, P. (2020, November). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In International Conference on Machine Learning (pp. 9929-9939). PMLR.

01 | Geometric Interpretation

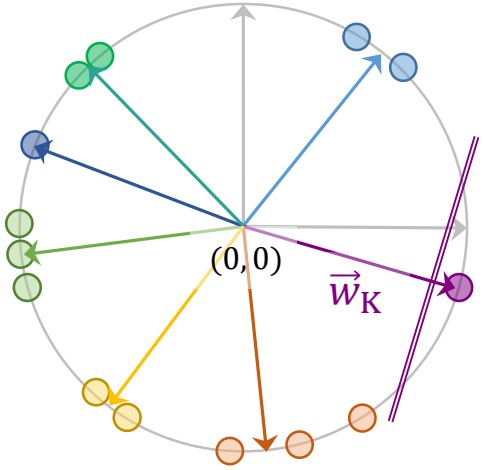
3–way classification



$$\max_{\vec{w}_c}\left\{\frac{1}{N_c}\sum_{i}^{N_c}\vec{w}_c\cdot\vec{x}_i\right\}$$

$solution$: $\{\vec{w}_1,\vec{w}_2,\vec{w}_3\}$

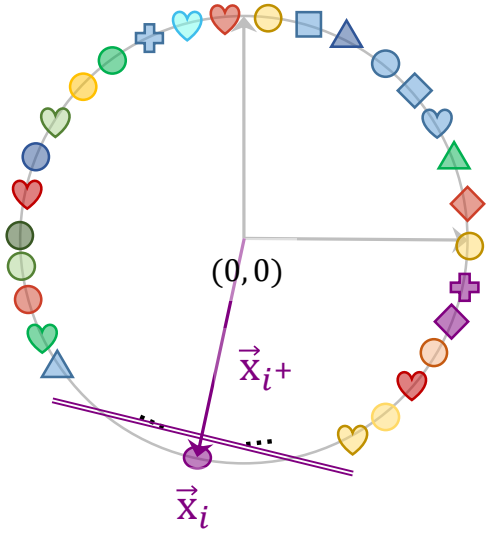
K–way classification



$$\max_{\vec{w}_k}\left\{\frac{1}{N_k}\sum_{i}^{N_k}\vec{w}_k\cdot\vec{x}_i\right\}$$

$solution$: $\{\vec{w}_1,\ldots,\vec{w}_K\}$

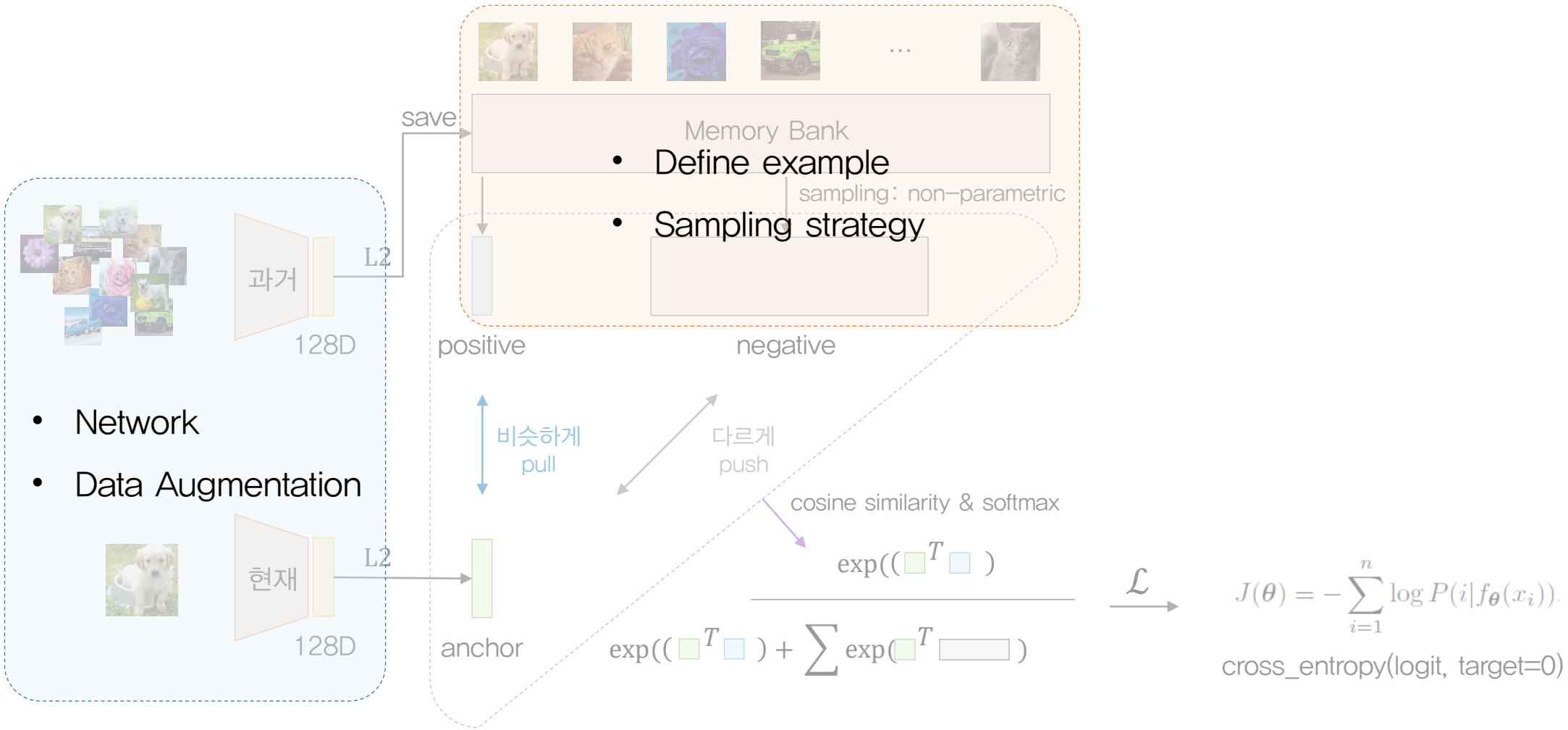
N–way classification (Instance discrimination)



$$\max_{\vec{x}_{i+}}\{\vec{x}_{i+}\cdot\vec{x}_i\}$$

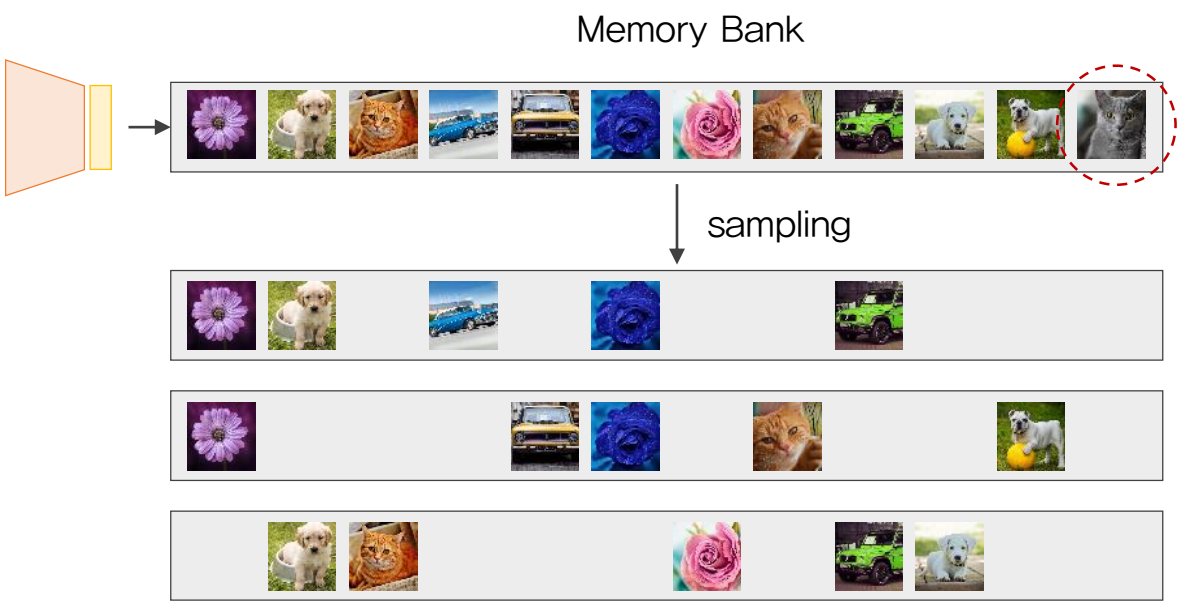
$solution$: $\{\vec{x}_{i+}\}_{i=1}^{N}$

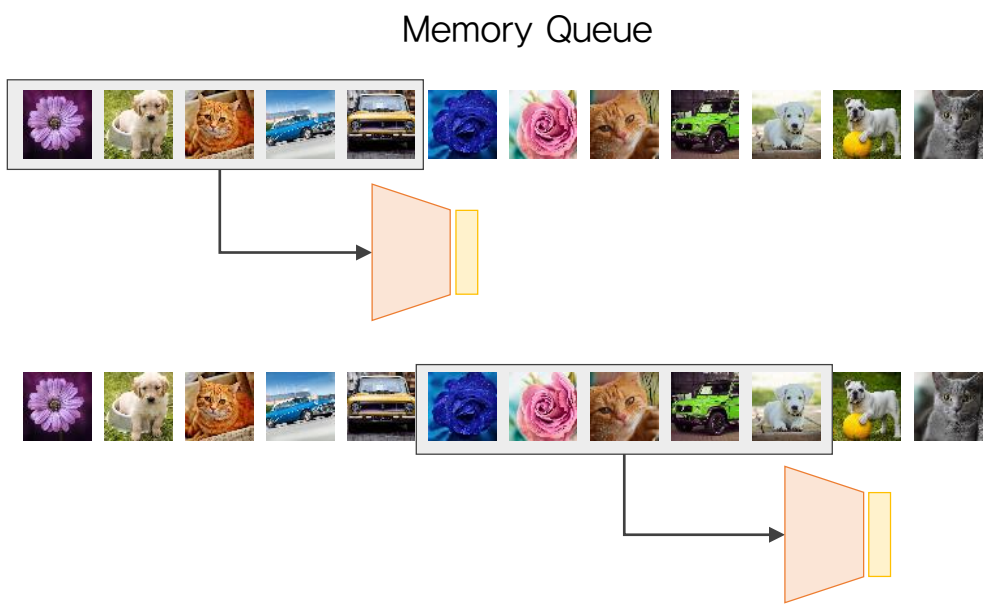
01 | Improve Performance



02 | MoCo

- Memory Bank의 단점
  - 관측치 전체에 대한 embedding 정보를 계속 저장하고 있어야 한다 – 메모리 이슈
  - 랜덤 샘플링을 통해 negative example을 구성 – 데이터 샘플별로 학습에 기여하는 정도가 다르다
- Memory Queue를 사용
  - First In First Out (FIFO)

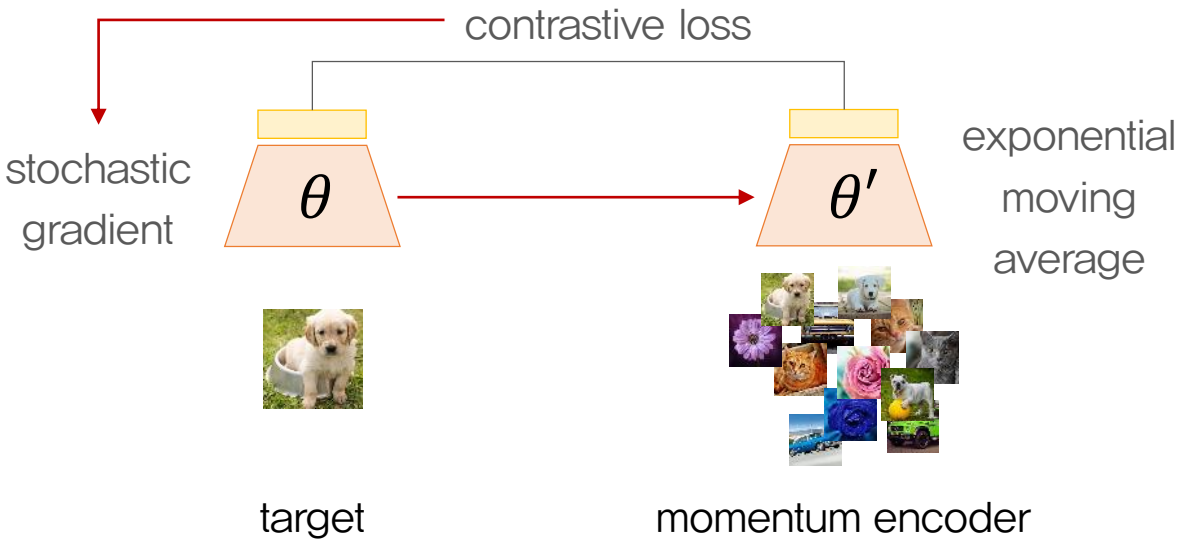




He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9729-9738).

02 | MoCo

- Momentum Encoding
  - 과거 모델 weight들의 가중평균으로 업데이트
  - 과거의 나 자신들이 'teacher'가 되는 mean teacher 방식과 동일 (semi-supervised model)
  - Exponential moving average, momentum encoding, on-the-fly ensemble, mean teacher 등등 다양한 이름으로 사용

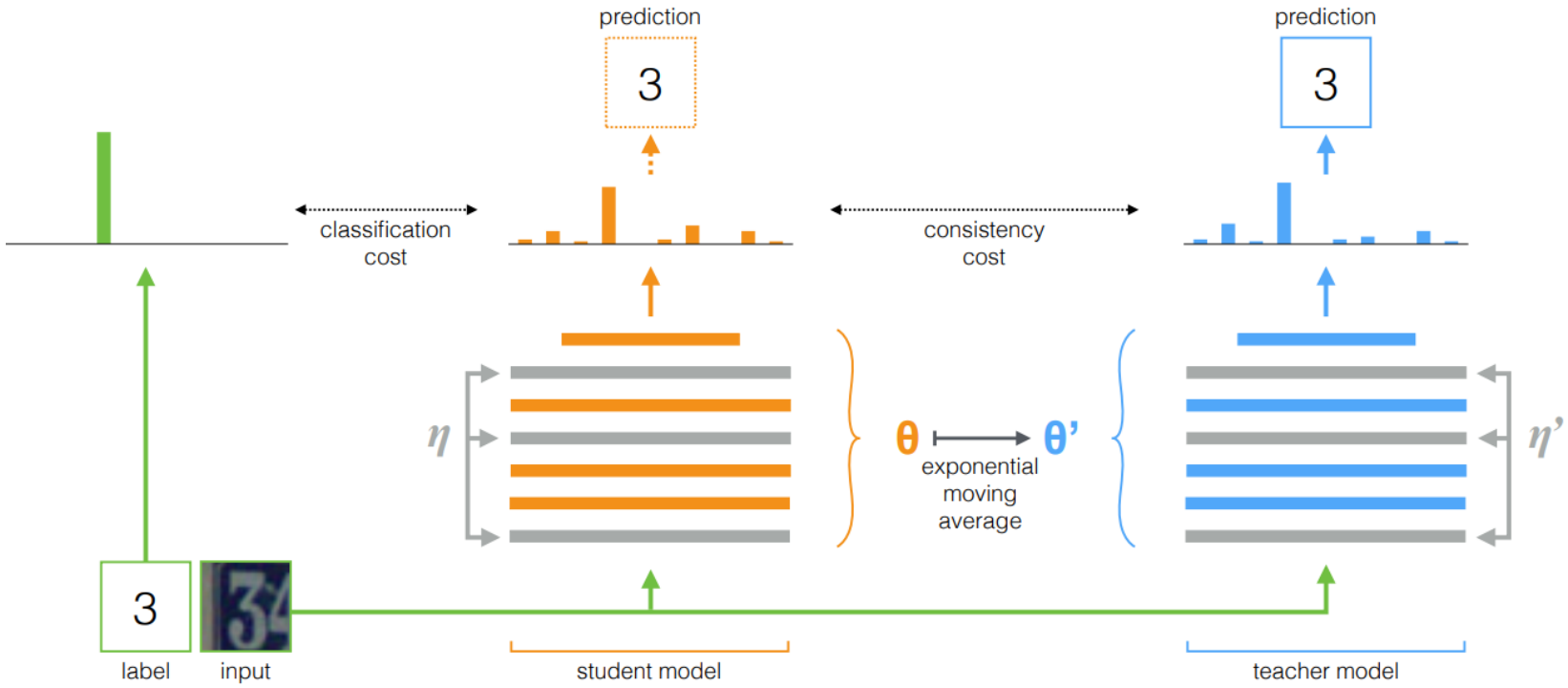


$$\theta' = m\theta' + (1-m)\theta$$

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9729-9738).

02 | MoCo

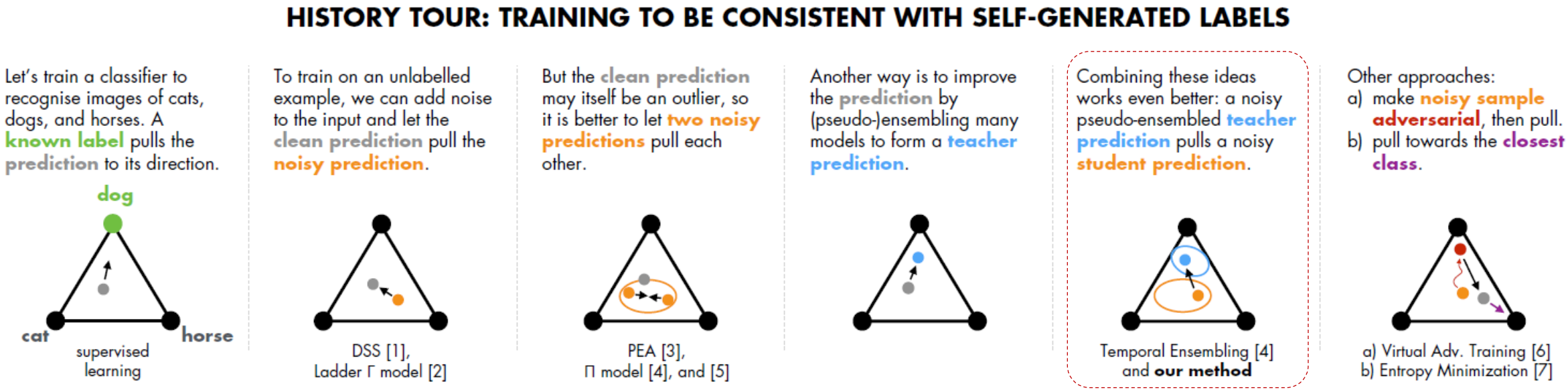
- Momentum Encoding
  - 과거 모델 weight들의 가중평균으로 업데이트
  - 과거의 나 자신들이 'teacher'가 되는 mean teacher 방식과 동일 (semi-supervised model)
  - Exponential moving average, momentum encoding, on-the-fly ensemble, mean teacher 등등 다양한 이름으로 사용



Tarvainen, A., & Valpola, H. (2017, December). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 1195-1204).

02 | MoCo

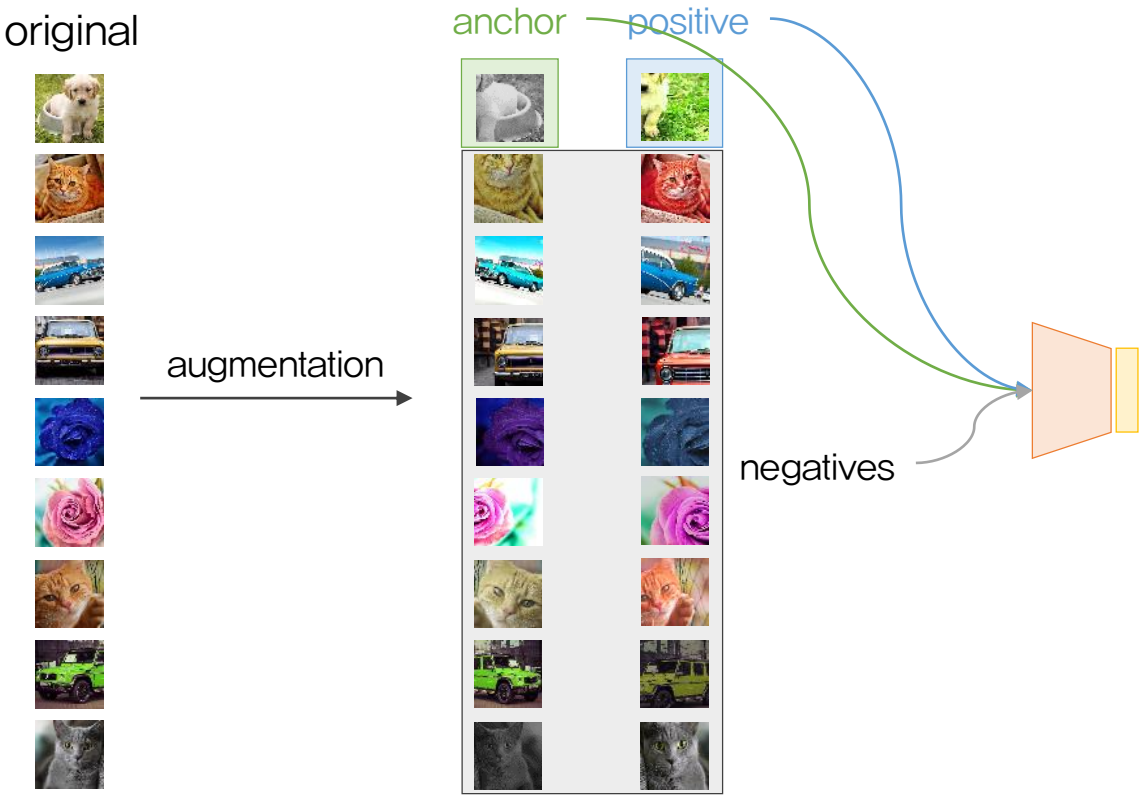
- Momentum Encoding
  - 과거 모델 weight들의 가중평균으로 업데이트
  - 과거의 나 자신들이 'teacher'가 되는 mean teacher 방식과 동일 (semi-supervised model)
  - Exponential moving average, momentum encoding, on-the-fly ensemble, mean teacher 등등 다양한 이름으로 사용



Tarvainen, A., & Valpola, H. (2017, December). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 1195-1204).

02 SimCLR

- Queue도 사용하지 않고 batch size를 늘려서 negative 생성을 batch 안에서 해결해보자
- 관측치별로 2번의 data augmentation을 적용



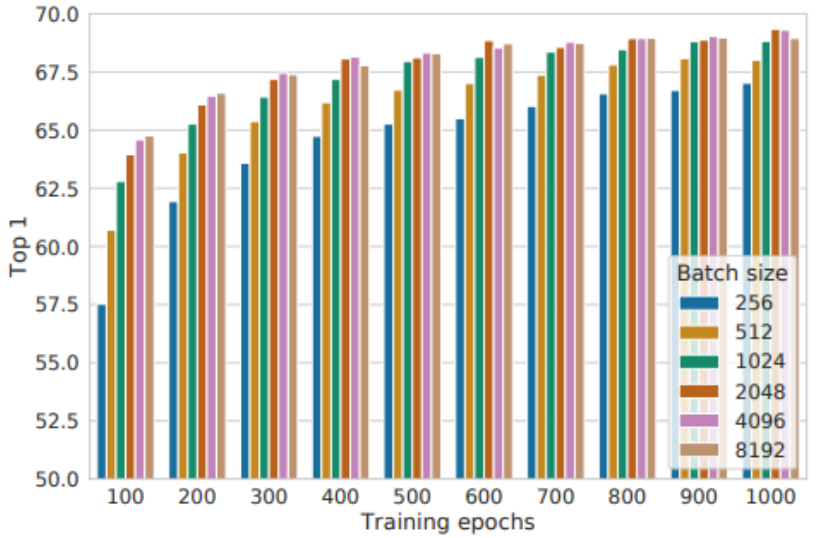


Figure 9. Linear evaluation models (ResNet-50) trained with different batch size and epochs. Each bar is a single run from scratch.[10]

Chen, X., Fan, H., Girshick, R., & He, K. (2020). Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297.

02 | SimCLR

- 다양한 data augmentation 기법을 활용 (Augmentation: "The whys and hows of data augmentation, 강현구" 세미나 참고)
  - ✓ Robust data distribution
- 더 좋은 feature를 만들기 위해 embedding layer를 더 깊게 쌓자

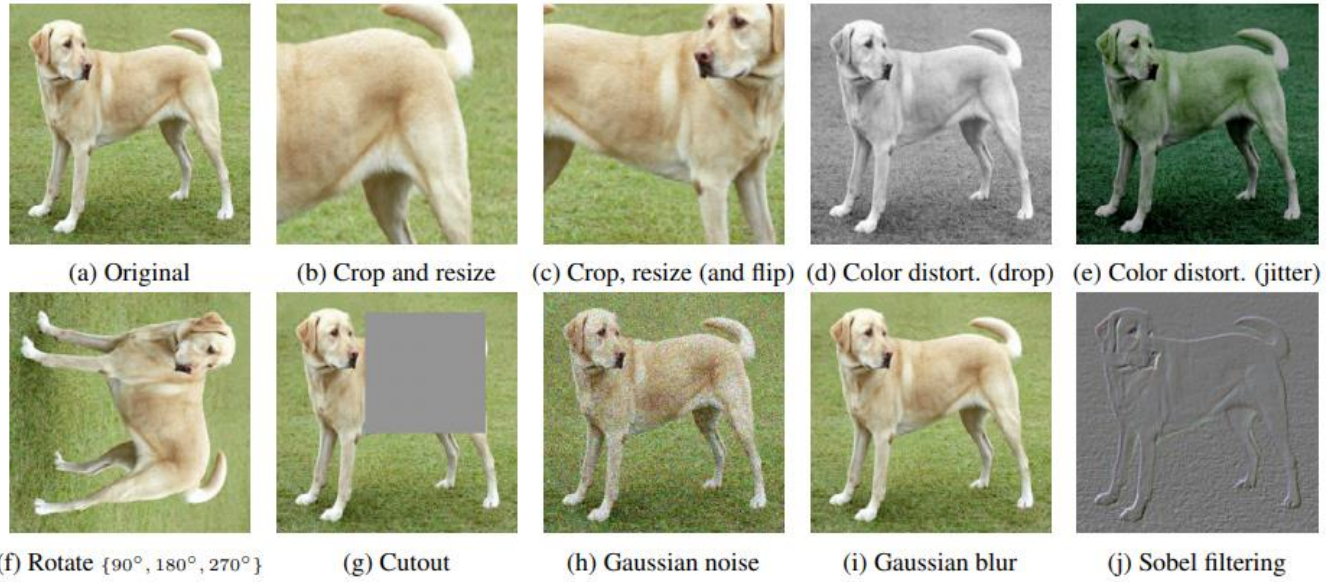
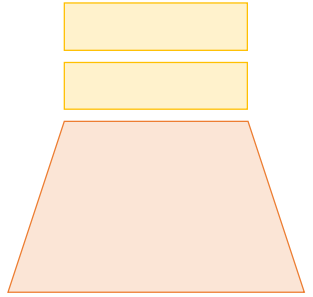


Figure 4. Illustrations of the studied data augmentation operators. Each augmentation can transform data stochastically with some internal parameters (e.g. rotation degree, noise level). Note that we *only* test these operators in ablation, the *augmentation policy used to train our models* only includes *random crop (with flip and resize)*, *color distortion*, and *Gaussian blur*. (Original image cc-by: Von.grzanka)



Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.

02 | MoCo v2

- SimCLR에서 사용한 깊은 embedding layer & data augmentation 기법을 MoCo에 적용
  - ✓ Facebook vs. Google
- 높은 분류 정확도를 보임 (Table 2)
- Batch size를 엄청 크게 사용하는 SimCLR보다 메모리 효율 측면에서 좋음 (Table 3)

case	unsup. pre-train MLP	aug+	cos	epochs	batch	ImageNet acc.
MoCo v1 [6]				200	256	60.6
SimCLR [2]	✓	✓	✓	200	256	61.9
SimCLR [2]	✓	✓	✓	200	8192	66.6
MoCo v2	✓	✓	✓	200	256	**67.5**
*results of **longer** unsupervised training follow:*						
SimCLR [2]	✓	✓	✓	1000	4096	69.3
MoCo v2	✓	✓	✓	800	256	**71.1**

Table 2. **MoCo *vs.* SimCLR**: ImageNet linear classifier accuracy (**ResNet-50, 1-crop 224×224**), trained on features from unsupervised pre-training. "aug+" in SimCLR includes blur and stronger color distortion. SimCLR ablations are from Fig. 9 in [2] (we thank the authors for providing the numerical results).

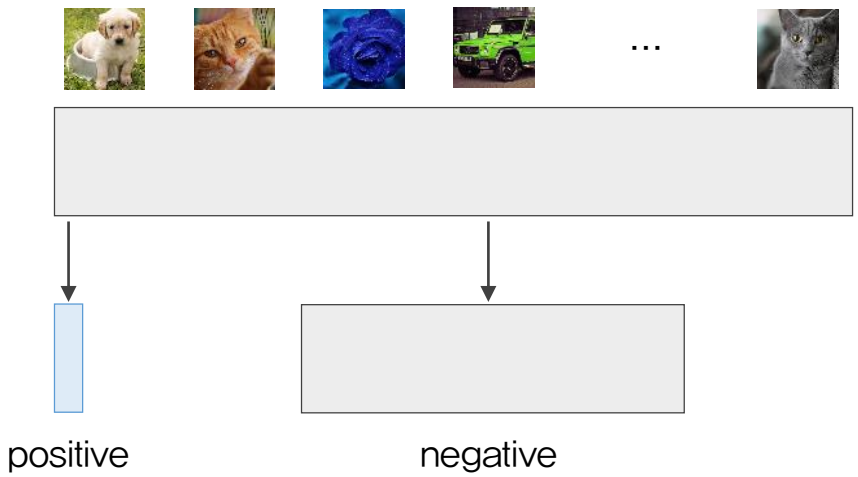
mechanism	batch	memory / GPU	time / 200-ep.
MoCo	256	**5.0G**	**53 hrs**
end-to-end	256	7.4G	65 hrs
end-to-end	4096	93.0G†	n/a

Table 3. **Memory and time cost** in 8 V100 16G GPUs, implemented in PyTorch. †: based on our estimation.

Chen, X., Fan, H., Girshick, R., & He, K. (2020). Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297.

03 Define Example Nicely

- 좀 더 좋은 positive를 정의할 수 없을까? 하나로 충분할까?
- 과연 적절한 negative일까? 더 좋은 샘플링 방법은 무엇일까?
  - ✓ Memory bank, queue, large batch – 3가지 방법 모두 결론적으로는 negative example 샘플링을 하는 것



03 Prototype to Support Positive

- 뭉쳐있는 유사한 여러 이미지를 대표할 수 있는 prototype (원형)을 positive로 사용하는 것
- 군집화 알고리즘(K-means)을 사용하여 prototype을 정의하고, contrastive learning을 수행
  - ✓ Fine-grained: $K = 100{,}000$
  - ✓ Coarse-grained: $K = 50{,}000$

} Hierarchical Semantic Structure

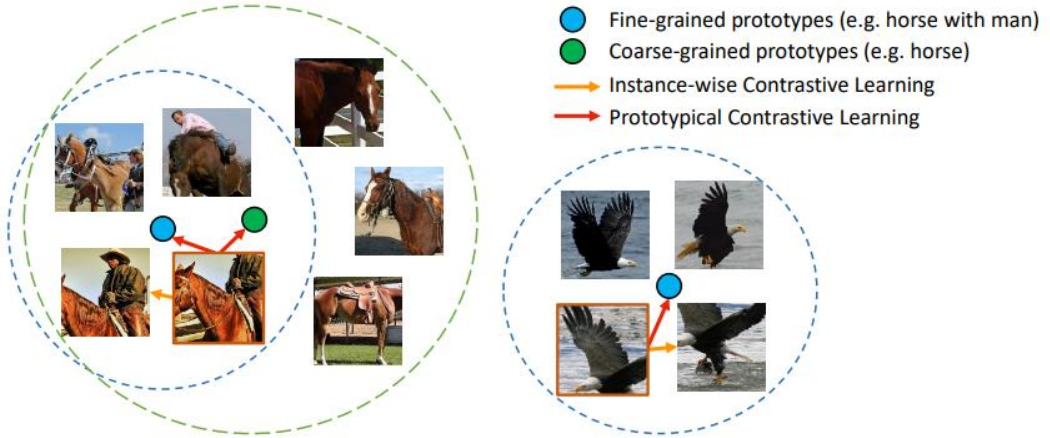


Figure 1: Illustration of Prototypical Contrastive Learning. Each instance is assigned to multiple prototypes with different granularity. PCL learns an embedding space which encodes the semantic structure of data.

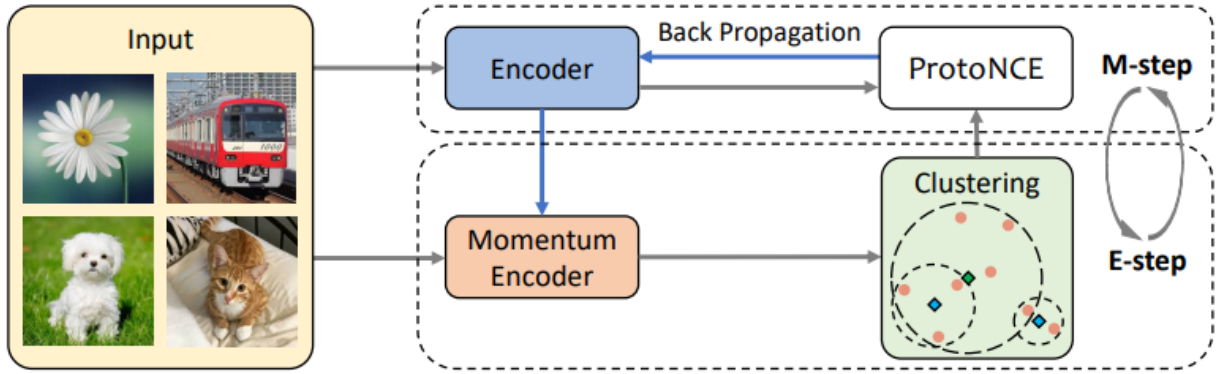
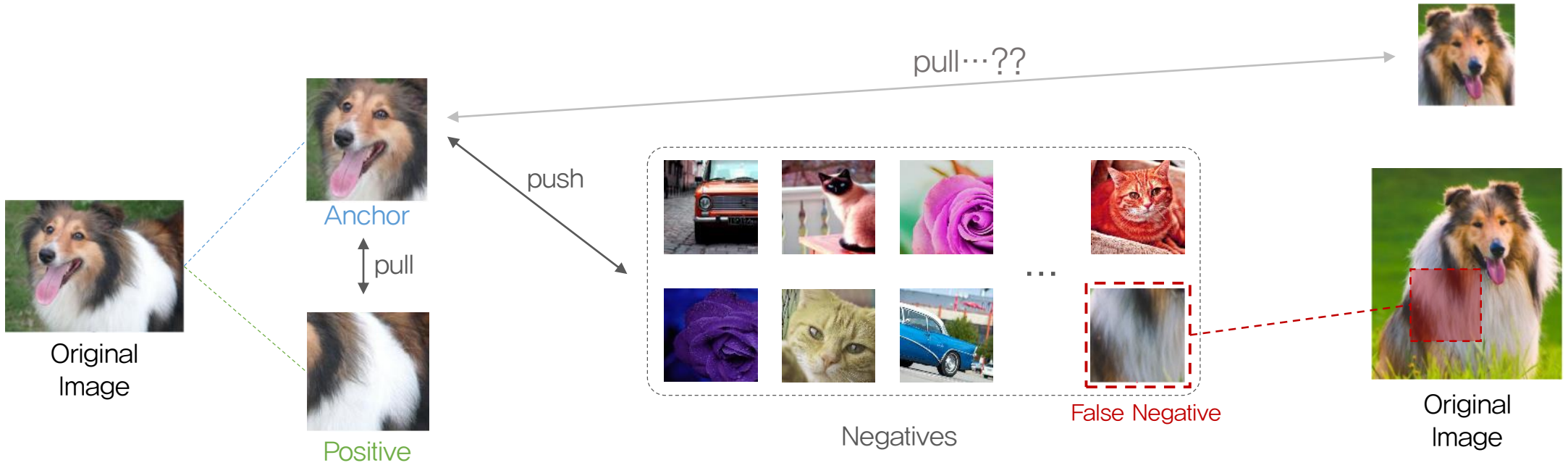


Figure 2: Training framework of Prototypical Contrastive Learning.

Li, J., Zhou, P., Xiong, C., Socher, R., & Hoi, S. C. (2020). Prototypical contrastive learning of unsupervised representations. arXiv preprint arXiv:2005.04966.

03 False Negative Cancellation

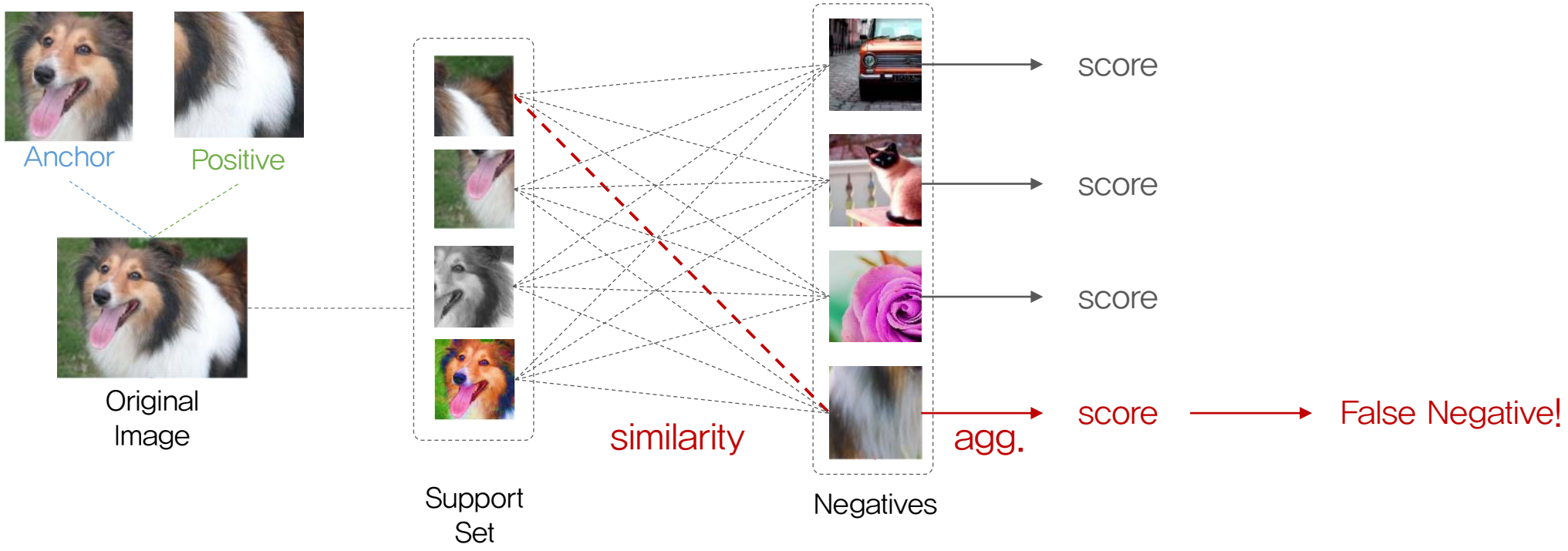
- Negative examples 중에서 anchor와 원래는 유사한 semantic structure를 갖는 이미지가 섞여 있을 수도 있지 않을까?
- False Negative란?
  - ✓ Anchor와 다른 이미지로부터 생성되어 negative로 정의되었으나, 원본 이미지 기준으로는 어느 정도의 semantic 유사성은 있는 이미지
  - ✓ Augmentation의 종류에 따라 anchor와 유사성이 낮게 나올 수 있기 때문에 True-False 구분 짓기 어려움



Huynh, T., Kornblith, S., Walter, M. R., Maire, M., & Khademi, M. (2020). Boosting Contrastive Self-Supervised Learning with False Negative Cancellation. arXiv preprint arXiv:2011.11765.

03 False Negative Cancellation

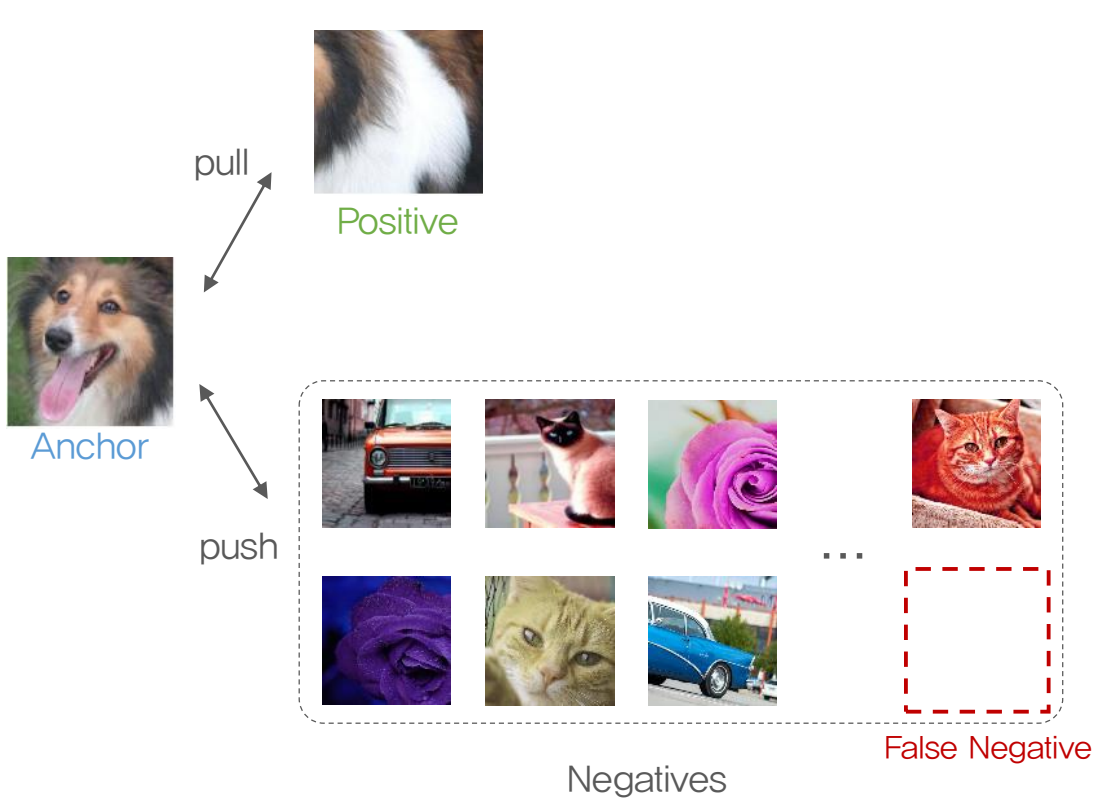
- 해결 방법
  - ✓ anchor에 augmentation을 여러 번 적용해서 negative examples과의 similarity score를 측정

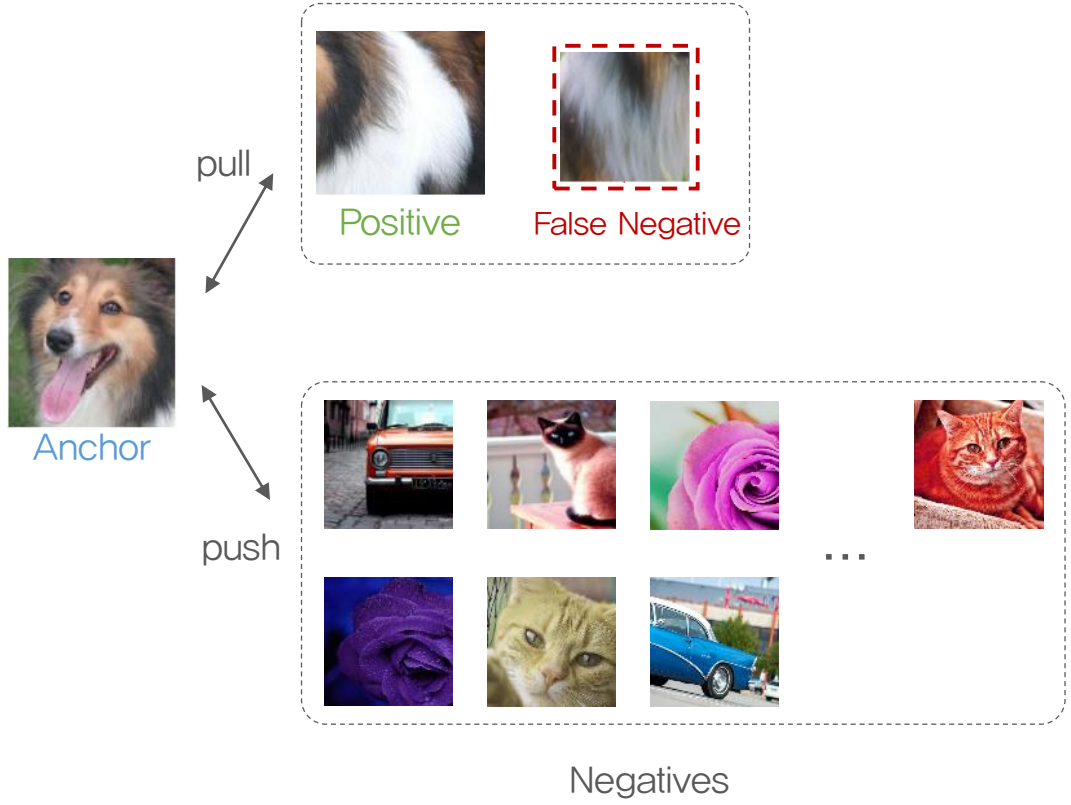


Huynh, T., Kornblith, S., Walter, M. R., Maire, M., & Khademi, M. (2020). Boosting Contrastive Self-Supervised Learning with False Negative Cancellation. arXiv preprint arXiv:2011.11765.

03 False Negative Cancellation

- 최종적으로 false negative를 제거하거나, positive에 포함시키는 전략을 사용할 수 있음

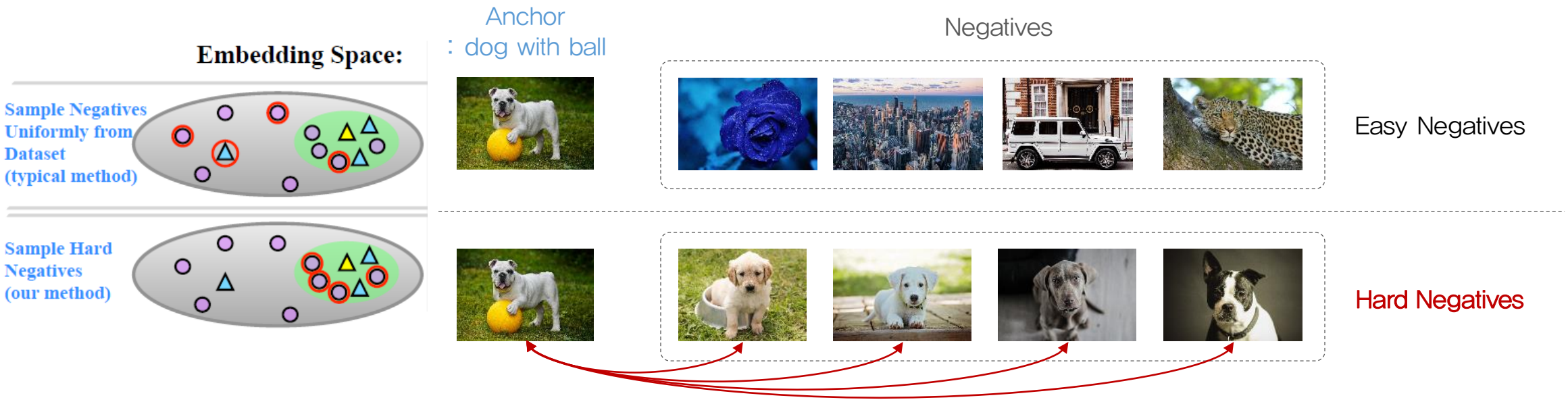




Huynh, T., Kornblith, S., Walter, M. R., Maire, M., & Khademi, M. (2020). Boosting Contrastive Self-Supervised Learning with False Negative Cancellation. arXiv preprint arXiv:2011.11765.

03 Sample Hard Negatives

- Instance discrimination을 잘 하기 위한 negative sample은 어디에 위치해 있어야 할까?
  - Anchor와 멀리 있는 easy negatives: 구분하기 쉽지만 instance discrimination을 위한 representation을 학습하기에는 부적합
  - Anchor와 가까이 있는 hard negatives를 선택해야 한다
  - 관측치별로 주변 density를 측정하여 샘플링에 반영



Robinson, J., Chuang, C. Y., Sra, S., & Jegelka, S. (2020). Contrastive learning with hard negative samples. arXiv preprint arXiv:2010.04592.

세미나 요약

1. We can learn sophisticated representation without label

2. Why NCE loss is exploited in contrastive learning

3. How to improve the contrastive learning?
   - ✓ Deepen network and use data augmentation: MoCo (Facebook) and SimCLR (Google)
   - ✓ Define example nicely: select hard, but not false negatives


Thank
you