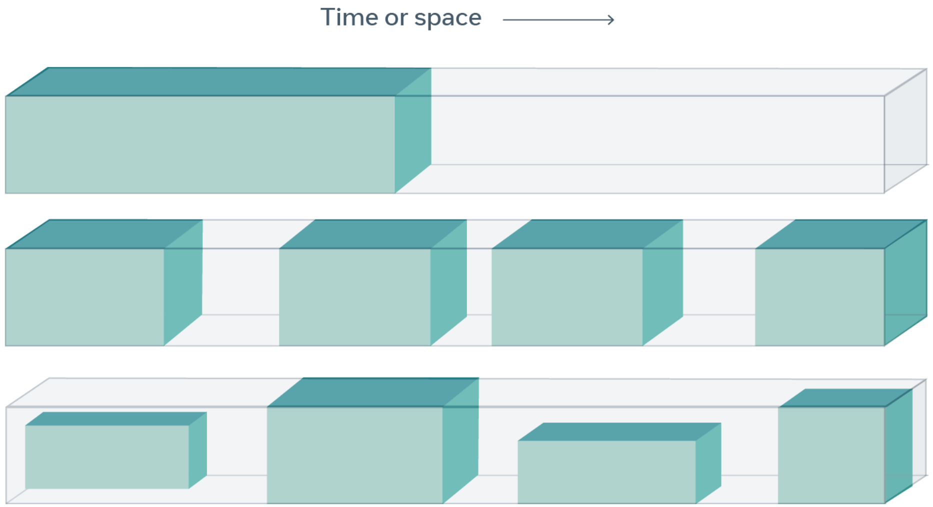
Contrastive Self-Supervised Learning

ML Reading Group – 9th Aug 2023

What’s covered in this talk?

- Introduction to self-supervised learning
- What is contrastive learning?
- Popular contrastive learning frameworks in computer vision
    - SimCLR
    - MoCo
- Multimodal contrastive learning
    - CLIP
- Future directions

Effectively Utilise Unlabeled Data with Self-Supervised Learning (SSL)



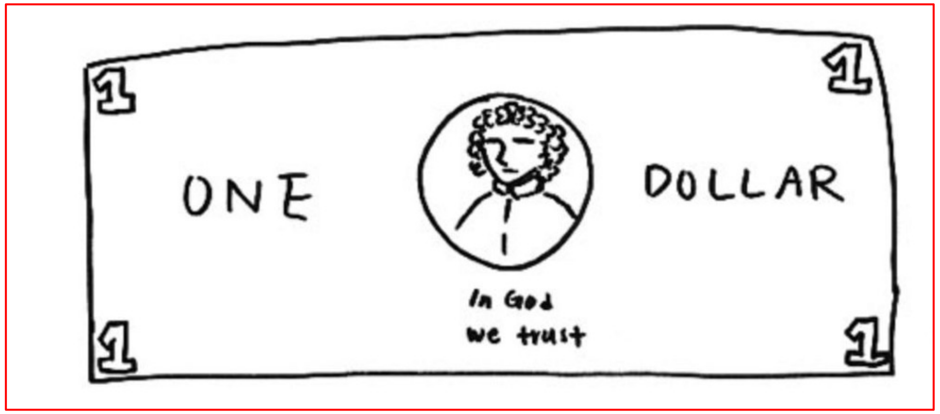
In SSL, model is trained to predict hidden parts of the input (in grey) from visible parts of the input using pre-text tasks

- Challenges in supervised learning
  - High cost of data annotation
  - Model is too specific to the trained task
  - Generalisation error
  - Spurious correlations

- Self supervised learning attempts to address above challenges by **learning from data w/o manual annotation**.

https://ai.meta.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/

SSL Intuition - The dollar bill experiment

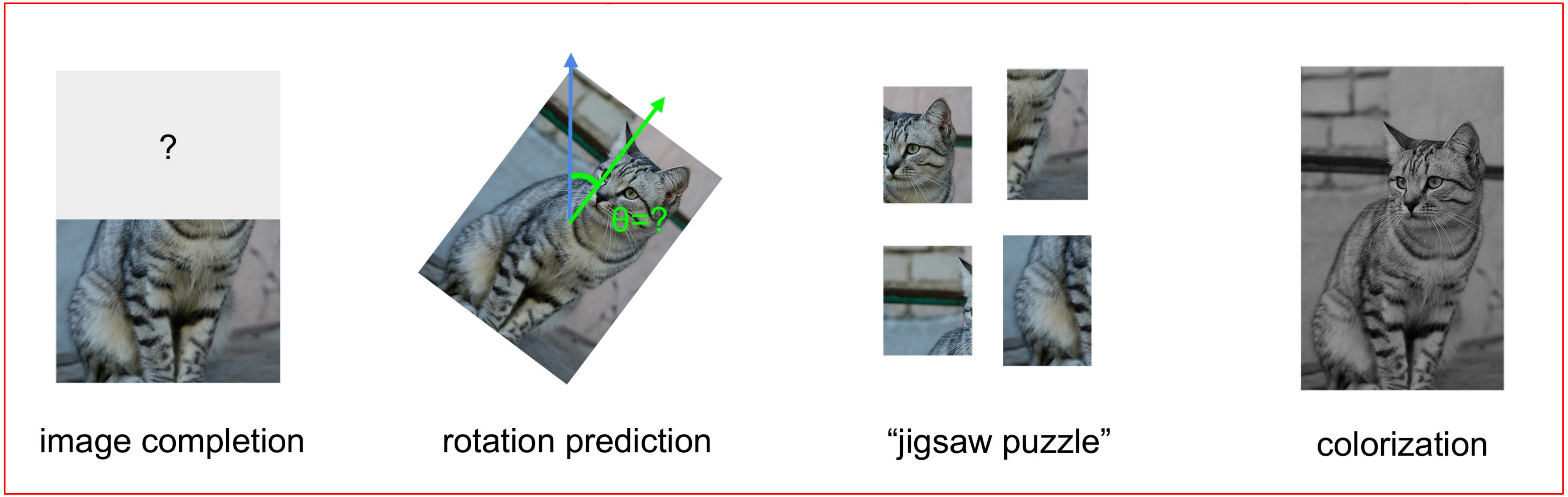


- Brain does not need complete information of a visual piece to differentiate from other
- It just needs rough representation of an image to discriminate from other.

http://cs231n.stanford.edu/2021/slides/2021/lecture_13.pdf
https://arxiv.org/pdf/2011.00362.pdf

SSL : Common Pretext Tasks in CV

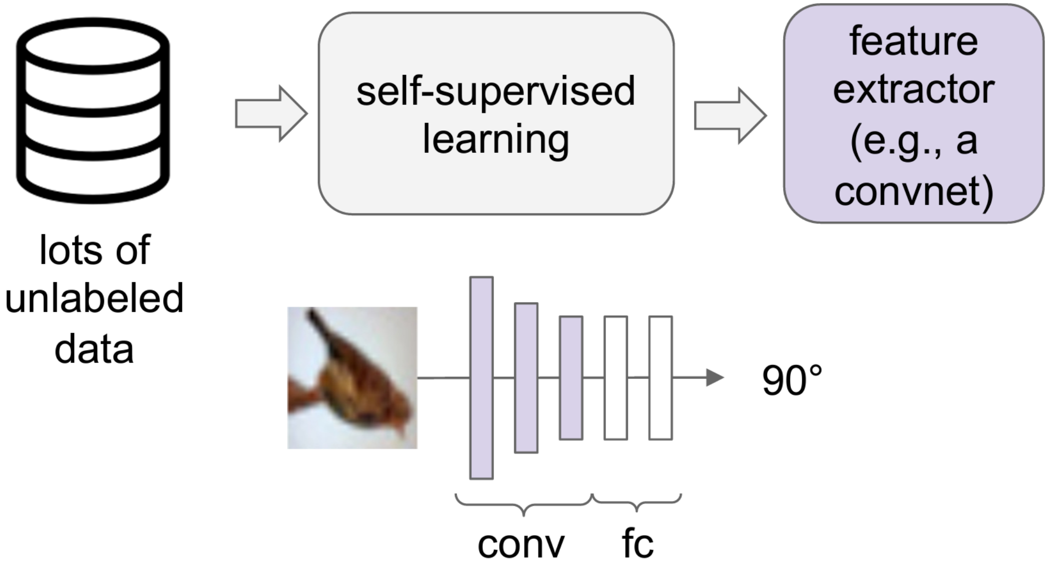
- SSL methods use “pretext” tasks to produce good features for downstream tasks



http://cs231n.stanford.edu/slides/2023/lecture_13.pdf

How to Evaluate a SSL Method?

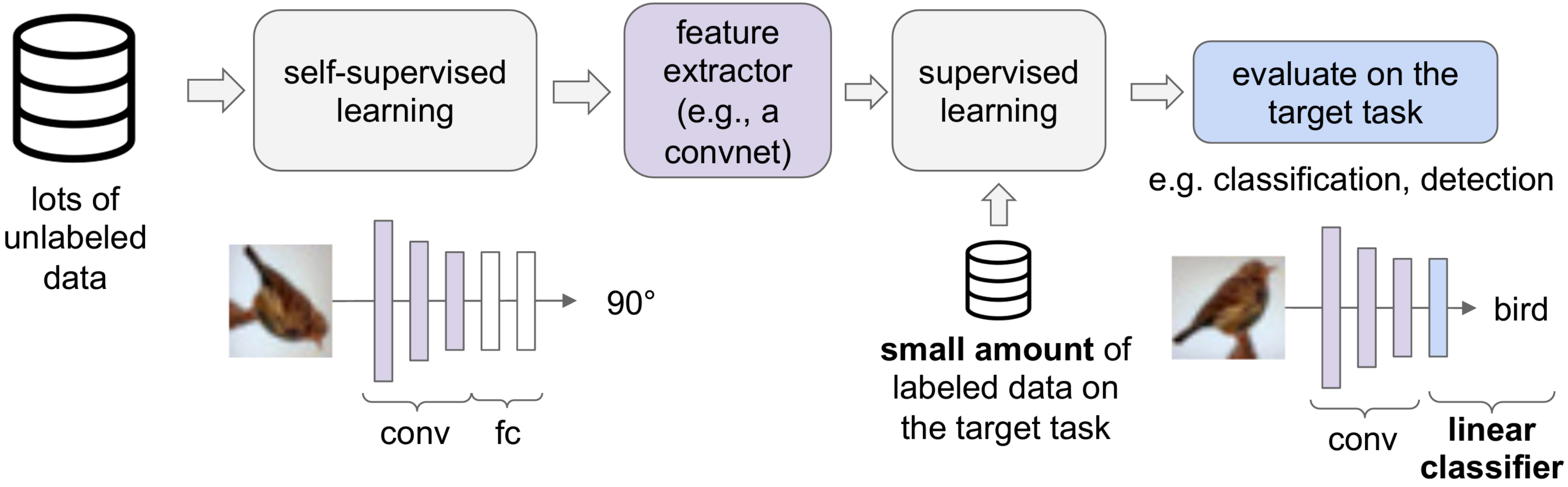
- Purpose of SSL method is to learn useful features for downstream tasks
  - It’s not necessary to get perfect score in pre-text tasks



http://cs231n.stanford.edu/slides/2023/lecture_13.pdf

How to Evaluate a SSL Method?

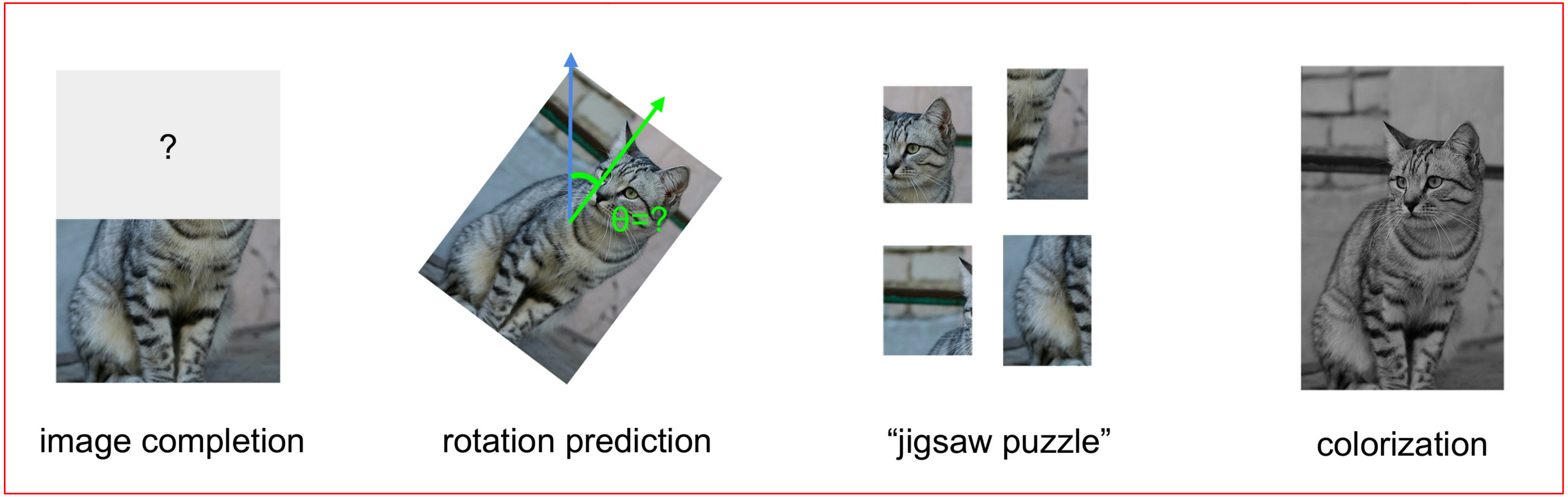
- Purpose of SSL method is to learn useful features for downstream tasks
  - It's not necessary to get perfect score in pre-text tasks



http://cs231n.stanford.edu/slides/2023/lecture_13.pdf

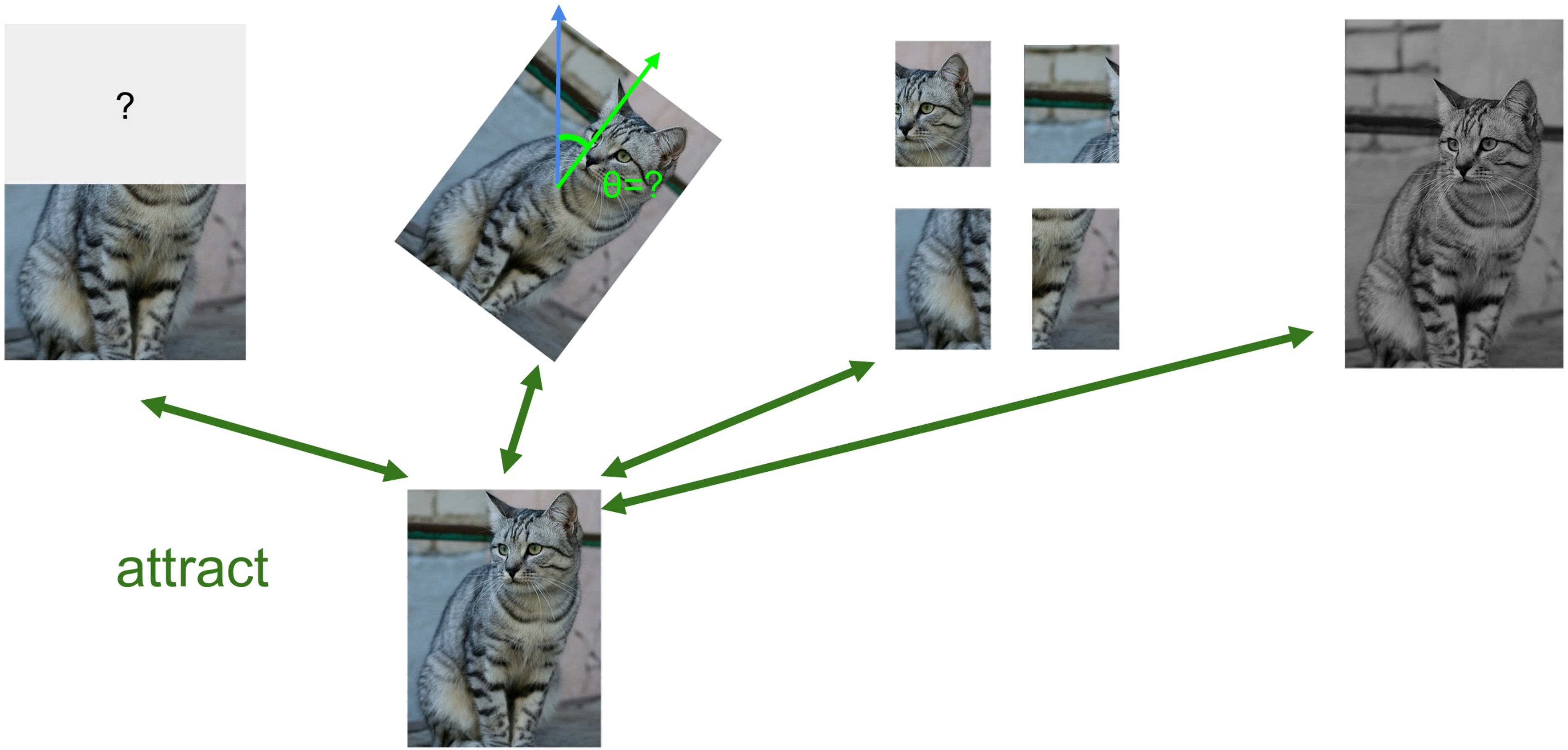
Can we do better than the common pretext tasks from image transformations?

- Risk of the learned features tied to specific pretext task?
  - How about a more general pretext task?



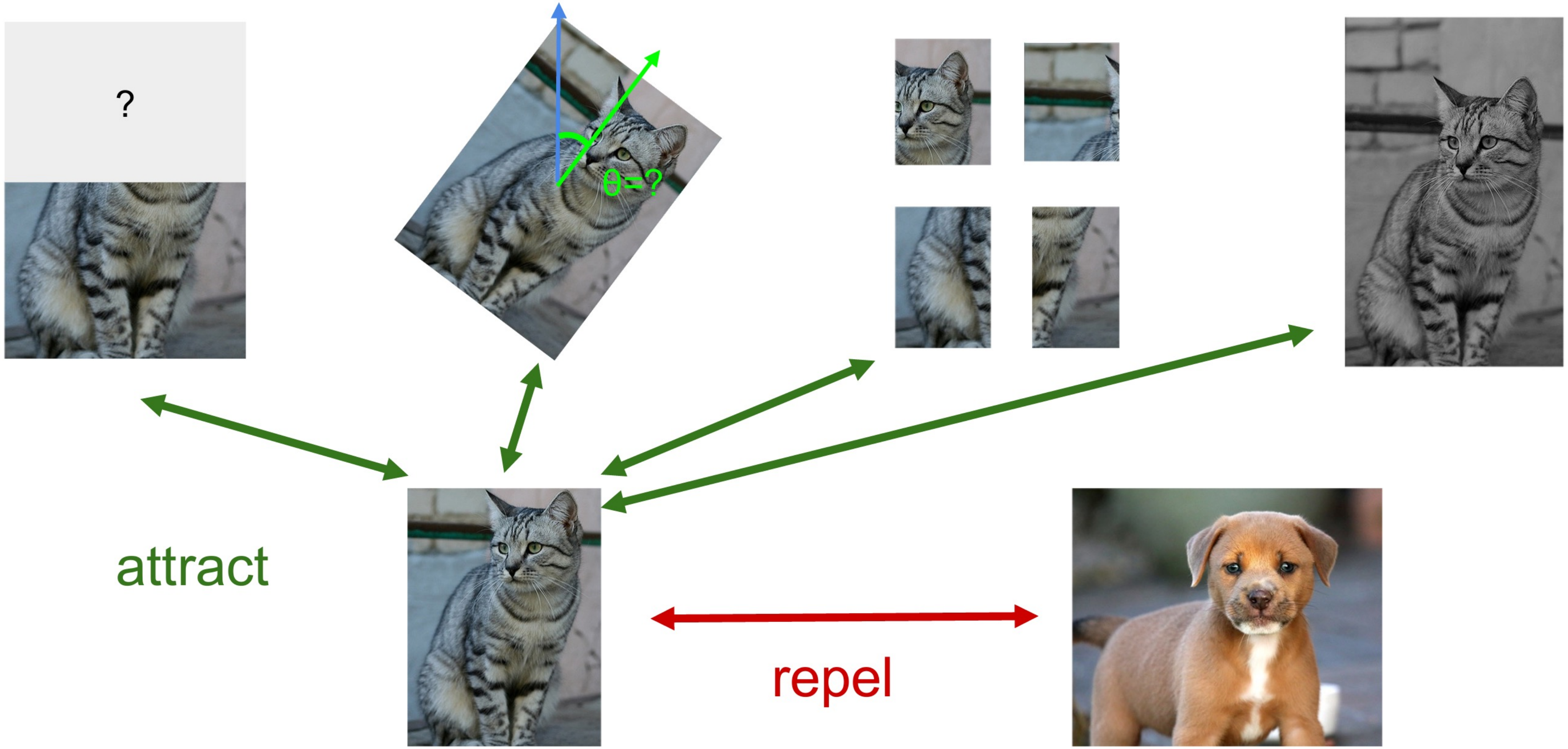
http://cs231n.stanford.edu/slides/2023/lecture_13.pdf

A more general pretext task?



http://cs231n.stanford.edu/slides/2023/lecture_13.pdf

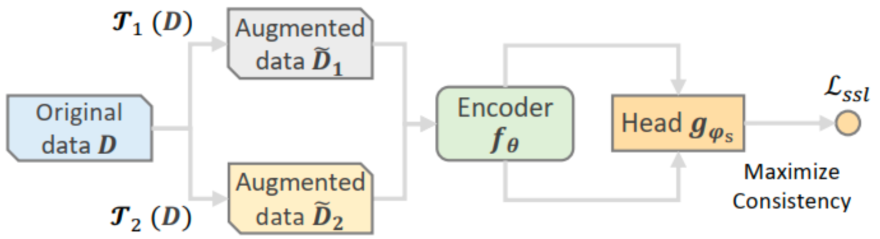
Contrastive representation learning!



http://cs231n.stanford.edu/slides/2023/lecture_13.pdf

Contrastive SSL : Definition and Training Objective

- The primary objective of contrastive learning is to ensure that representations of similar samples are brought closer together in the learned space, while representations of diverse or dissimilar samples are pushed farther apart.



$$f_\theta^* = \arg\min_{f_\theta, g_{\phi_s}} \boxed{\mathcal{L}_{ssl}}\Big(g_{\phi_s}(f_\theta(\boxed{\tilde{\mathcal{D}}_1}), f_\theta(\boxed{\tilde{\mathcal{D}}_2}))\Big)$$

Estimates the mutual information between views.

Views are created by imposing different transformations on the original data.

https://arxiv.org/pdf/2203.15876.pdf

Mathematical Intuition of Contrastive Learning

- Mathematically, contrastive learning aims to learn an encoder f for a given data point x, such that,

$$\text{score}(f(x), f(x^+)) >> \text{score}(f(x), f(x^-))$$

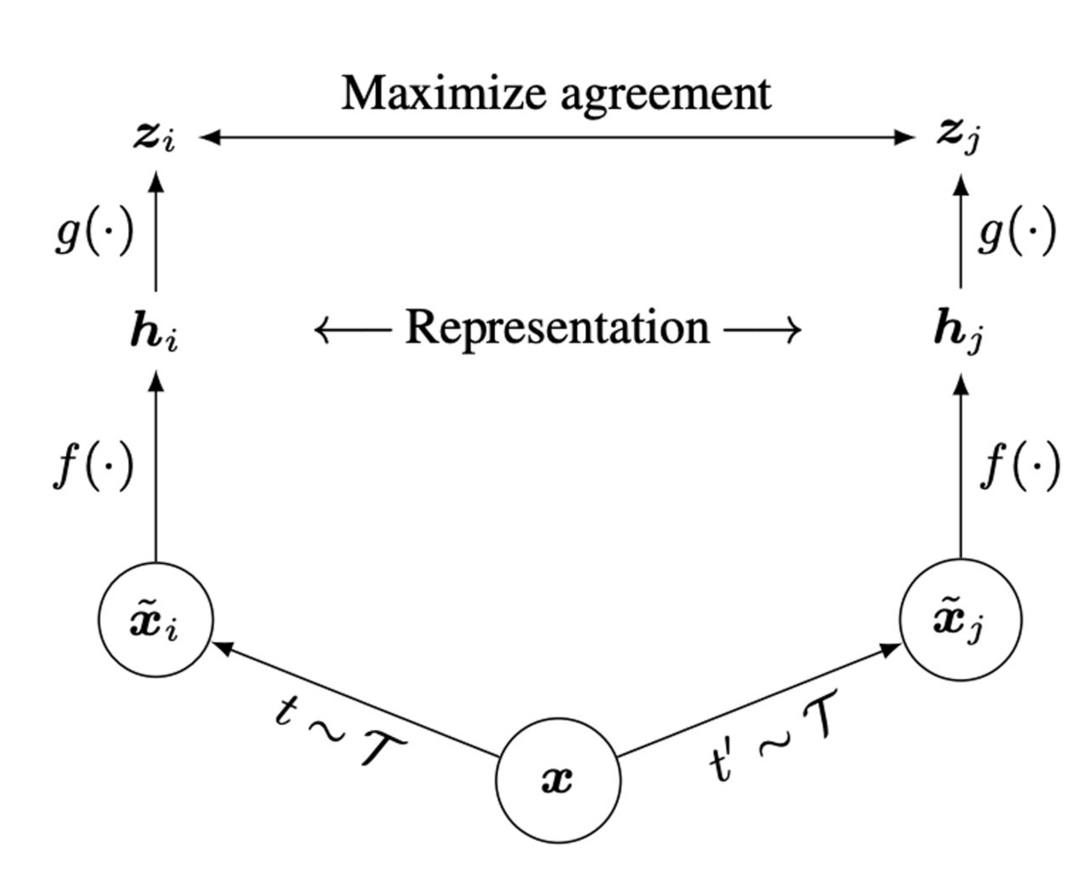
- For 1 positive and (N-1) negative samples, typical contrastive loss (N-way SoftMax classifier) :

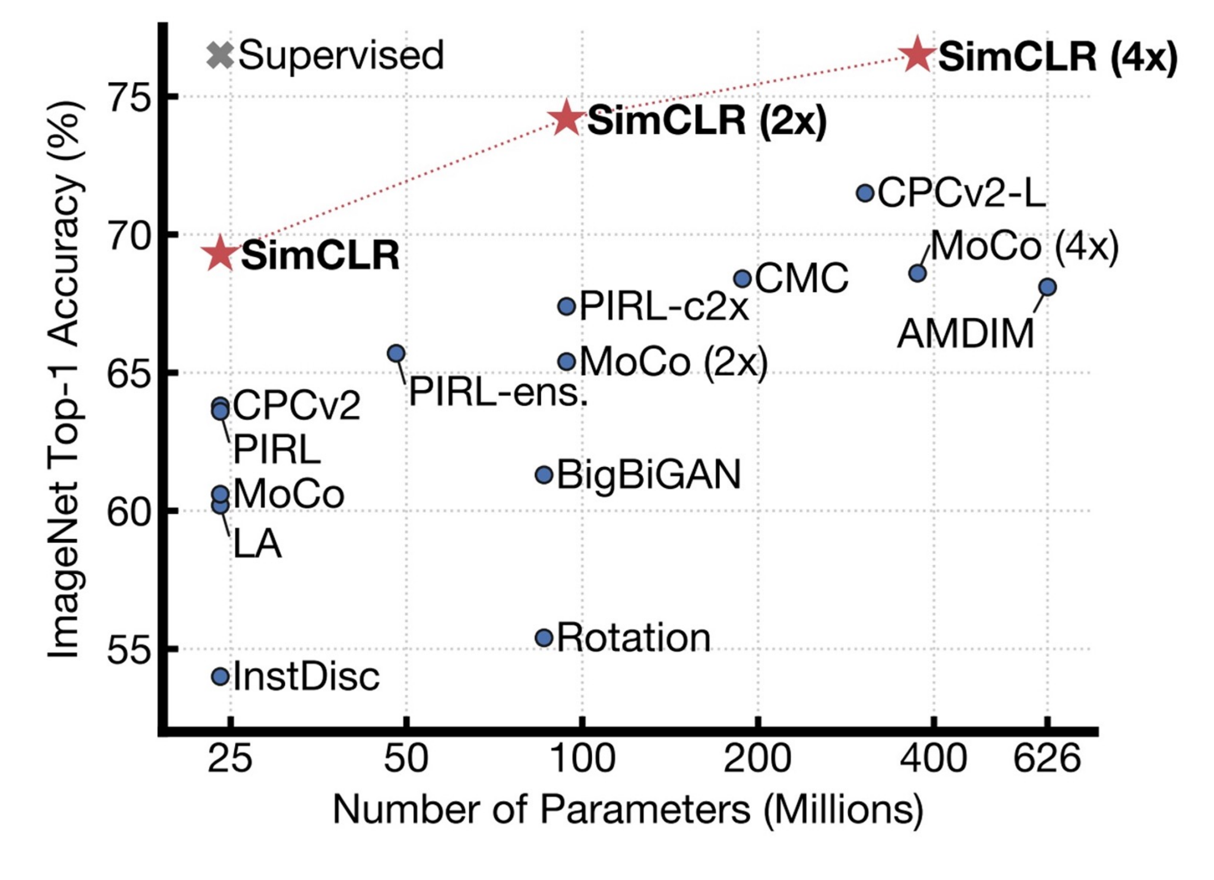
$$L = -\mathbb{E}_X \left[\log \frac{\exp(s(f(x), f(x^+)))}{\exp(s(f(x), f(x^+))) + \sum_{j=1}^{N-1} \exp(s(f(x), f(x_j^-)))} \right]$$

http://cs231n.stanford.edu/slides/2023/lecture_13.pdf

Popular Contrastive Learning Frameworks in CV

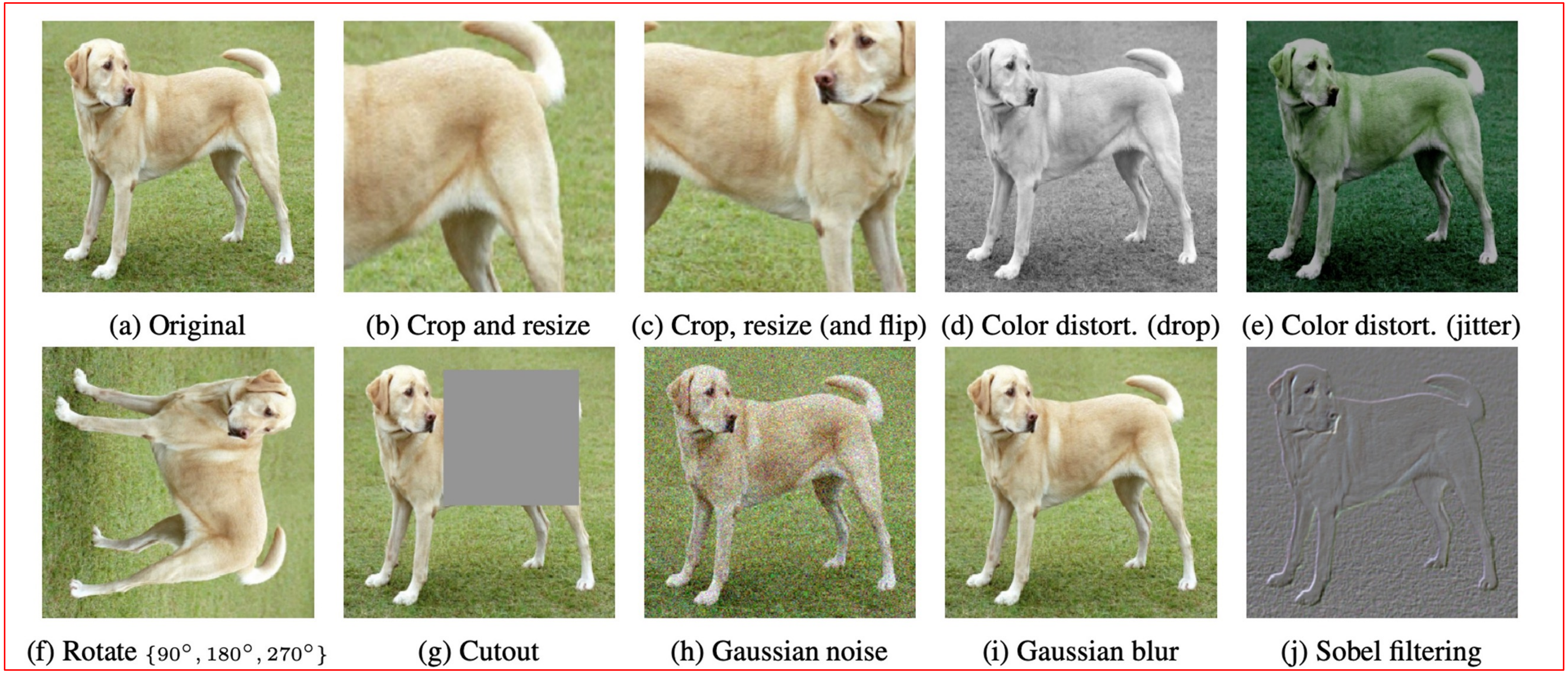
SimCLR – A Simple Framework for Contrastive Learning of Visual Representations (2020)





https://arxiv.org/pdf/2002.05709.pdf

SimCLR – Positive samples were generated from these data augmentation methods



https://arxiv.org/pdf/2002.05709.pdf

SimCLR : Main Learning Algorithm

Algorithm 1 SimCLR's main learning algorithm.

input: batch size N, constant τ, structure of f, g, $\mathcal{T}$.
for sampled minibatch $\{\boldsymbol{x}_k\}_{k=1}^{N}$ **do**
  for all $k \in \{1, \ldots, N\}$ **do**
    $\boxed{\text{draw two augmentation functions } t \sim \mathcal{T}, t' \sim \mathcal{T}}$
    # the first augmentation
    $\boxed{\tilde{\boldsymbol{x}}_{2k-1} = t(\boldsymbol{x}_k)}$
    $\boldsymbol{h}_{2k-1} = f(\tilde{\boldsymbol{x}}_{2k-1})$ # representation
    $\boldsymbol{z}_{2k-1} = g(\boldsymbol{h}_{2k-1})$ # projection
    # the second augmentation
    $\boxed{\tilde{\boldsymbol{x}}_{2k} = t'(\boldsymbol{x}_k)}$
    $\boldsymbol{h}_{2k} = f(\tilde{\boldsymbol{x}}_{2k})$ # representation
    $\boldsymbol{z}_{2k} = g(\boldsymbol{h}_{2k})$ # projection
  end for
  for all $i \in \{1, \ldots, 2N\}$ and $j \in \{1, \ldots, 2N\}$ **do**
    $s_{i,j} = \boldsymbol{z}_i^\top \boldsymbol{z}_j / (\|\boldsymbol{z}_i\| \|\boldsymbol{z}_j\|)$ # pairwise similarity
  end for
  define $\ell(i,j)$ **as** $\boxed{\ell(i,j) = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(s_{i,k}/\tau)}}$
  $\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]$
  update networks f and g to minimize $\mathcal{L}$
end for
return encoder network $f(\cdot)$, and throw away $g(\cdot)$

Generate a positive pair by sampling data augmentation functions

Iterate through and use each of the 2N sample as reference, compute average loss

InfoNCE loss:
Use all non-positive samples in the batch as x^-

https://arxiv.org/pdf/2002.05709.pdf
http://cs231n.stanford.edu/slides/2023/lecture_13.pdf

SimCLR : Performance on ImageNet on Semi-Supervised Setup

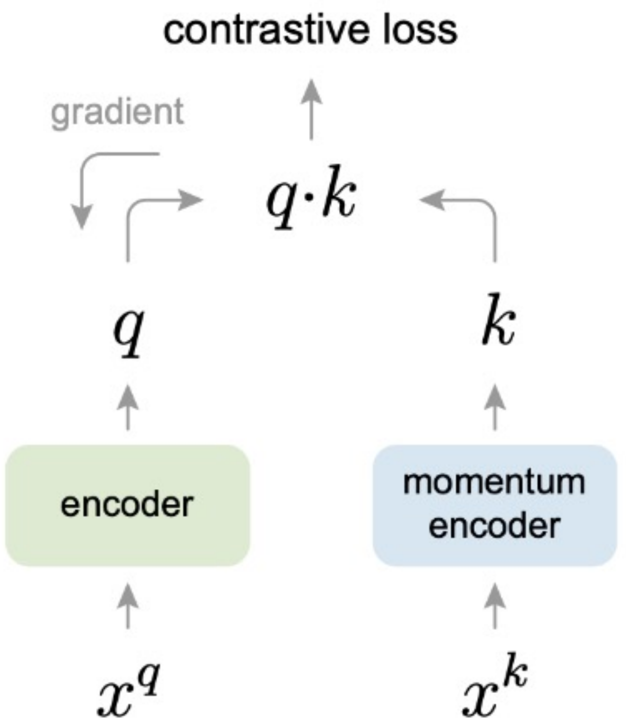
Method	Architecture	Label fraction 1% Top 5	Label fraction 10% Top 5
Supervised baseline	ResNet-50	48.4	80.4
Methods using other label-propagation:			
Pseudo-label	ResNet-50	51.6	82.4
VAT+Entropy Min.	ResNet-50	47.0	83.4
UDA (w. RandAug)	ResNet-50	-	88.5
FixMatch (w. RandAug)	ResNet-50	-	89.1
S4L (Rot+VAT+En. M.)	ResNet-50 (4×)	-	91.2
Methods using representation learning only:			
InstDisc	ResNet-50	39.2	77.4
BigBiGAN	RevNet-50 (4×)	55.2	78.8
PIRL	ResNet-50	57.2	83.8
CPC v2	ResNet-161(∗)	77.9	91.2
SimCLR (ours)	ResNet-50	75.5	87.8
SimCLR (ours)	ResNet-50 (2×)	83.0	91.2
SimCLR (ours)	ResNet-50 (4×)	**85.8**	**92.6**

Table 7. ImageNet accuracy of models trained with few labels.

- The feature encoder was trained on full ImageNet data with the proposed SimCLR framework.
- The model was finetuned with 1% and 10% labeled data from ImageNet respectively.

https://arxiv.org/pdf/2002.05709.pdf
http://cs231n.stanford.edu/slides/2023/lecture_13.pdf

MoCo :
Momentum Contrast for Unsupervised Visual Representation Learning



(c) MoCo

- Contrastive Learning is benefitted by large number of negative examples
- Though, mini-batch size can restrict the number of negative samples
- MoCo addresses this by maintaining a large queue of negative samples and updating the negative encoder weights with momentum update instead of backpropagation!

https://ankeshanand.com/blog/2020/01/26/contrative-self-supervised-learning.html
https://arxiv.org/pdf/1911.05722.pdf

MoCo : Main Algorithm

Generate a positive pair by sampling data augmentation functions

No gradient through the key

Algorithm 1 Pseudocode of MoCo in a PyTorch-like style.

```
# f_q, f_k: encoder networks for query and key
# queue: dictionary as a queue of K keys (CxK)
# m: momentum
# t: temperature

f_k.params = f_q.params # initialize
for x in loader: # load a minibatch x with N samples
   x_q = aug(x) # a randomly augmented version
   x_k = aug(x) # another randomly augmented version

   q = f_q.forward(x_q) # queries: NxC
   k = f_k.forward(x_k) # keys: NxC
   k = k.detach() # no gradient to keys

   # positive logits: Nx1
   l_pos = bmm(q.view(N,1,C), k.view(N,C,1))

   # negative logits: NxK
   l_neg = mm(q.view(N,C), queue.view(C,K))

   # logits: Nx(1+K)
   logits = cat([l_pos, l_neg], dim=1)

   # contrastive loss, Eqn.(1)
   labels = zeros(N) # positives are the 0-th
   loss = CrossEntropyLoss(logits/t, labels)

   # SGD update: query network
   loss.backward()
   update(f_q.params)

   # momentum update: key network
   f_k.params = m*f_k.params+(1-m)*f_q.params

   # update dictionary
   enqueue(queue, k) # enqueue the current minibatch
   dequeue(queue) # dequeue the earliest minibatch
```

bmm: batch matrix multiplication; mm: matrix multiplication; cat: concatenation.

Use the running queue of keys as the negative samples

InfoNCE loss

Update f_k through momentum

Update the FIFO negative sample queue

https://arxiv.org/pdf/1911.05722.pdf

MoCo v2

- MoCo v2 utilized best of both MoCo v1 and SimCLR.
  - **From SimCLR**: non-linear projection head and strong data augmentation.
  - **From MoCo v1**: momentum-updated queues that allow training on a large number of negative samples (no TPU required!).

Let’s see how it performed against SimCLR.

http://cs231n.stanford.edu/slides/2023/lecture_13.pdf

MoCo v2 vs SimCLR

- MoCo v2 outperforms SimCLR with smaller batch size

case	MLP	aug+	cos	epochs	batch	ImageNet acc.
	unsup. pre-train					
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).

- MoCo v2 has a smaller memory footprint!

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.

https://arxiv.org/pdf/2003.04297.pdf

Multimodal Contrastive Learning

CLIP : Contrastive Language–Image Pre-training

- Trains contrastive pre-training model on image and text data
- It achieves competitive zero-shot performance on a variety of image classification datasets

Food101

guacamole (90.1%) Ranked 1 out of 101 labels

✓ a photo of **guacamole**, a type of food.

✕ a photo of **ceviche**, a type of food.

✕ a photo of **edamame**, a type of food.

✕ a photo of **tuna tartare**, a type of food.

✕ a photo of **hummus**, a type of food.

Stanford Cars

2012 Honda Accord Coupe (63.3%) Ranked 1 out of 196 labels

✓ a photo of a **2012 honda accord coupe**.

✕ a photo of a **2012 honda accord sedan**.

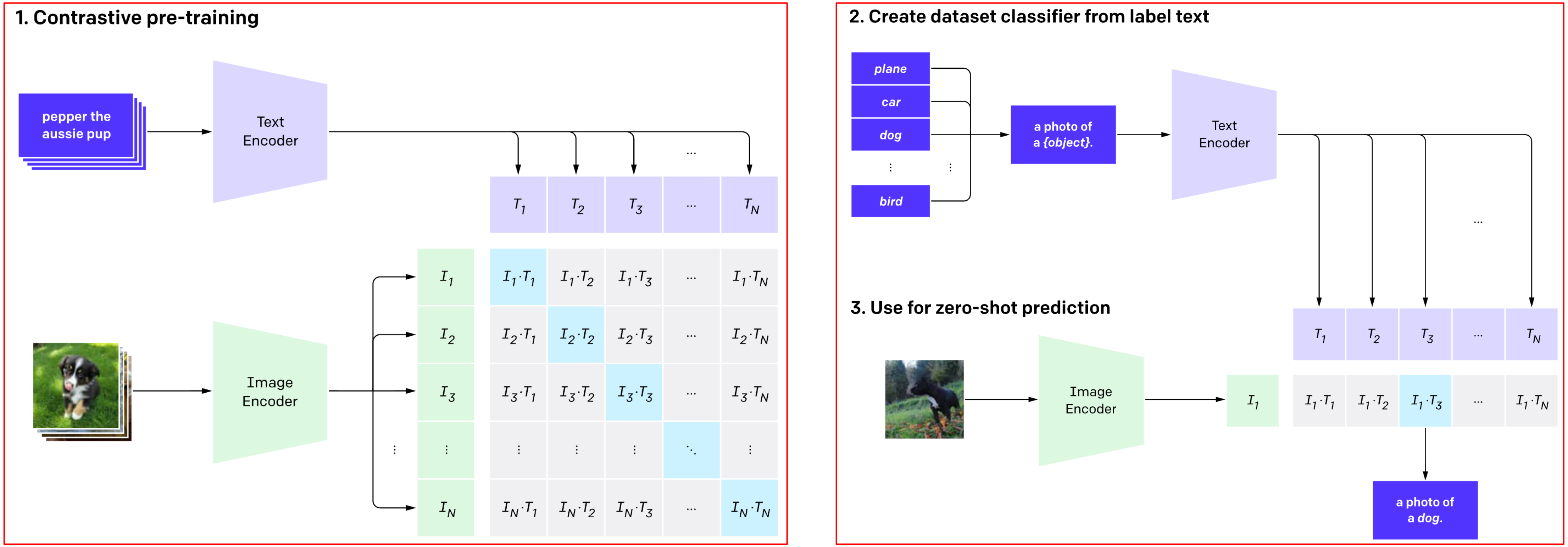
✕ a photo of a **2012 acura tl sedan**.

✕ a photo of a **2012 acura tsx sedan**.

✕ a photo of a **2008 acura tl type-s**.

https://openai.com/research/clip

CLIP : Approach



https://openai.com/research/clip

Future Directions

- How to effectively apply contrastive approaches to different modalities like time-series, graph, etc?
    - Time-series and graph data are more heterogenous compared to image data.
- How to use contrastive methods in multi-modal data?

Thank you for your attention!
Any questions?