

限られたデータからの深層学習

東京工業大学 情報理工学院 助教
井上 中順

はじめに



- ・限られたデータからの深層学習

第1部 背景と問題設定

- 様々な学習の枠組みの整理

第2部 Self-Supervised Learning

- ラベルなしデータでの事前学習

第3部 Formula-driven Approach

- 自然画像なしでの事前学習



自己紹介



現職：東京工業大学 情報理工学院 助教

略歴

2014.3 博士(工学)取得, 指導教員: 篠田浩一教授

2014.4 現職

2019.4 研究室設立 (テニュアトラック)

※ 学生募集中です！

専門

マルチメディア情報処理, 映像認識, 音声認識, 画像認識



第1部 背景と問題設定

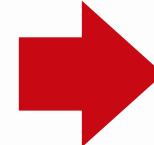
背景



近年：深層学習技術の発展、様々な分野へのAI応用が加速

基礎研究

物体認識、動作認識、
シーン認識など



応用展開

情報検索、ロボティクス、医療
自動運転、セキュリティなど

課題：ネットワークの学習には大量の教師付きデータが必要

教師付きデータ

教師あり学習には データ と 正解ラベル の準備が必要

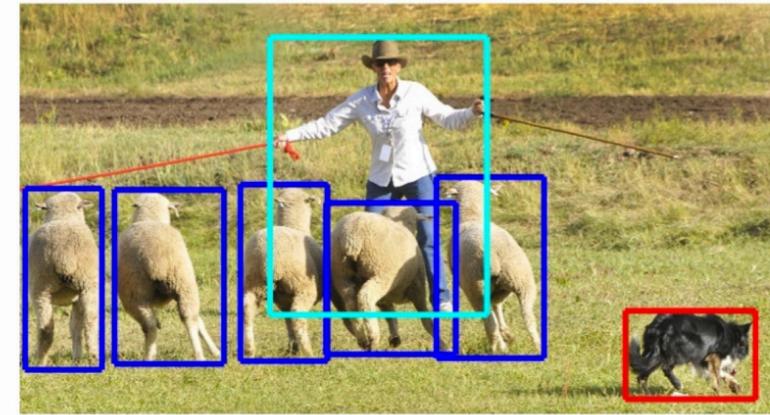
例) 物体検出

データ

画像

正解ラベル

物体の矩形位置



※ 画像は MS COCO Datasetより

問題点：大量の正解ラベル付与には多大なコストがかかる

限られたデータ

実応用ではデータやラベルが限られていることが多い

例) 監視カメラ・車載カメラ

→ 大量の映像はあるが、ラベルがないものが多い

医療・福祉

→ 記録は残っているが、症例が少ないものが多くある

Webサービス

→ ユーザーデータはあるが、曖昧なタグが多い

学習フレームワーク



様々な条件下での「学習」が提案されている

教師あり
(Supervised)

教師なし
(Unsupervised)

半教師あり 弱教師あり 自己教師あり
(Semi-Supervised) (Weakly-Supervised) (Self-Supervised)

Few-Shot

Zero-Shot

メタ学習
(Meta)

距離学習
(Metric)

対照学習
(Contrastive)

継続学習
(Continual)

強化学習
(Reinforcement)

限られたデータからの学習



様々な条件下での「学習」が提案されている

ラベルの量	Supervised	Semi-supervised	Unsupervised
データの量	(Many-Shot)	Few-Shot	Zero-Shot
ラベルの完全さ	Strongly-	Weakly-	Randomly-
ラベル付与方式	(Manually-)	Self-supervised	
実データの有無	Data-driven	Formula-driven?	

限られたデータからの学習

問題設定
を解説

様々な条件下での「学習」が提案されている

ラベルの量	Supervised	Semi-supervised	Unsupervised
データの量	(Many-Shot)	Few-Shot	Zero-Shot
ラベルの完全さ	Strongly-	Weakly-	Randomly-
ラベル付与方式	(Manually-)	Self-supervised	
実データの有無	Data-driven	Formula-driven?	

今日の本題

限られたデータからの学習



様々な条件下での「学習」が提案されている

ラベルの量	Supervised	Semi-supervised	Unsupervised
データの量	(Many-Shot)	Few-Shot	Zero-Shot
ラベルの完全さ	Strongly-	Weakly-	Randomly-
ラベル付与方式	(Manually-)	Self-supervised	
実データの有無	Data-driven	Formula-driven?	

最初に問題設定を解説

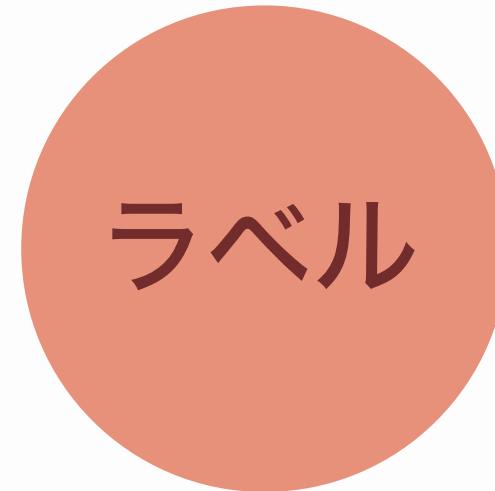
Supervised Learning



全ての画像にラベルが付与されている状況



+



Supervised Learning

全ての画像にラベルが付与されている状況



画像

y_i : motorbike

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ \textcolor{red}{1} \\ 0 \\ 0 \end{pmatrix}$$

airplane
bus
car
motorbike
train
truck

ラベルの例: 物体名

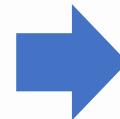
Supervised Learning

全ての画像にラベルが付与されている状況

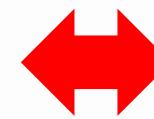


画像

$f(x_i)$



$$\begin{pmatrix} -1.61 \\ 2.90 \\ -0.50 \\ 0.11 \\ 0.52 \\ \vdots \end{pmatrix} \rightarrow \begin{pmatrix} 0.01 \\ 0.03 \\ 0.02 \\ 0.91 \\ 0.01 \\ 0.02 \end{pmatrix}$$



y_i : motorbike

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

airplane
bus
car
motorbike
train
truck

(確率)

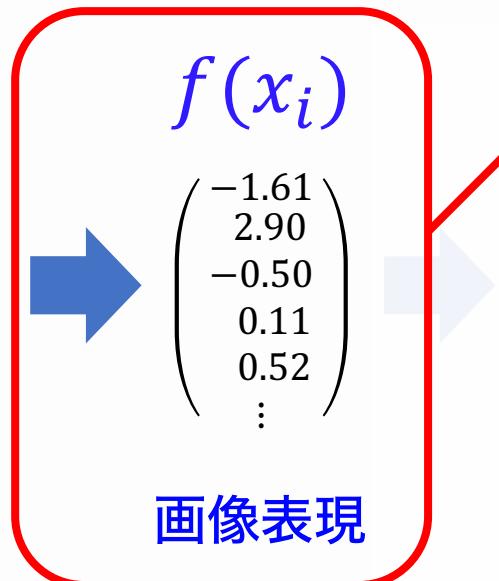
ラベルの例: 物体名

Supervised Learning

全ての画像にラベルが付与されている状況



画像



多層のニューラルネット
例) ResNet
EfficientNetV2
Vision Transformer
(ViT, DeiT, etc.)
gMLP, MLP-Mixer

Supervised Learning

全ての画像にラベルが付与されている状況

1層～数層のヘッダ (x_i)

例) Linear \circ Softmax

2層MLP \circ Softmax

画像

画像表現

(確率)

$$\begin{pmatrix} 0.01 \\ 0.03 \\ 0.02 \\ 0.91 \\ 0.01 \\ 0.02 \end{pmatrix}$$

y_i : motorbike

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

airplane
bus
car
motorbike
train
truck

ラベルの例: 物体名

Supervised Learning

全ての画像にラベルが付与されている状況

損失を計算 $f(x_i)$

例) Cross Entropy Loss

(NLL Loss)

L1 Loss

MSE Loss

画像

画像表現

x_i
-11
2.90
-0.50
0.11
0.52
⋮

y_i : motorbike

$$\begin{pmatrix} 0.01 \\ 0.03 \\ 0.02 \\ 0.91 \\ 0.01 \\ 0.02 \end{pmatrix}$$

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

airplane
bus
car
motorbike
train
truck

(確率)

ラベルの例: 物体名

Cross Entropy Loss

入力: 画像 x_i , ラベル y_i ($i = 1, 2, \dots, N$)

損失の計算:

$$z_i = f(x_i)$$

損失の計算に
データが必要

$$L = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(z_i^T y_i)}{\sum_j \exp(z_{ij})}$$

softmax

損失の計算に
ラベルが必要

y_i : motorbike

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

airplane
bus
car
motorbike
train
truck

ラベルの例: 物体名

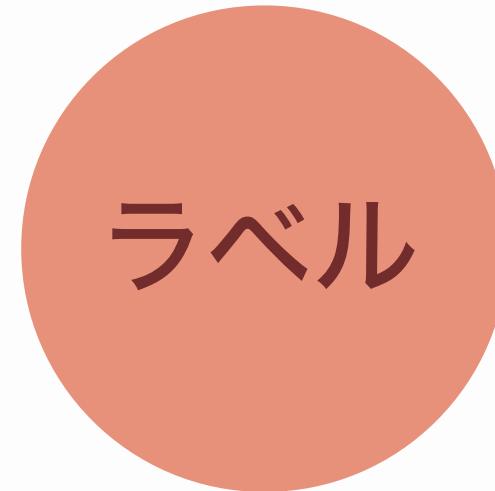
Supervised Learning



全ての画像にラベルが付与されている状況



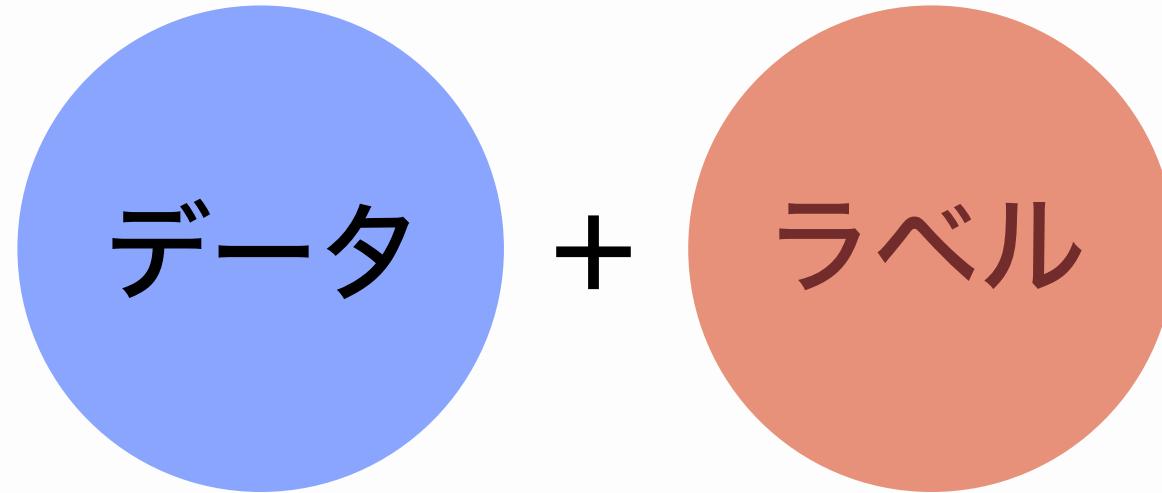
+



Supervised Learning

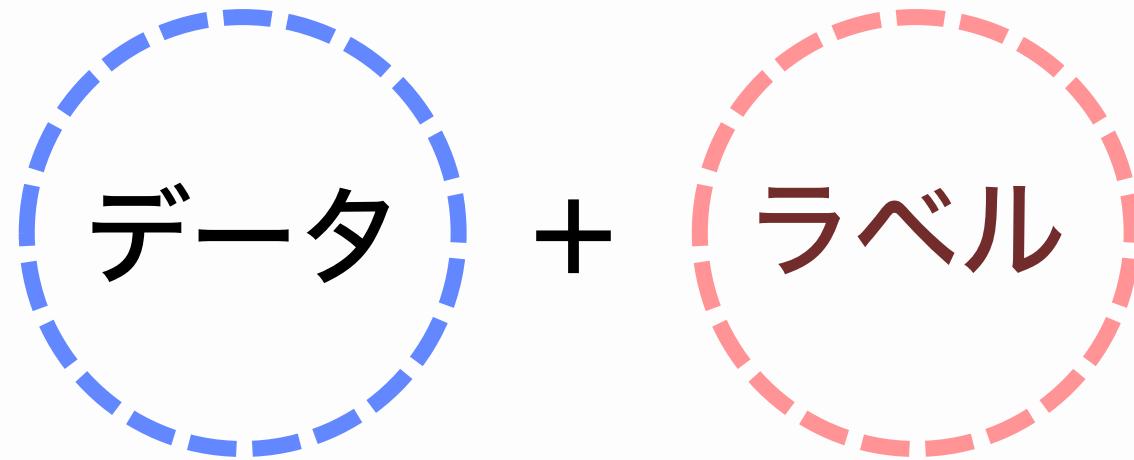


高い精度を達成するには大量のデータが必要となる



限られたデータからの学習

実際の多くの問題ではデータもラベルも限りがある



限られたデータからの学習

様々な条件下での「学習」が提案されている

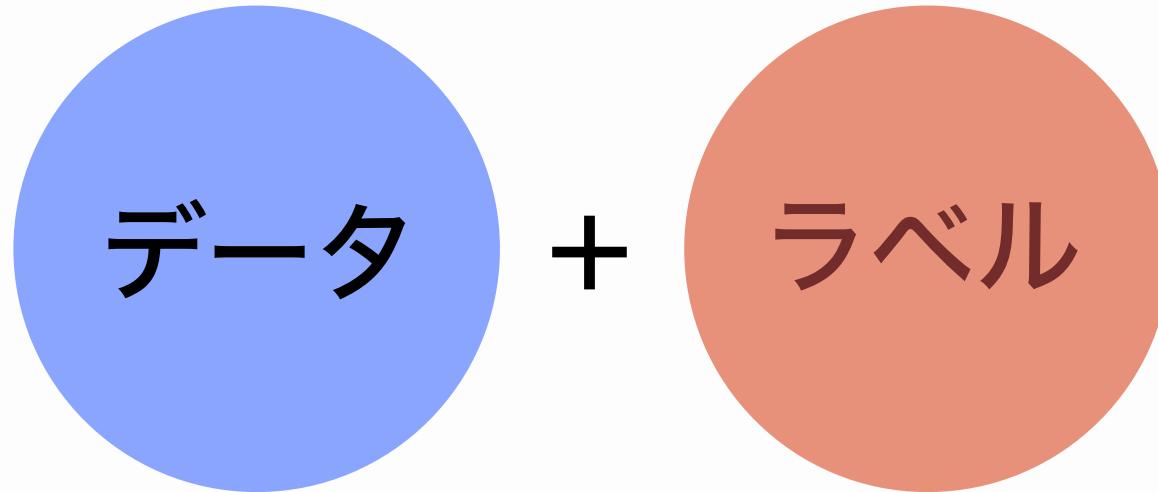
ラベルの量	Supervised	Semi-supervised	Unsupervised
データの量	(Many-Shot)	Few-Shot	Zero-Shot
ラベルの完全さ	Strongly-	Weakly-	Randomly-
ラベル付与方式	(Manually-)	Self-supervised	
実データの有無	Data-driven	Formula-driven?	

まずはそれぞれの問題設定を簡単に紹介します

Supervised Learning



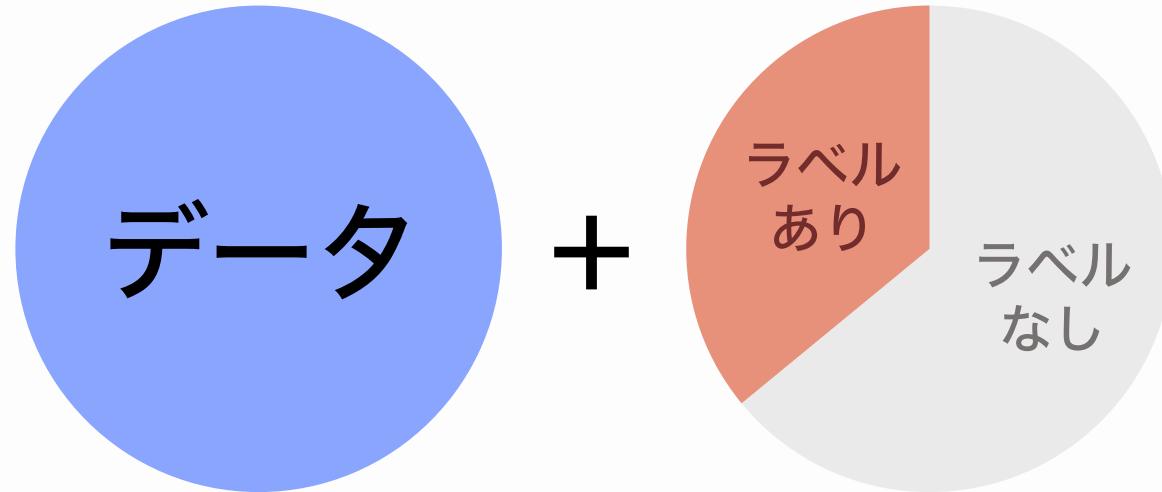
全てのデータにラベルが付与されている状態
ラベル付与率は100%



Semi-supervised Learning



一部のデータにのみラベルが付与されている状態
ラベル付与率は100%未満



Unsupervised Learning



データにラベルが付与されていない状態
ラベル付与率は0%



学習フレームワーク



様々な条件下での「学習」が提案されている

ラベルの量	Supervised	Semi-supervised	Unsupervised
データの量	(Many-Shot)	Few-Shot	Zero-Shot
ラベルの完全さ	Strongly-	Weakly-	Randomly-
ラベル付与方式	(Manually-)	Self-supervised	
実データの有無	Data-driven	Formula-driven?	

まずはそれぞれの問題設定を簡単に紹介します

学習フレームワーク



様々な条件下での「学習」が提案されている

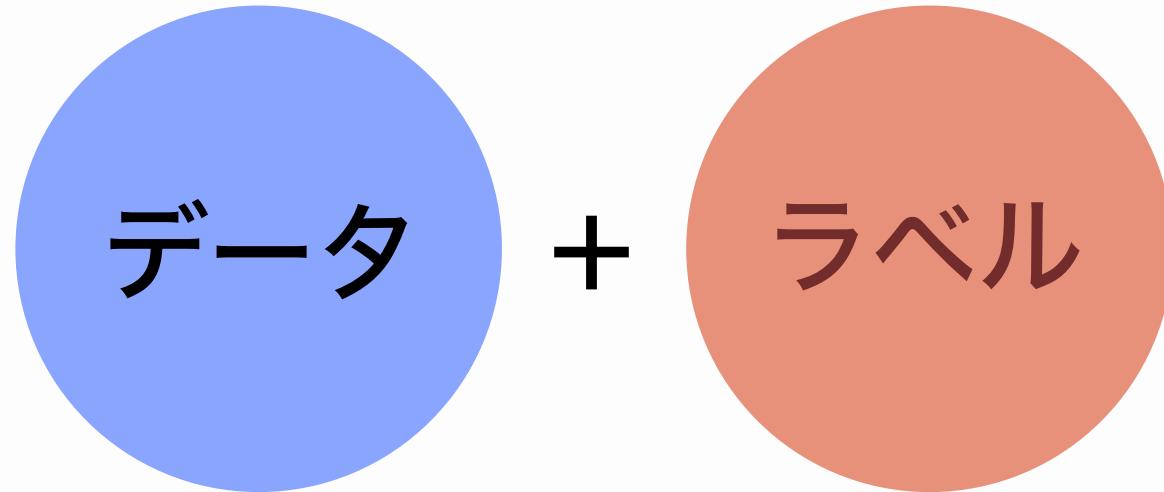
ラベルの量	Supervised	Semi-supervised	Unsupervised
データの量	(Many-Shot)	Few-Shot	Zero-Shot
ラベルの完全さ	Strongly-	Weakly-	Randomly-
ラベル付与方式	(Manually-)	Self-supervised	
実データの有無	Data-driven	Formula-driven?	

まずはそれぞれの問題設定を簡単に紹介します

(Many-Shot Learning)



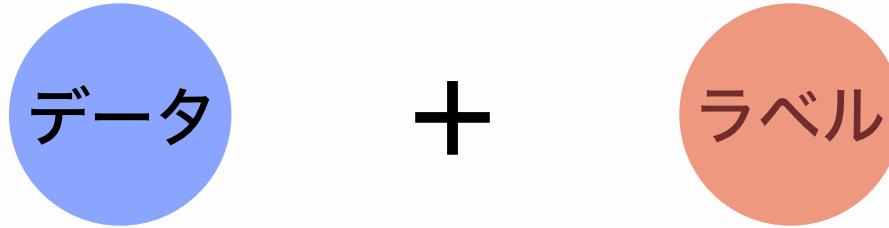
大量のデータがある状態, 通常はラベルもあると仮定



Few-Shot Learning

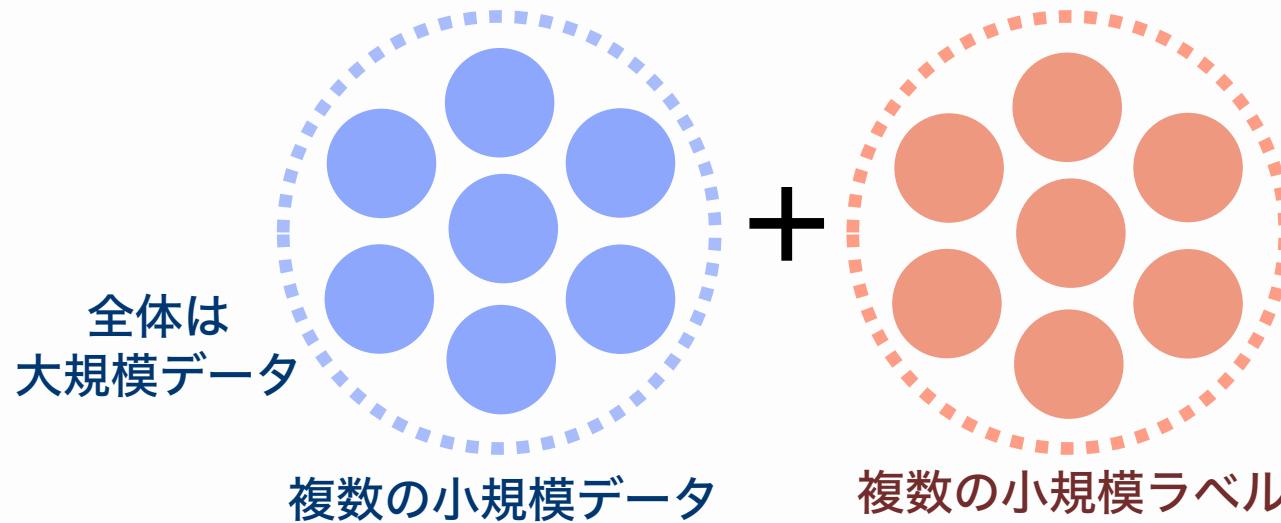


データ量が少ない状態
※小さい問題が沢山ある場合も含んでいる → メタ学習



Meta Learning

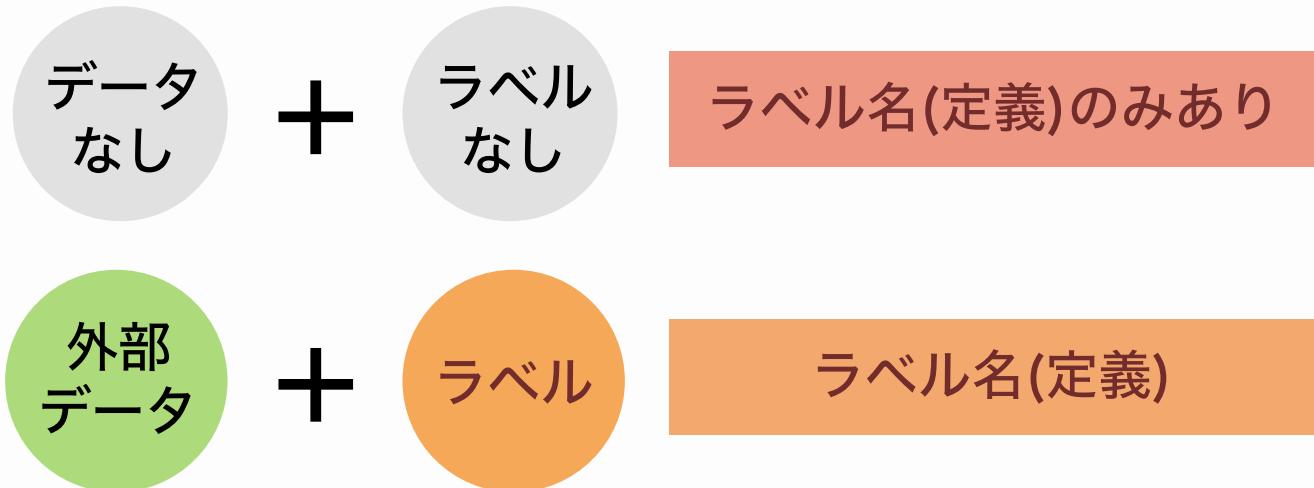
別の大規模データで小規模な学習-テストを繰り返すことで、
「どう学習すれば良いか」を学習する



Zero-Shot Learning



直接的な学習データがない状態
ただし、ラベル間の関係など外部データがある状態



学習フレームワーク



様々な条件下での「学習」が提案されている

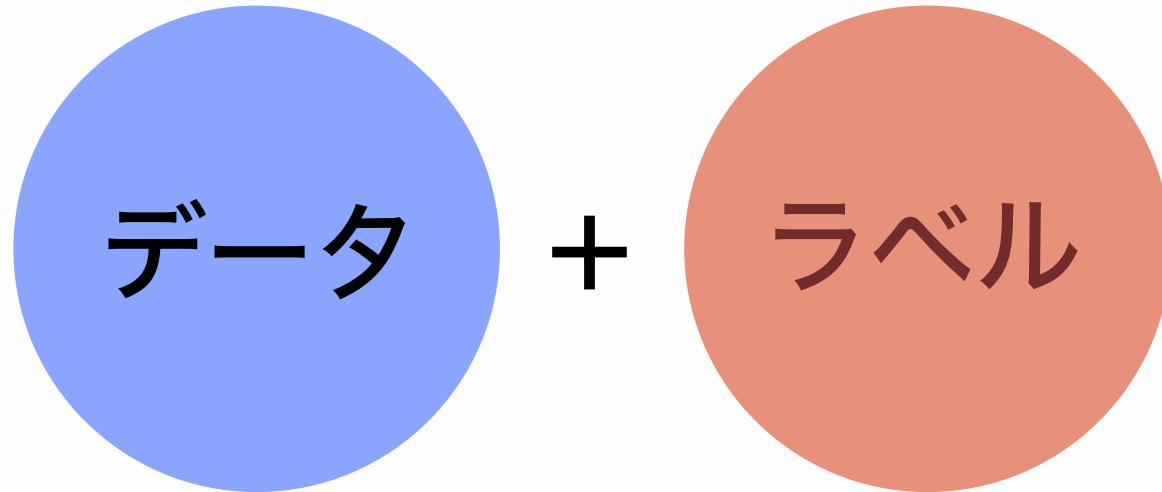
ラベルの量	Supervised	Semi-supervised	Unsupervised
データの量	(Many-Shot)	Few-Shot	Zero-Shot
ラベルの完全さ	Strongly-	Weakly-	Randomly-
ラベル付与方式	(Manually-)	Self-supervised	
実データの有無	Data-driven	Formula-driven?	

まずはそれぞれの問題設定を簡単に紹介します

Strongly-supervised Learning



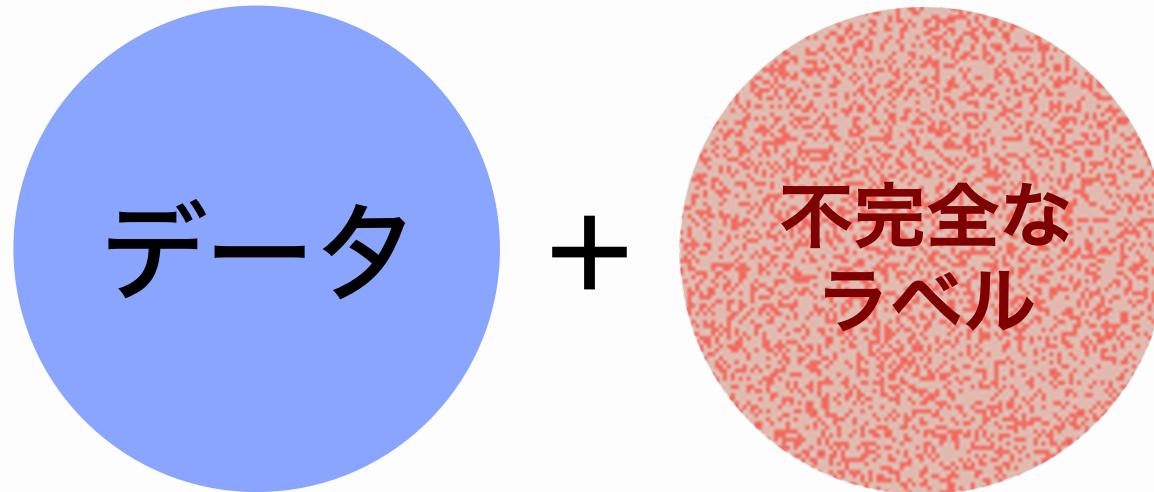
完全なラベルがある状態



Weakly-supervised Learning



不完全なラベルがある状態



Weakly-supervised Learning



不完全なラベルがある状態

[弱教師の例]

画像セグメンテーション

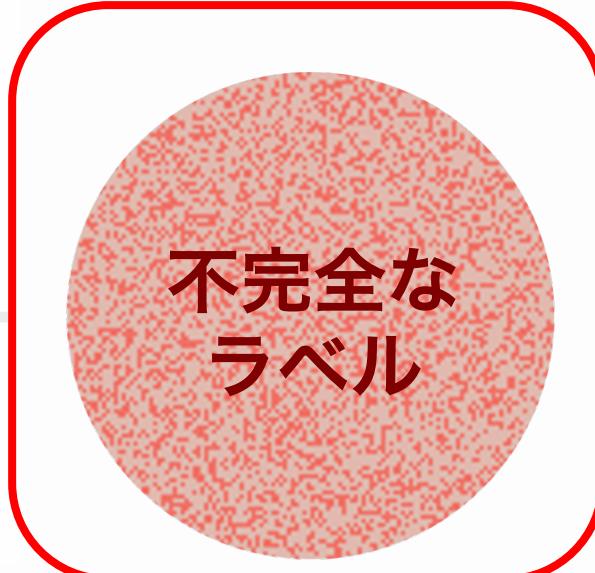
→ 物体ラベルがあるがマスクがない

画像分類

→ ユーザータグのみある

一般の問題

→ ラベルにノイズがある



不完全な
ラベル

学習フレームワーク



様々な条件下での「学習」が提案されている

ラベルの量	Supervised	Semi-supervised	Unsupervised
データの量	(Many-Shot)	Few-Shot	Zero-Shot
ラベルの完全さ	Strongly-	Weakly-	Randomly-
ラベル付与方式	(Manually-)	Self-supervised	
実データの有無	Data-driven	Formula-driven?	

まずはそれぞれの問題設定を簡単に紹介します

学習フレームワーク



様々な条件下での「学習」が提案されている

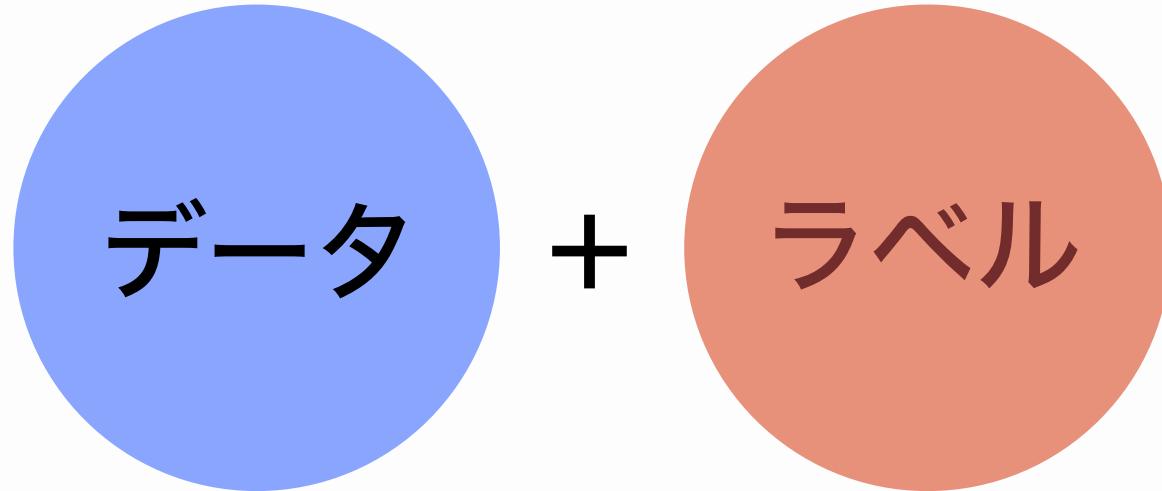
ラベルの量	Supervised	Semi-supervised	Unsupervised
データの量	(Many-Shot)	Few-Shot	Zero-Shot
ラベルの完全さ	Strongly-	Weakly-	Randomly-
ラベル付与方式	(Manually-)	Self-supervised	
実データの有無	Data-driven	Formula-driven?	

まずはそれぞれの問題設定を簡単に紹介します

Manual Supervision



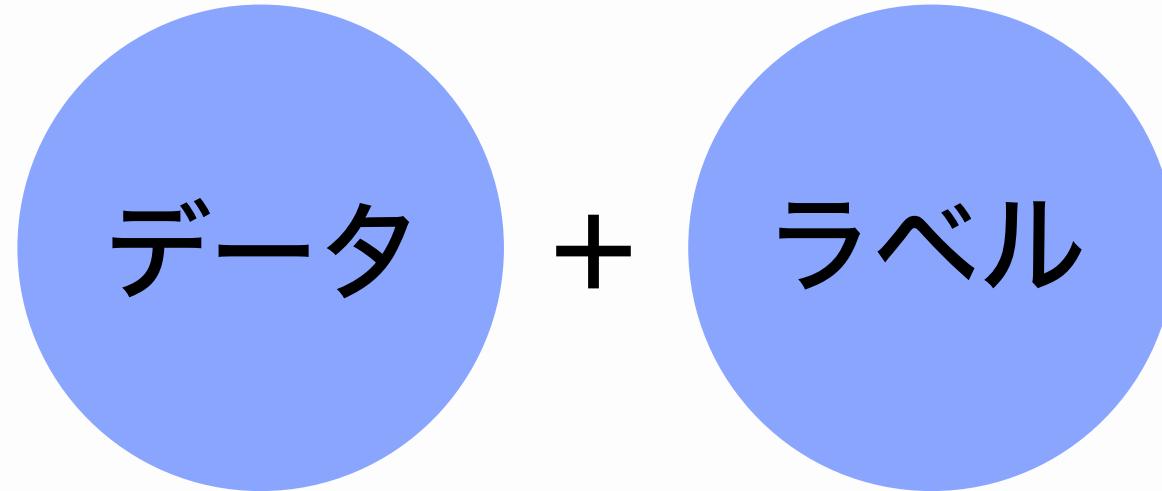
人手で付与されたラベル (を用いた学習)



Self-supervised Learning



自己的に付与されたラベルを用いた学習



Self-supervised Learning



自己的に付与されたラベルを用いた学習

「ラベル」はデータから生成できる？

- Pretext Taskを作つて解く

例) ジグソーパズル,
マスク部分の復元タスク

- 最近は対照学習が主流

例) SimCLR, MoCo

c.f., AutoEncoder

昔からデータを“ラベル”として使つていた

ラベル

学習フレームワーク



様々な条件下での「学習」が提案されている

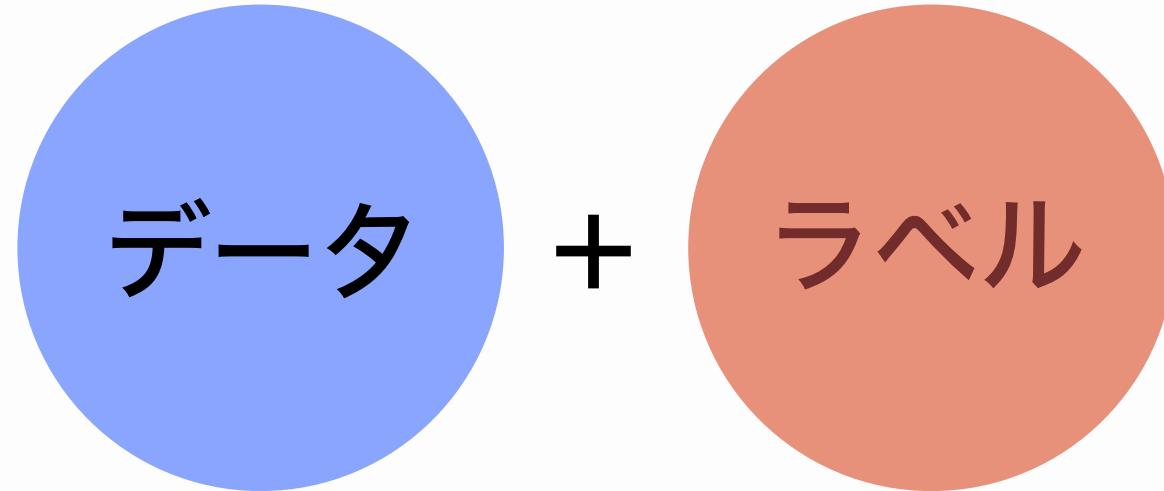
ラベルの量	Supervised	Semi-supervised	Unsupervised
データの量	(Many-Shot)	Few-Shot	Zero-Shot
ラベルの完全さ	Strongly-	Weakly-	Randomly-
ラベル付与方式	(Manually-)	Self-supervised	
実データの有無	Data-driven	Formula-driven?	

まずはそれぞれの問題設定を簡単に紹介します

Data-driven Approach



実データに基づいたアプローチ



Formula-driven Approach?

「データ」も「ラベル」も何らかの規則で生成できる？

生成規則(式)



データ

+

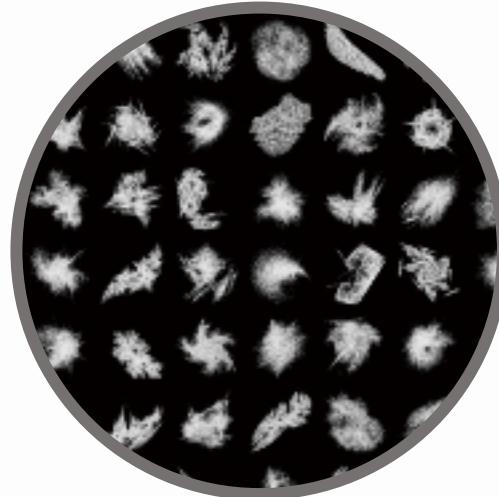
ラベル

Pre-training without Natural Images



自然画像を用いない事前学習

生成規則(式)



+

フラクタル
ラベル

Pre-training without Natural Images



Best Paper
Honorable Mention

Pre-training without Natural Images
Hirokatsu Kataoka, Kazushige Okayasu, Asato Matsumoto,
Eisuke Yamagata, Ryosuke Yamada, Nakamasa Inoue,
Akio Nakamura, Yutaka Satoh

産総研
TDU 東京電機大学

筑波大学
University of Tsukuba

東京工業大学
Tokyo Institute of Technology

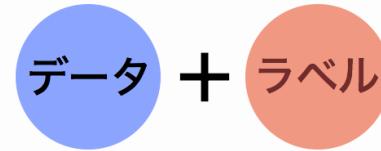
産総研 片岡さん, 佐藤先生, 電機大 中村先生らとの共同研究
ACCV20では賞を頂いたもので, プロジェクトは進行中です!

高い精度を求めるなら？

事前学習後、少量のデータでFine-tuningをするのが良い
最近はラベル無しでも事前学習が可能



大量データ(自己教師)



少量データ

→ これで大規模なラベル付けが回避できる

学習フレームワーク



様々な条件下での「学習」が提案されている

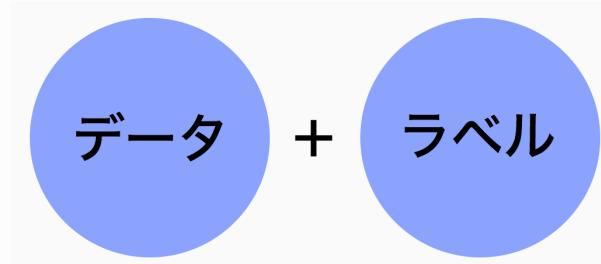
ラベルの量	Supervised	Semi-supervised	Unsupervised
データの量	(Many-Shot)	Few-Shot	Zero-Shot
ラベルの完全さ	Strongly-	Weakly-	Randomly-
ラベル付与方式	(Manually-)	Self-supervised	
実データの有無	Data-driven	Formula-driven?	

今日の本題

ここからの2つのトピック

第2部 Self-Supervised Learning

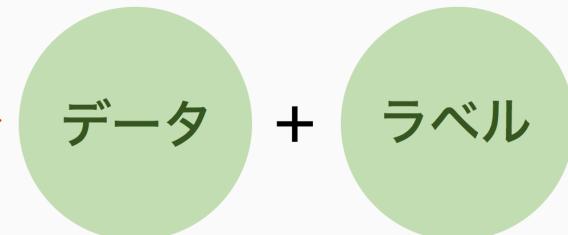
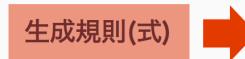
- ラベルはデータから生成できるか？
- Pretext Tasks, 対照学習



第3部 Formula-driven Approach

- データは数式から生成できるか？
- パーリンノイズ, フラクタル

生成規則(式)

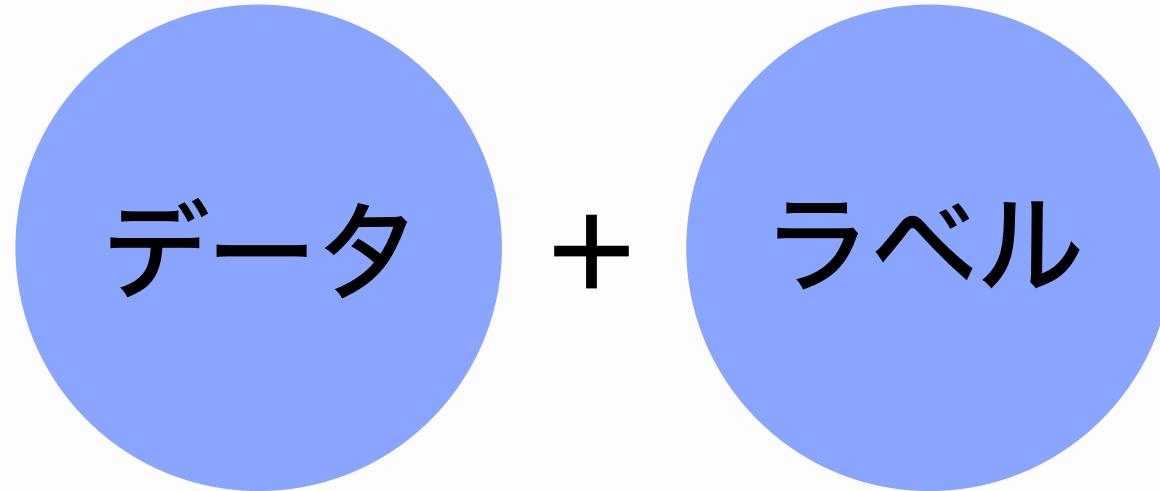


第2部 Self-Supervised Learning

Self-supervised Learning

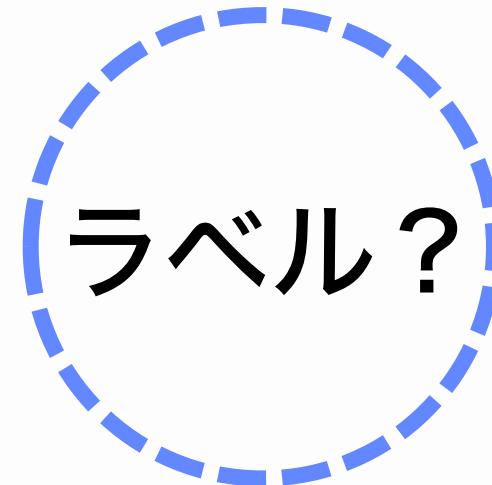


自己的に付与されたラベルを用いた学習



Pretext Tasks

アイデア：画像のみが与えられた状態で疑似的な問題を作る



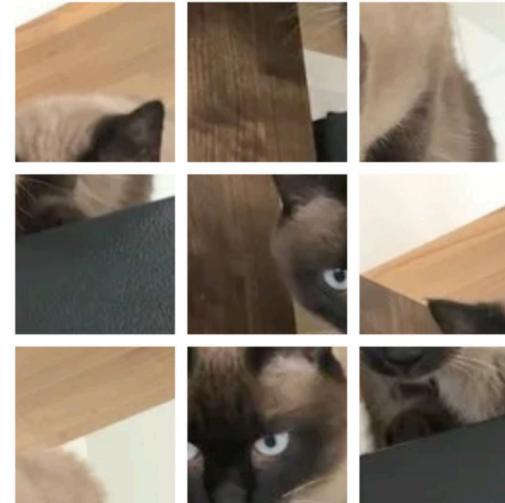
Pretext Tasks

アイデア：画像のみが与えられた状態で疑似的な問題を作る



単純なジグソーパズル

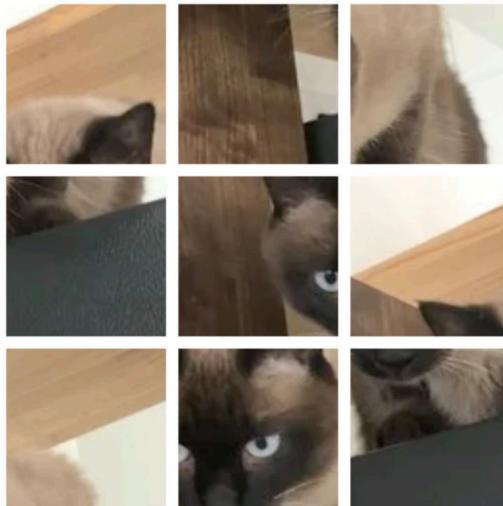
アイデア：画像のみが与えられた状態で疑似的な問題を作る



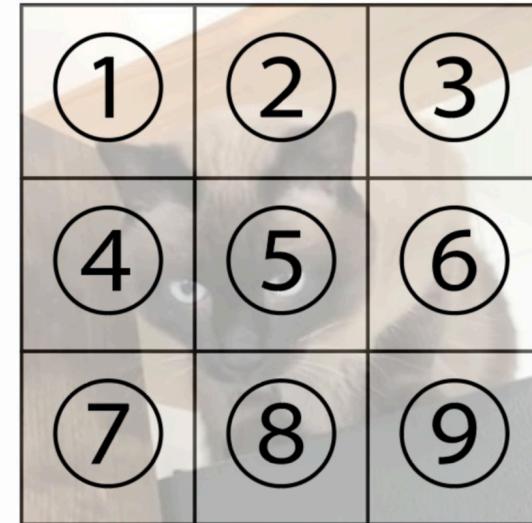
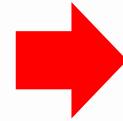
ジグソーパズル

単純なジグソーパズル

何を“当てれば”元の画像に戻せるか？



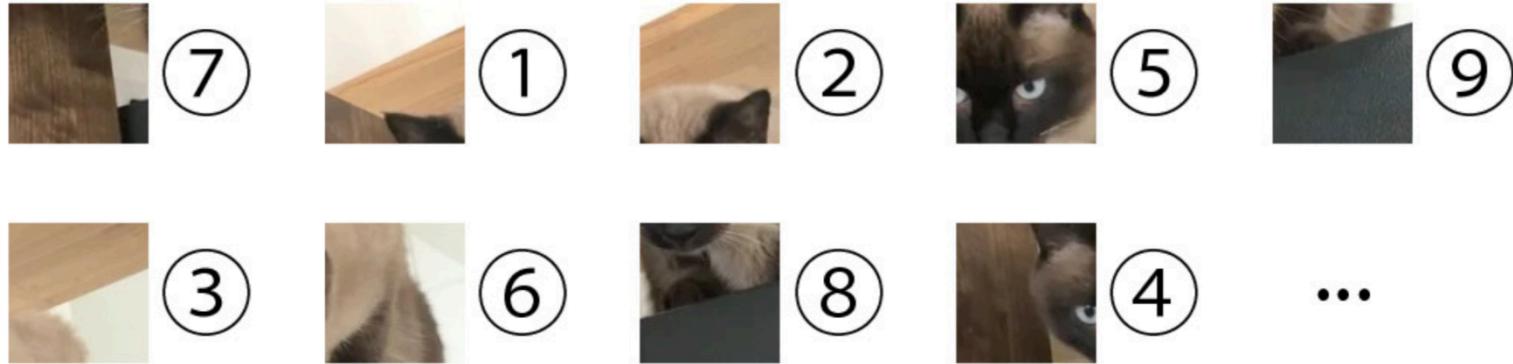
ジグソーパズル



元の位置の番号

単純なジグソーパズル

元の位置の番号を当てる9クラス分類問題



※実際にはもっと複雑なLossを設計します

単純なジグソーパズル

元の画像の位置を教師信号として使う



画像

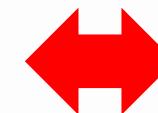
$f(x_i)$



$$\begin{pmatrix} -1.61 \\ 2.90 \\ -0.50 \\ 0.11 \\ 0.52 \\ \vdots \end{pmatrix} \rightarrow \begin{pmatrix} 0.05 \\ 0.09 \\ 0.02 \\ 0.13 \\ 0.61 \\ \vdots \end{pmatrix}$$

画像表現

$y_i : ⑤$



$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ \vdots \end{pmatrix} \begin{matrix} \text{位置①} \\ \text{位置②} \\ \text{位置③} \\ \text{位置④} \\ \text{位置⑤} \\ \vdots \end{matrix}$$

(確率)

ラベル: 元の位置

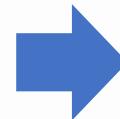
Supervised Learning

全ての画像にラベルが付与されている状況と同じ



画像

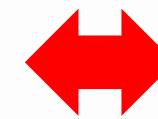
$$f(x_i)$$



$$\begin{pmatrix} -1.61 \\ 2.90 \\ -0.50 \\ 0.11 \\ 0.52 \\ \vdots \end{pmatrix} \rightarrow \begin{pmatrix} 0.01 \\ 0.03 \\ 0.02 \\ 0.91 \\ 0.01 \\ 0.02 \end{pmatrix}$$

画像表現

y_i : motorbike



$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

airplane
bus
car
motorbike
train
truck

(確率)

ラベルの例: 物体名

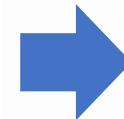
単純なジグソーパズル

元の画像の位置を教師信号として使う



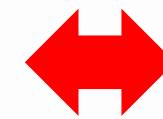
画像

$f(x_i)$



$$\begin{pmatrix} -1.61 \\ 2.90 \\ -0.50 \\ 0.11 \\ 0.52 \\ \vdots \end{pmatrix} \rightarrow \begin{pmatrix} 0.05 \\ 0.09 \\ 0.02 \\ 0.13 \\ 0.61 \\ \vdots \end{pmatrix}$$

画像表現



$y_i : ⑤$

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ \vdots \end{pmatrix} \leftarrow \begin{array}{l} \text{位置①} \\ \text{位置②} \\ \text{位置③} \\ \text{位置④} \\ \text{位置⑤} \\ \vdots \end{array}$$

(確率)

ラベル: 元の位置

Self-Supervised Learning



なぜ「自己教師あり」と呼ばれているのか？

教師なし学習の
問題設定



画像

$f(x_i)$



$$\begin{pmatrix} -1.61 \\ 2.90 \\ -0.50 \\ 0.11 \\ 0.52 \\ \vdots \end{pmatrix}$$

画像表現

(確率)

$$\begin{pmatrix} 0.05 \\ 0.09 \\ 0.02 \\ 0.13 \\ 0.61 \\ \vdots \end{pmatrix}$$

$y_i : ⑤$

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ \vdots \end{pmatrix}$$

ラベル: 元の位置

教師あり学習の
損失が使える

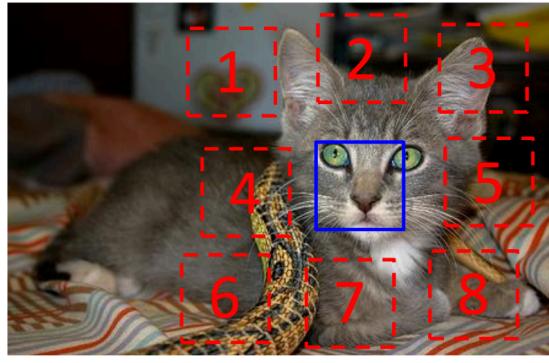
2017年頃までは教師なし表現
学習という呼び方が多かった

単純なクラスタリングとは異なり、
「教師信号」を利用している

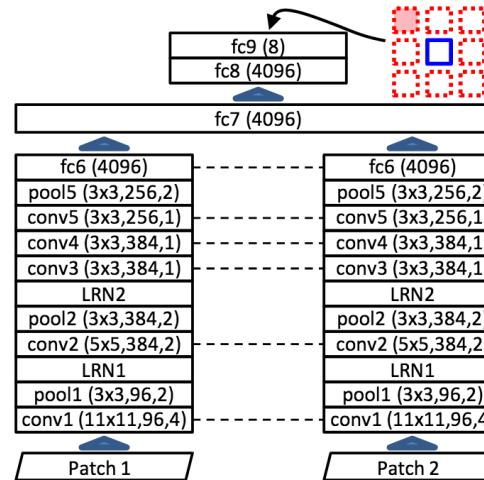
- 自動的に生成された教師信号を使った教師なし学習？ → わかりやすい名前が定着
- 最近はContrastive Learningにより「教師なし」という言い方に戻ることが多い

Context Prediction

画像の相対位置の番号を当てる分類問題, 性能は低い



$$X = (\text{[cat eye]}, \text{[cat ear]}); Y = 3$$

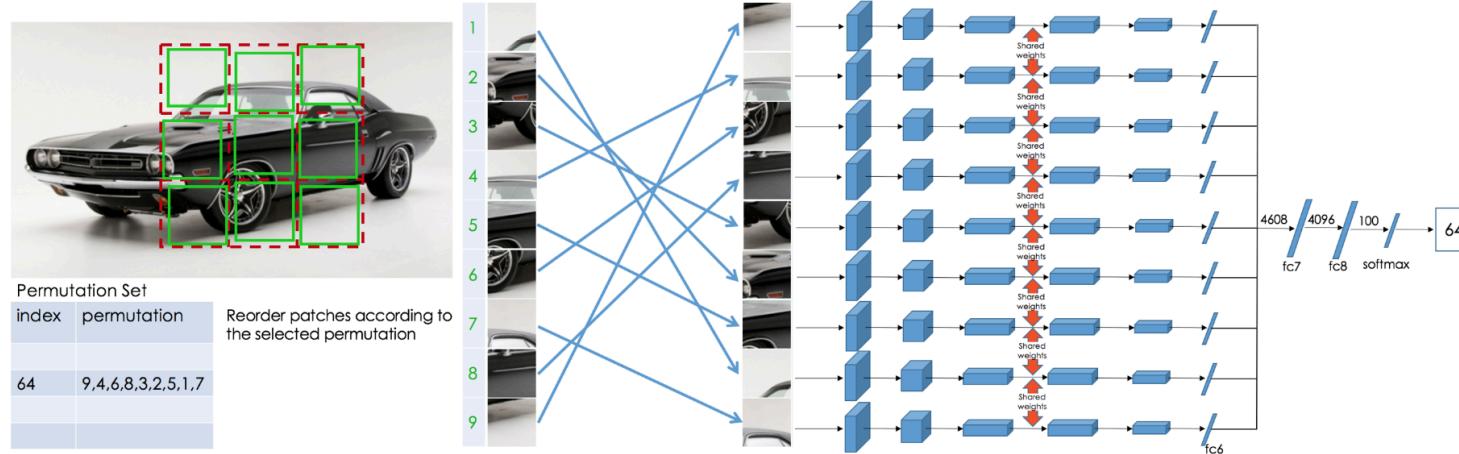


図は論文より引用

C. Doersch, et al., "Unsupervised visual representation learning by context prediction", Proc. ICCV 2015.

Jigsaw Puzzles

画像の相対位置の番号を当てる分類問題, 性能は中程度



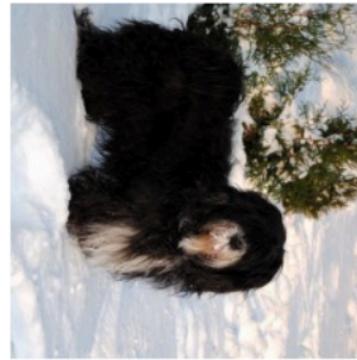
M. Noroozi and P. Favaro, Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles, ECCV, 2016. 図は論文より引用

Rotation

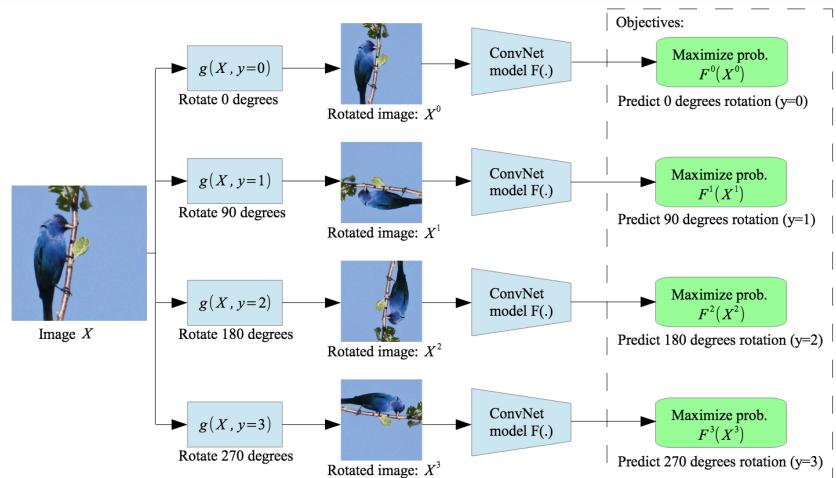
画像の回転角度を当てる、性能は中程度



90° rotation

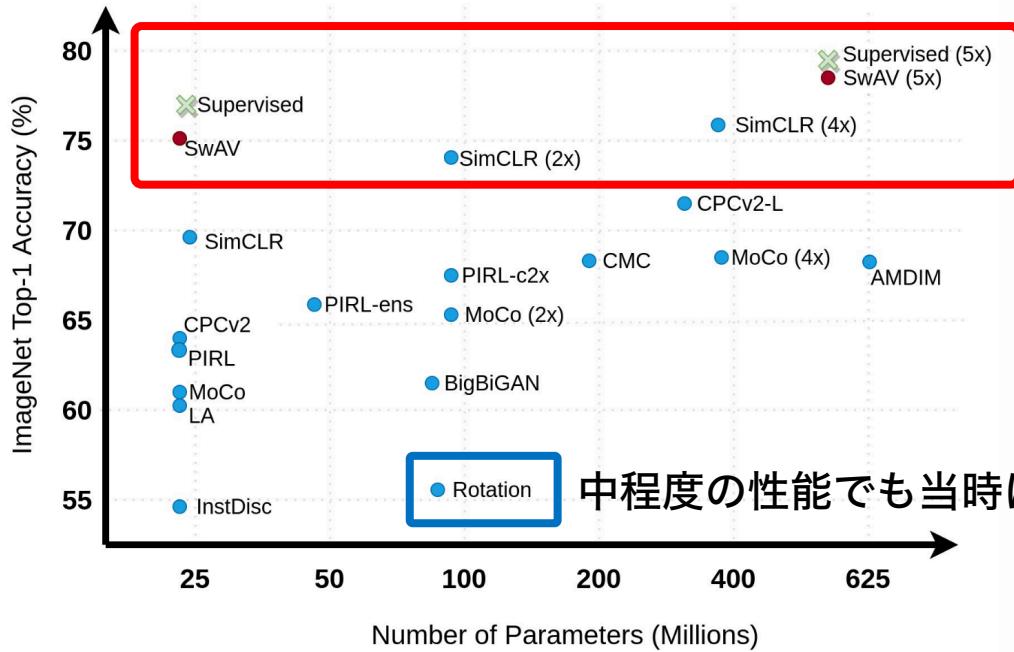


270° rotation



S. Gidaris, et al., UNSUPERVISED REPRESENTATION LEARNING BY PREDICTING IMAGE ROTATIONS, ICLR 2018. 図は論文より引用

ImageNet Top-1 Accuracy



ドメイン一致の
教師あり学習に迫る

Contrastive Learning
(対照学習)の発展

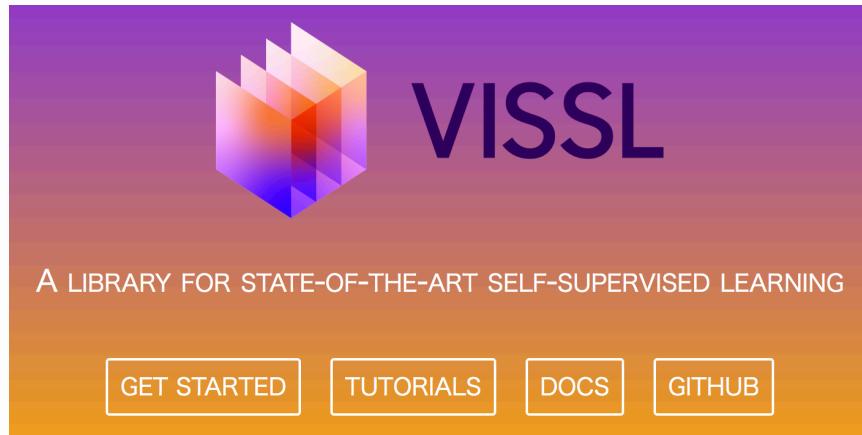
中程度の性能でも当時はすごかった

A. Jaiswal, et al, "A SURVEY ON CONTRASTIVE SELF-SUPERVISED LEARNING," arXiv2011.00362. より改変

Contrastive Learning



SimCLRとMoCoが有名なモデル, 最近はVISSLで簡単



VISSL: <https://vissl.ai/>

- SwAV, SimCLR, MoCo(v2), PIRL, NPID, NPID++, DeepClusterV2, ClusterFit, RotNet, Jigsawなどの実装がまとまっている
- Vision Transformerもある

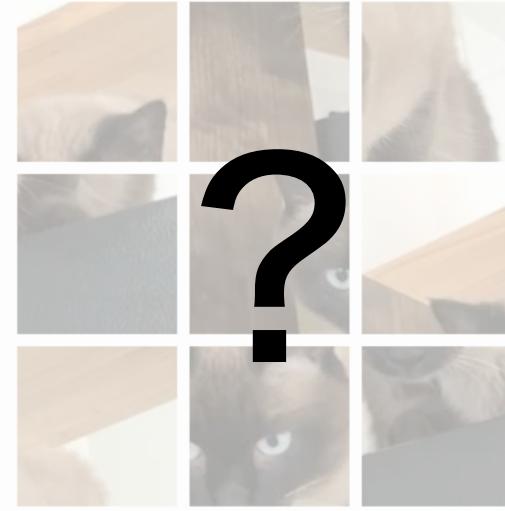
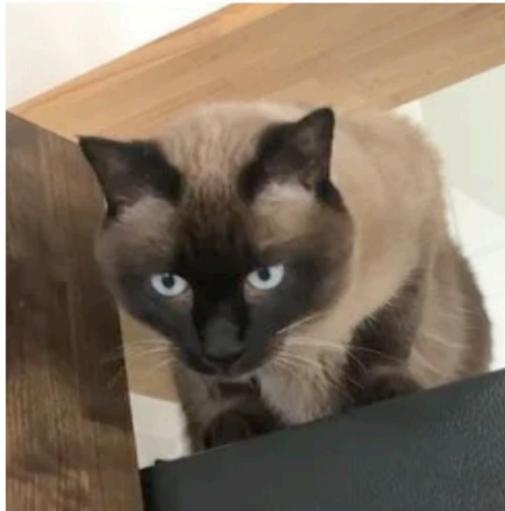
T. Chen, et al., A Simple Framework for Contrastive Learning of Visual Representations, ICML 2020.
K. He, et al., Momentum Contrast for Unsupervised Visual Representation Learning, CVPR 2020.

Contrastive Learning



今日はどういうパズルを解いてるか？を考える（初学者向け）

※Noise Contrastive Estimator (NCE) loss の話をするのが正統派



Contrastive Learning



2つの画像を照らし合わせました。元画像は一緒？

問題 1



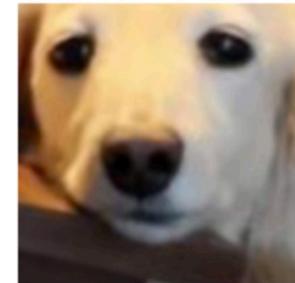
?
=



問題 2



?
=



Contrastive Learning



2つの画像を照らし合わせました。元画像は一緒？

問題 1



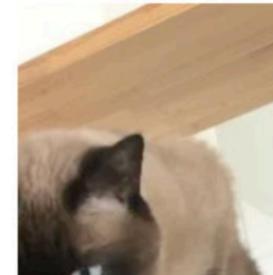
?

=



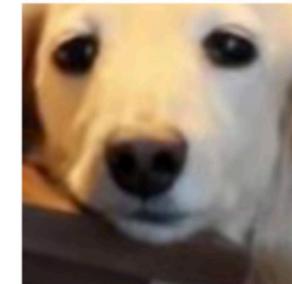
同じ画像から切り取られた
ものなので一緒

問題 2



?

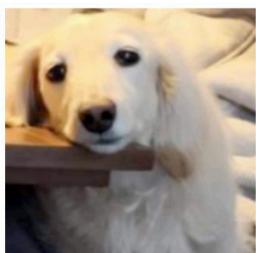
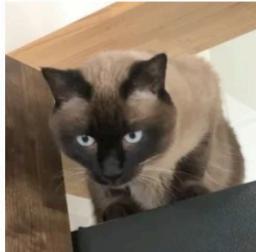
=



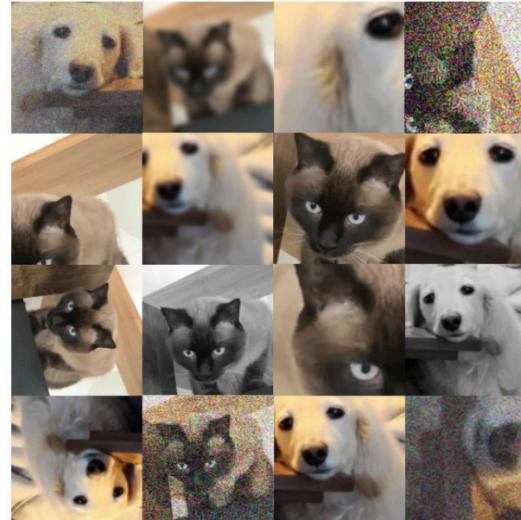
異なる画像から切り取られた
ものなので違う

Contrastive Learning

「照らし合わせ」のパズルを作る



元画像

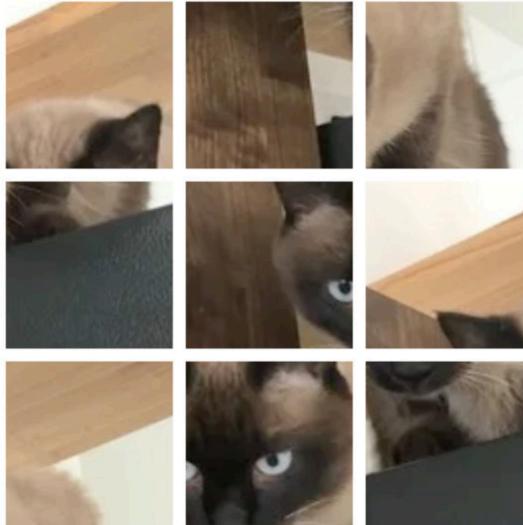


データ拡張

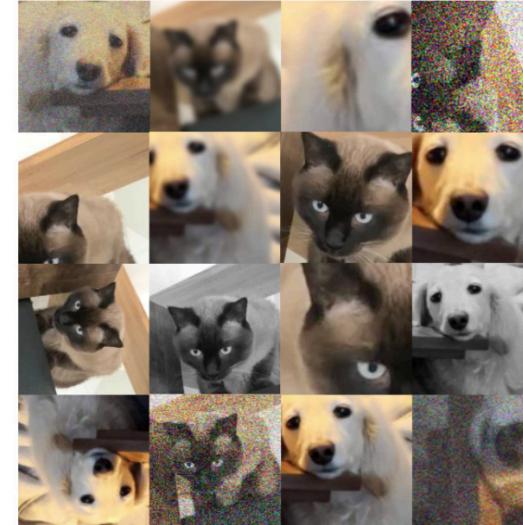


ランダムに2枚を
照らし合わせる

“パズル”的違い



ジグソーパズル



「照らし合わせ」のパズル
→ 対照学習ではこれを解く

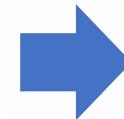
単純なジグソーパズル

元の画像の位置を教師信号として使う

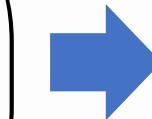


画像

$f(x_i)$



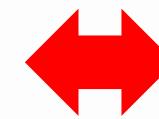
$$\begin{pmatrix} -1.61 \\ 2.90 \\ -0.50 \\ 0.11 \\ 0.52 \\ \vdots \end{pmatrix}$$



$$\begin{pmatrix} 0.05 \\ 0.09 \\ 0.02 \\ 0.13 \\ 0.61 \\ \vdots \end{pmatrix}$$

(確率)

$y_i : ⑤$



$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ \vdots \end{pmatrix}$$

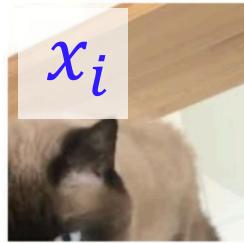
位置①
位置②
位置③
位置④
位置⑤
⋮

ラベル: 元の位置

Loss for Contrastive Learning



コサイン類似度を計算する



x_i



$f(x_i)$

$$\begin{pmatrix} -1.61 \\ 2.90 \\ -0.50 \\ 0.11 \\ 0.52 \\ \vdots \end{pmatrix}$$



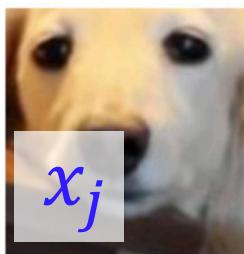
z_i

$$\begin{pmatrix} 1.41 \\ 2.90 \\ -1.97 \\ 0.35 \\ 0.65 \\ \vdots \end{pmatrix}$$



Similarity

$$s_{ij} = \frac{z_i^T z_j}{|z_i| |z_j|}$$



x_j



$f(x_j)$

$$\begin{pmatrix} -1.20 \\ -1.56 \\ -0.75 \\ 0.49 \\ 0.11 \\ \vdots \end{pmatrix}$$



z_j

$$\begin{pmatrix} 1.71 \\ 2.46 \\ -0.53 \\ -0.73 \\ 0.36 \\ \vdots \end{pmatrix}$$

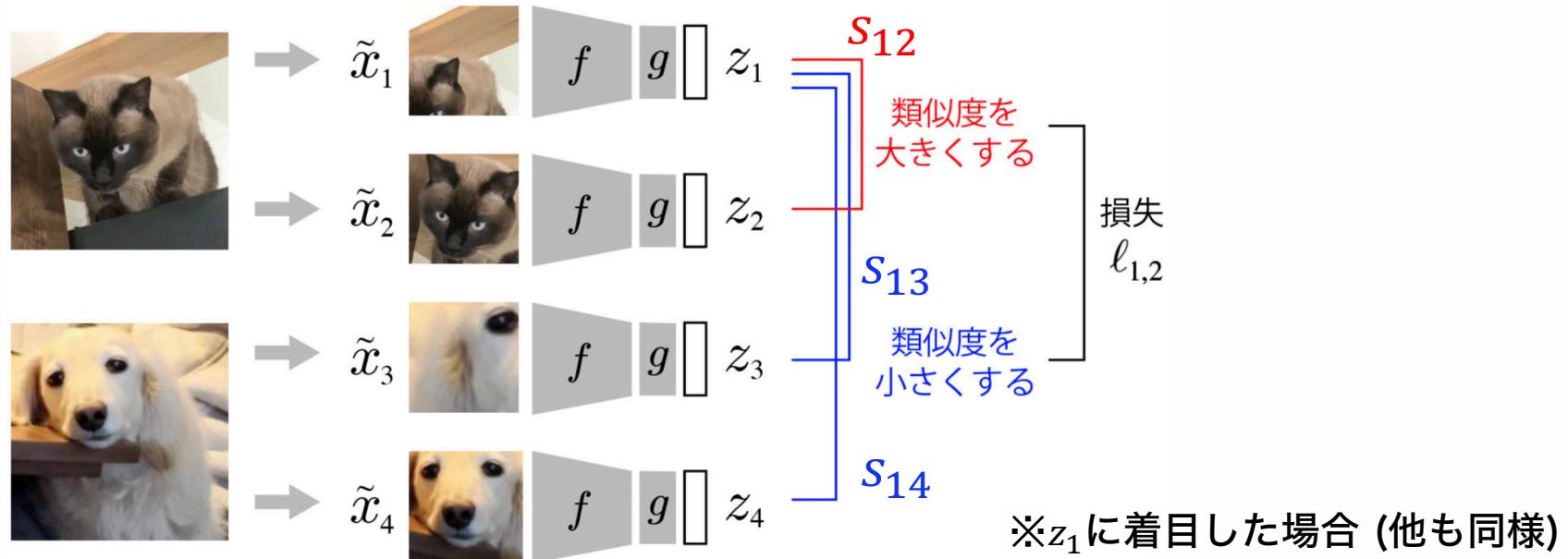


この値を小さくすべきか、
大きくすべきか？

画像表現

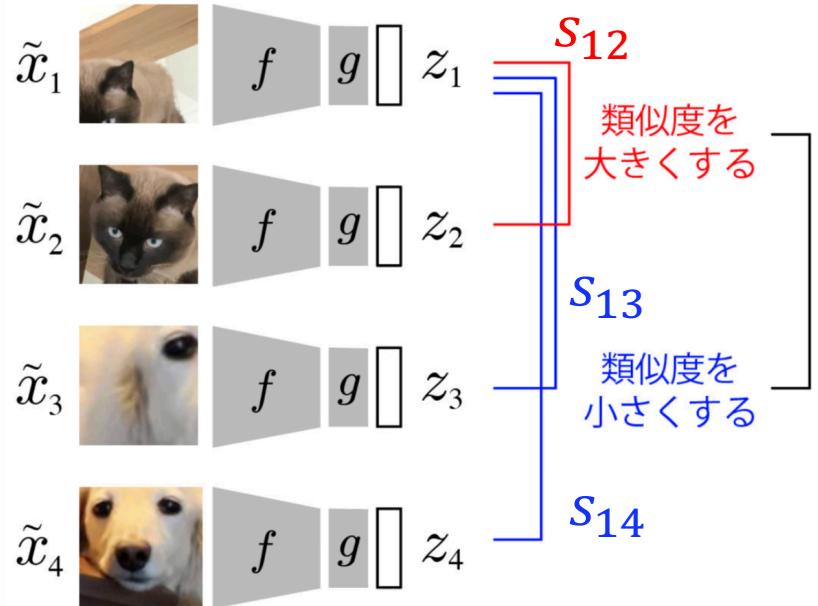
Loss for Contrastive Learning

2枚の画像の場合



NT-Xent Loss

2枚の画像の場合

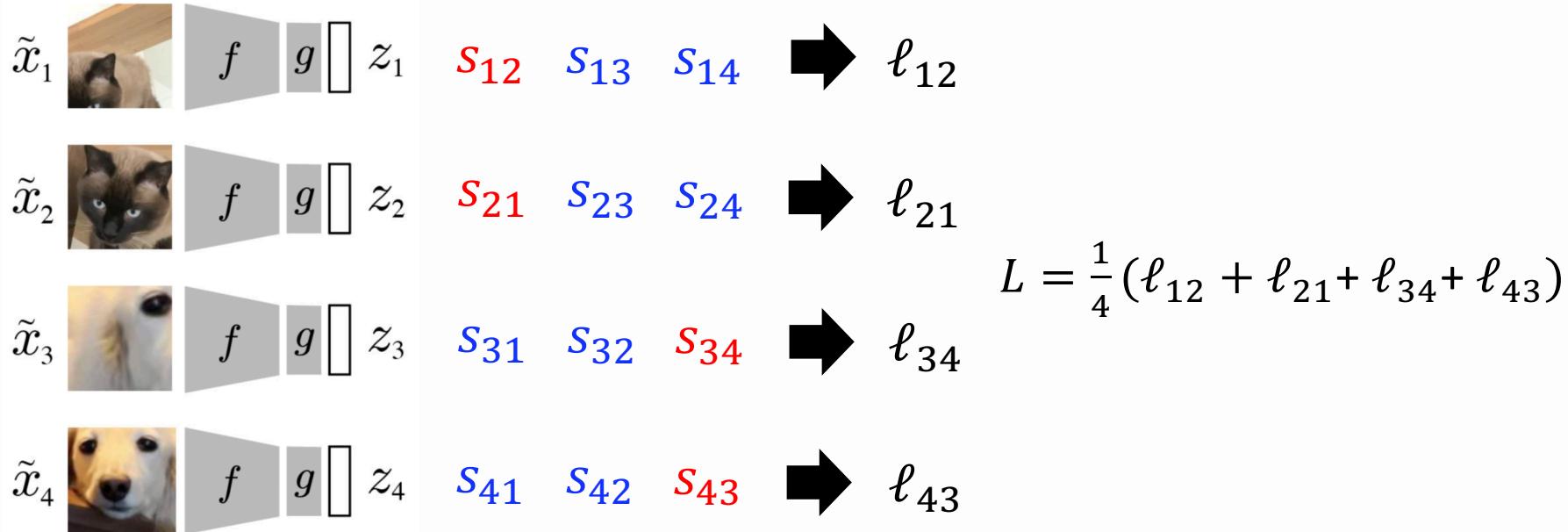


$$\ell_{1,2} = -\frac{\exp(s_{12}/\tau)}{\exp(s_{12}/\tau) + \exp(s_{13}/\tau) + \exp(s_{14}/\tau)}$$

s_{12} を s_{13}, s_{14} よりも 相対的に 大きくする
 τ は温度パラメータ

NT-Xent Loss

2枚の画像の場合の損失



NT-Xent Loss



ミニバッチに画像が n 枚ある場合の損失

$$\tilde{x}_1 \quad \begin{array}{c} \text{Siamese Cat Image} \\ f \quad g \end{array} \quad z_1 \quad s_{12} \quad s_{13} \quad s_{14} \quad \dots \quad s_{1,2n} \quad \rightarrow \quad \ell_{12}$$

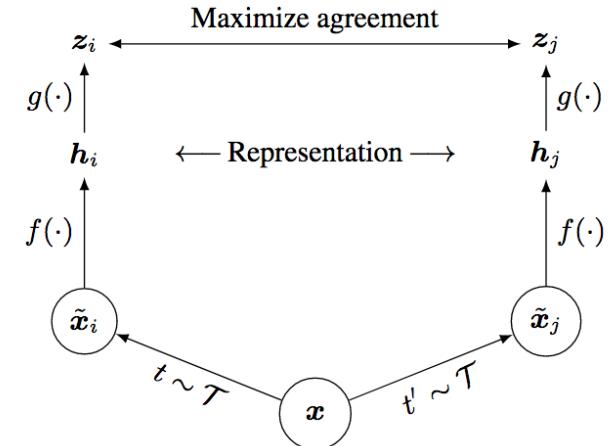
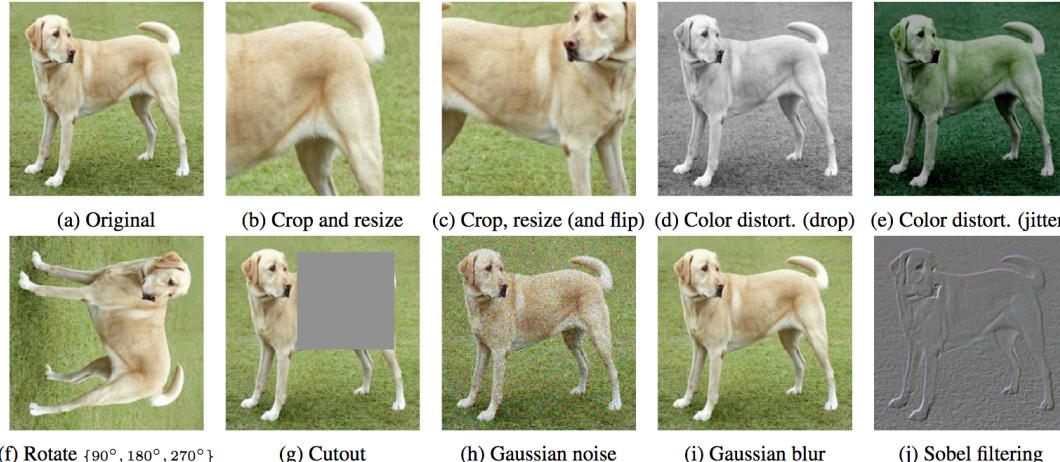
$$\tilde{x}_2 \quad \begin{array}{c} \text{Siamese Cat Image} \\ f \quad g \end{array} \quad z_2 \quad s_{21} \quad s_{23} \quad s_{24} \quad \dots \quad s_{1,2n} \quad \rightarrow \quad \ell_{21}$$

:

:

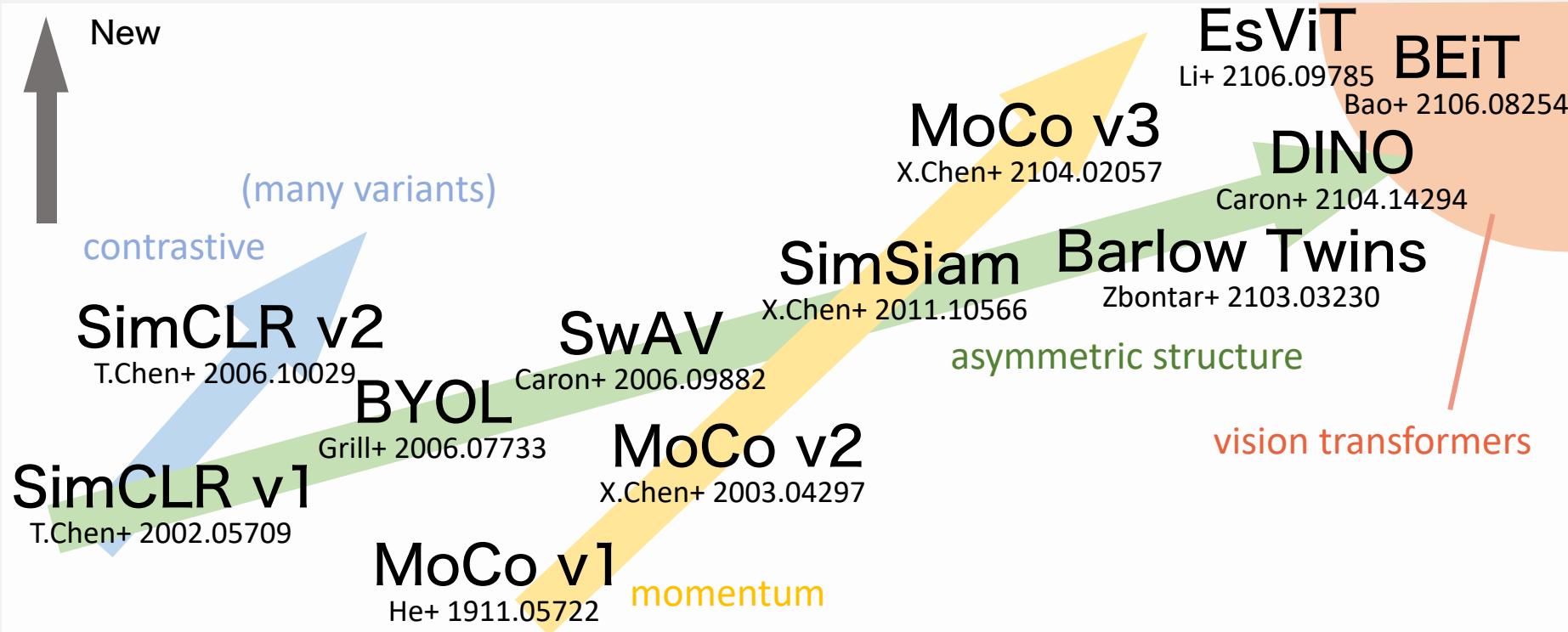
$$L = \frac{1}{2n} \sum_{k=1}^n (\ell_{2k-1,2k} + \ell_{2k,2k-1})$$

Simple Framework for Contrastive Learning

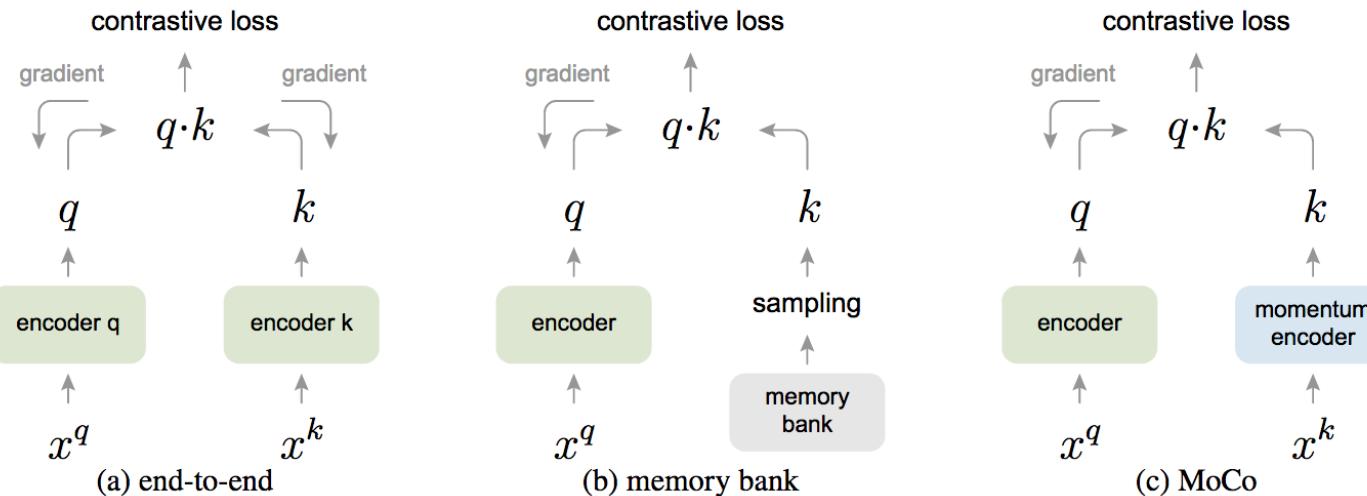


T. Chen, et al., A Simple Framework for Contrastive Learning of Visual Representations, ICML 2020.

最近の代表的な手法

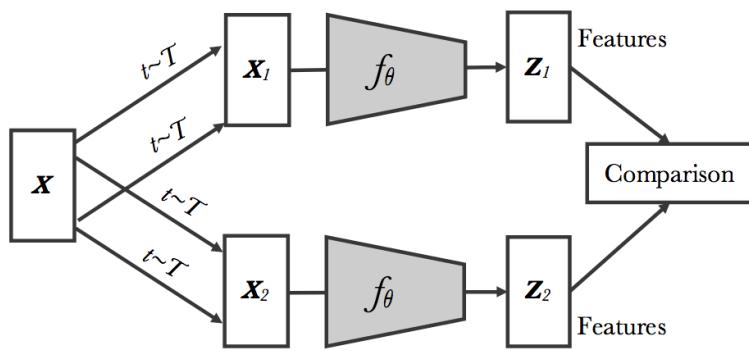


Momentum Contrast

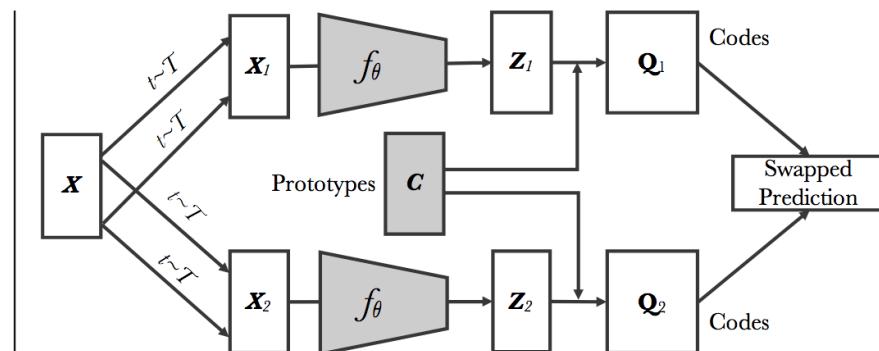


K. He, et al., Momentum Contrast for Unsupervised Visual Representation Learning, CVPR 2020.

Swapping Assignments



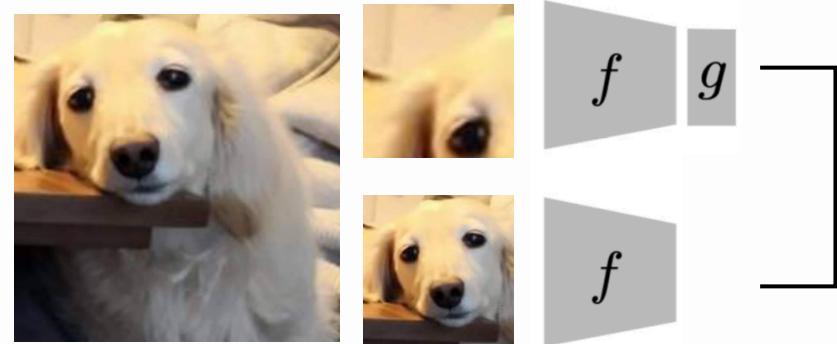
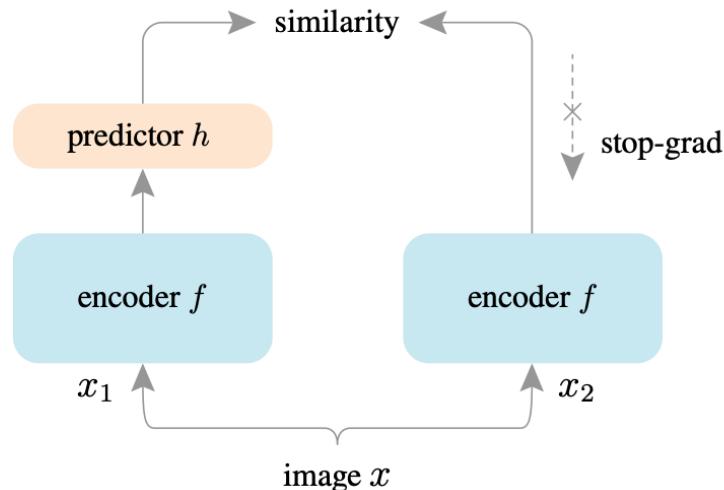
Contrastive instance learning



Swapping Assignments between Views (Ours)

M Caron, et al., Unsupervised Learning of Visual Features by Contrasting Cluster Assignments, NeurIPS 2020.

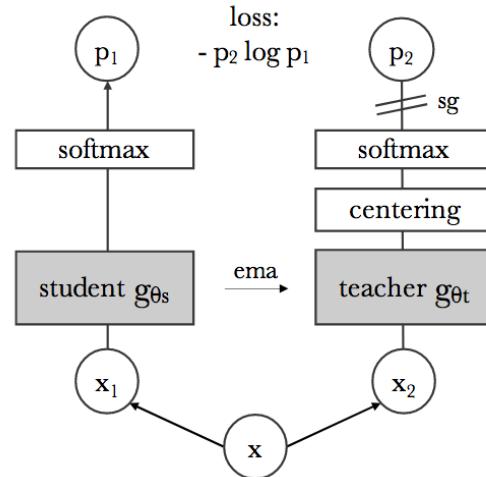
片側Stop-gradと非対称性が本質的



正例ペアのみで損失が計算できる

X. Chen and K. He, Exploring Simple Siamese Representation Learning, CVPR 2021.

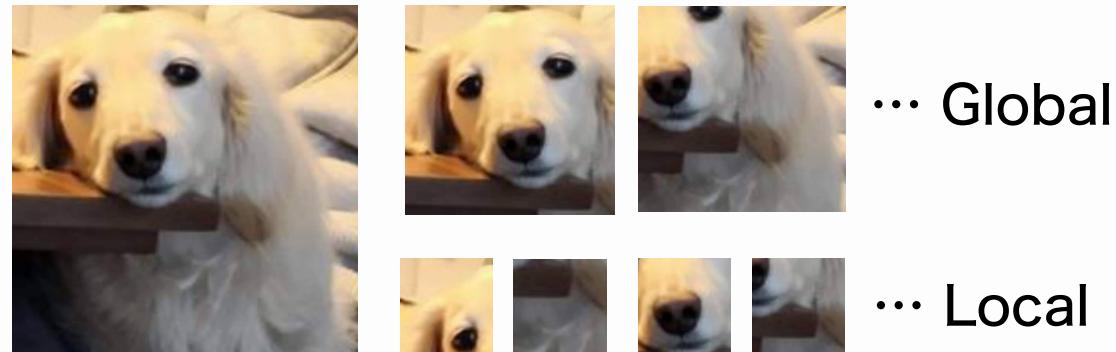
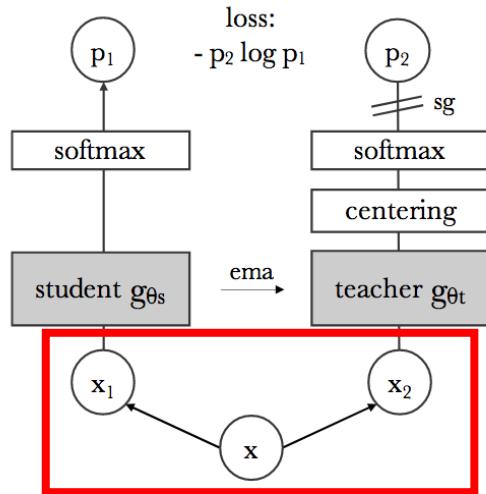
Self-distillation with no label (DINO)



ImageNet Top 1 Acc.					
Method	Arch.	Param.	im/s	Linear	k -NN
SCLR [12]	RN50w4	375	117	76.8	69.3
SwAV [10]	RN50w2	93	384	77.3	67.3
BYOL [30]	RN50w2	93	384	77.4	–
DINO	ViT-B/16	85	312	78.2	76.1
SwAV [10]	RN50w5	586	76	78.5	67.1
BYOL [30]	RN50w4	375	117	78.6	–
BYOL [30]	RN200w2	250	123	79.6	73.9
DINO	ViT-S/8	21	180	79.7	78.3
SCLRv2 [13]	RN152w3+SK	794	46	79.8	73.1
DINO	ViT-B/8	85	63	80.1	77.4

M Caron, et al., Emerging Properties in Self-Supervised Vision Transformers, arXiv2104.14294.

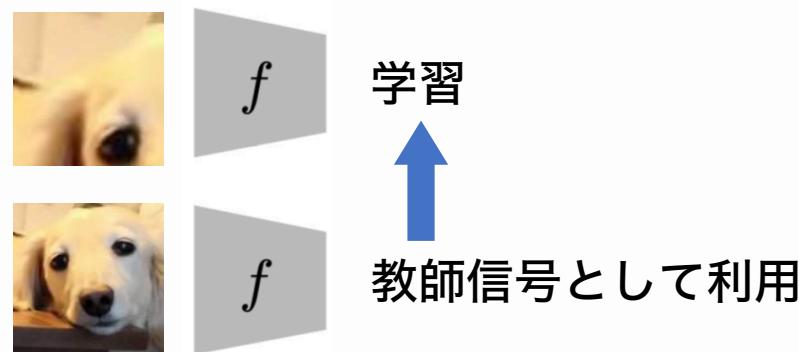
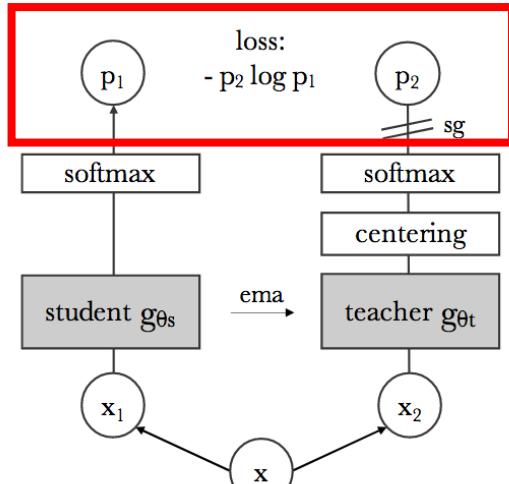
Global and local views



画像1枚からクロップ, globalとlocalを分ける

M Caron, et al., Emerging Properties in Self-Supervised Vision Transformers, arXiv2104.14294.

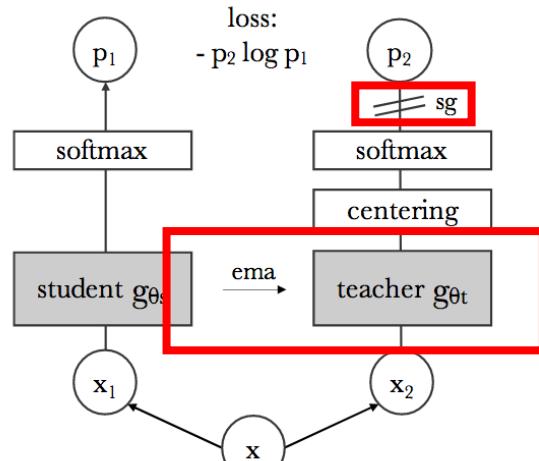
Student and teacher, local to global



基本的にはglobalの画像が教師信号となる

M Caron, et al., Emerging Properties in Self-Supervised Vision Transformers, arXiv2104.14294.

Self-distillation with a momentum encoder



Momentum encoderを作つておき、
Teacher Networkとして利用する

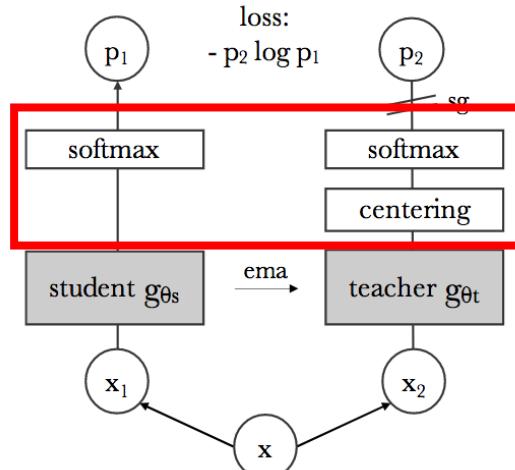
$$\theta_t \leftarrow \lambda \theta_t + (1 - \lambda) \theta_s$$

(exponential moving average)

Lossからは勾配を計算しない

M Caron, et al., Emerging Properties in Self-Supervised Vision Transformers, arXiv2104.14294.

Softmaxの計算は温度付き, 教師側はCenteringも導入



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

temperature parameter

$$g_t(x) \leftarrow g_t(x) + c$$

bias

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

M Caron, et al., Emerging Properties in Self-Supervised Vision Transformers, arXiv2104.14294.

ネットワーク構造はXCiTなどを利用すると良い

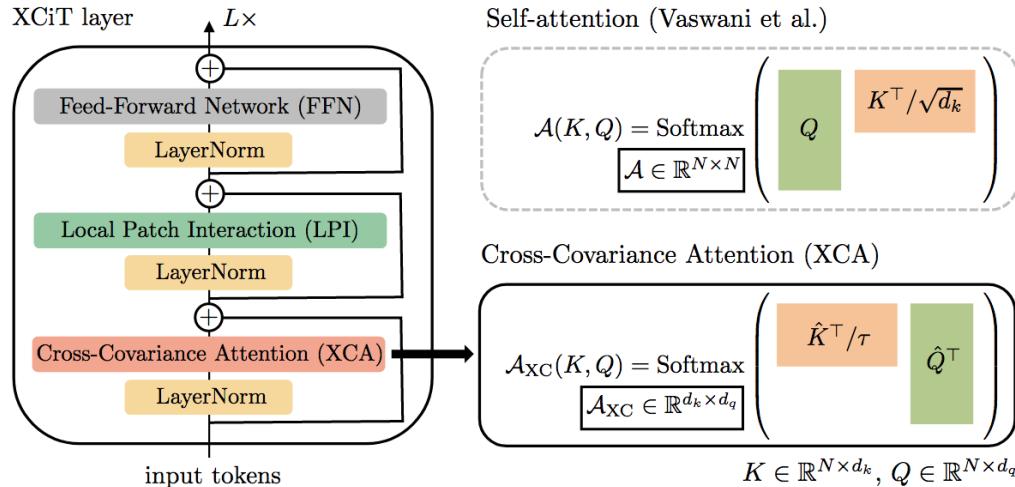
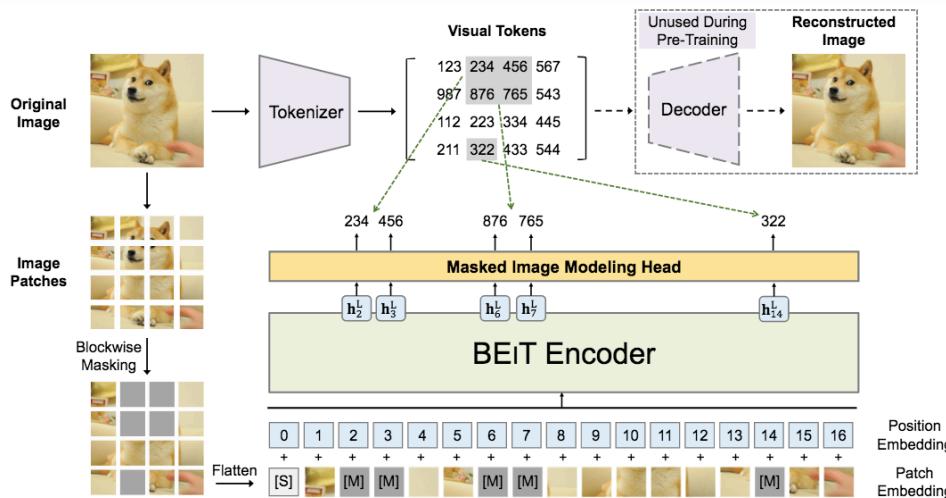


Table 3: **Self-supervised learning.** Top-1 acc. on ImageNet 1k. We report with a crop-ratio 0.875 for consistency with DINO. For the last row it is set to 1.0 (improves from 80.7% to 80.9%). All models are trained for 300 epochs.

SSL Method	Model	#params	FLOPs	Linear	<i>k</i> -NN
MoBY [76]	Swin-T [44]	29M	4.5B	75.0	-
DINO [12]	ResNet-50 [28]	23M	4.1B	74.5	65.6
DINO [12]	ViT-S/16 [22]	22M	4.6B	76.1	72.8
DINO [12]	ViT-S/8 [22]	22M	22.4B	79.2	77.2
DINO [12]	XCiT-S12/16	26M	4.9B	77.8	76.0
DINO [12]	XCiT-S12/8	26M	18.9B	79.2	77.1
DINO [12]	ViT-B/16 [22]	87M	17.5B	78.2	76.1
DINO [12]	ViT-B/8 [22]	87M	78.2B	80.1	77.4
DINO [12]	XCiT-M24/16	84M	16.2B	78.8	76.4
DINO [12]	XCiT-M24/8	84M	64.0B	80.3	77.9
DINO [12]	XCiT-M24/8/384	84M	188.0B	80.9	-

A. El-Nouby, et al., XGiT: Cross-Covariance Image Transformers, arXiv2106.09681.

BERTを利用すると良い？→この辺りから未解決の問題

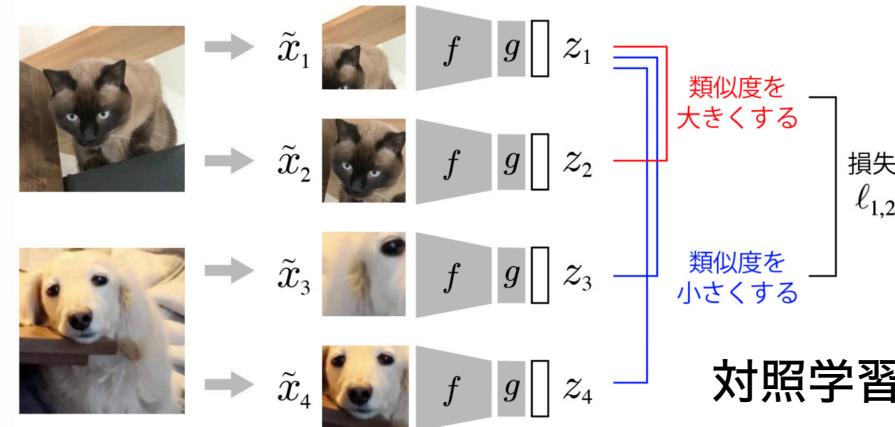
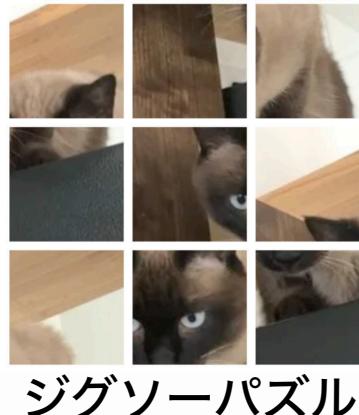


Models	CIFAR-100	ImageNet
<i>Training from scratch (i.e., random initialization)</i>		
ViT ₃₈₄ (Dosovitskiy et al., 2020)	48.5*	77.9
DeiT (Touvron et al., 2020)	n/a	81.8
<i>Supervised Pre-Training on ImageNet-1K (using labeled data)</i>		
ViT ₃₈₄ (Dosovitskiy et al., 2020)	87.1	77.9
DeiT (Touvron et al., 2020)	90.8	81.8
<i>Self-Supervised Pre-Training on ImageNet-1K (without labeled data)</i>		
iGPT-1.36B [†] (Chen et al., 2020a)	n/a	66.5
ViT ₃₈₄ -JFT300M [†] (Dosovitskiy et al., 2020)	n/a	79.9
DINO (Caron et al., 2021)	91.7	82.8
MoCo v3 (Chen et al., 2021)	87.1	n/a
BEiT (ours)	90.1	83.2
<i>Self-Supervised Pre-Training, and Intermediate Fine-Tuning on ImageNet-1K</i>		
BEiT (ours)	91.8	83.2

H. Bao, et al., BEiT: BERT Pre-Training of Image Transformers, arXiv2106.08254 .

第2部 まとめ

- Self-Supervised Learning
 - アイデア：画像のみが与えられた状態で疑似的な問題を作る
 - 対照学習：データ拡張を利用して画像表現を学習

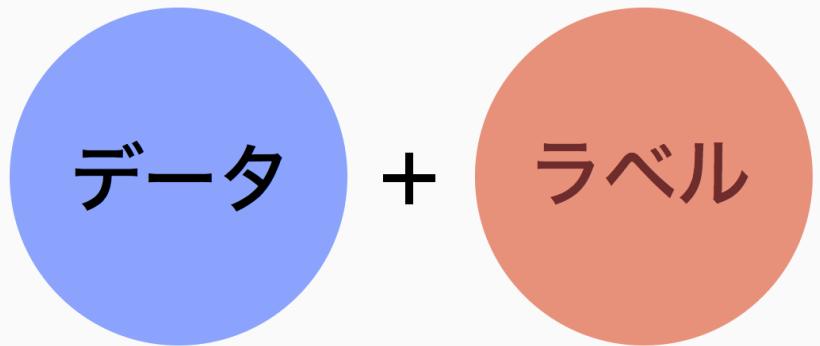


第3部 Formula-driven Approach

大規模事前学習のはじまり



Supervised Learning

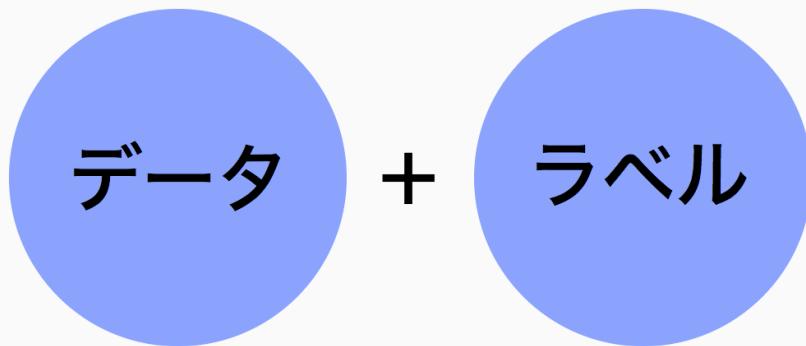


画像に関する大規模サービスを
運営している企業が
画像検索サービスで有利となった

大規模なデータにラベル付け
↑
Data-driven Approachの成功

大規模事前学習の今

Self-Supervised Learning



大規模なデータのみあれば良い

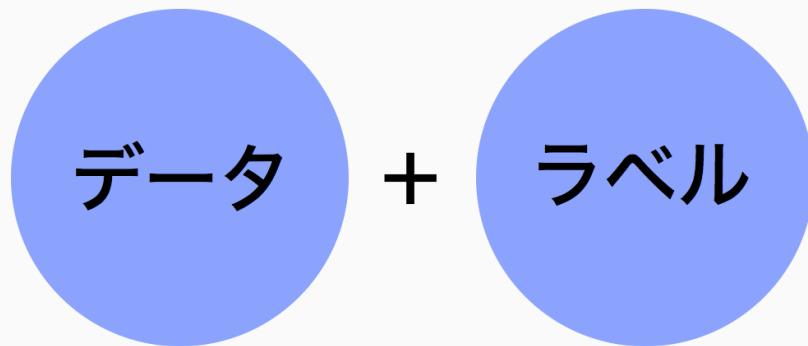


Data-driven Approachの終着点？

画像に関する大規模サービスを
運営している企業が
あらゆるタスクで有利となりつつある？
(新しいタスクのfine-tuningに強いかもしれない)

大規模事前学習のこれから？

Self-Supervised Learning

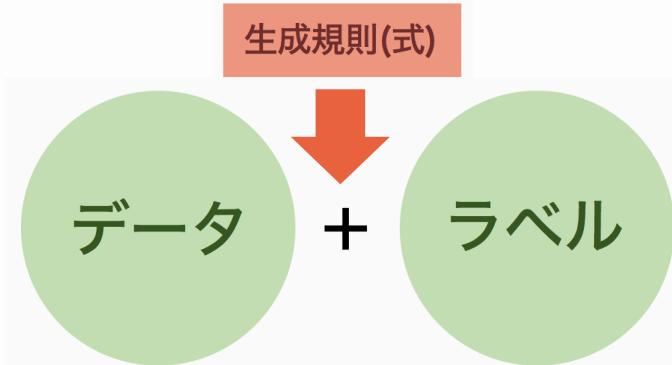


大規模なデータのみあれば良い



Data-driven Approachの終着点？ Formula-driven Approachの目標

新しい事前学習方式？

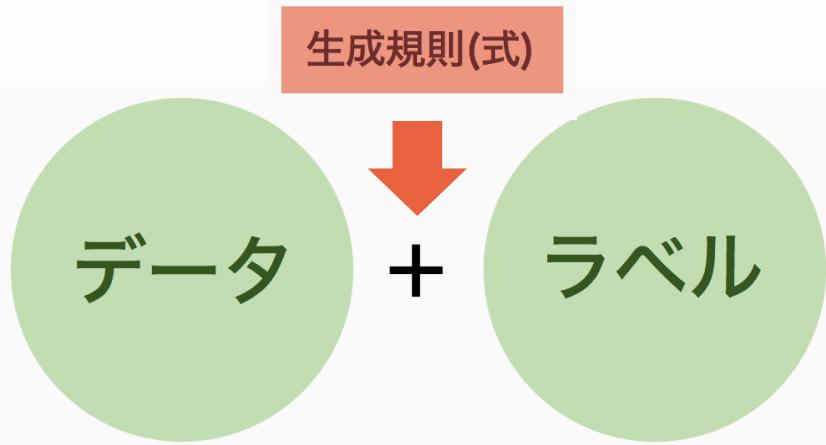


無限のデータを数式から生成

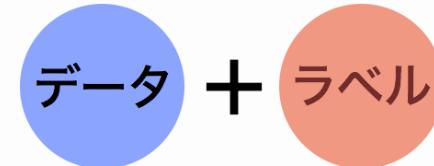


Formula-driven Approach

何らかの生成規則により「データ」と「ラベル」を作る



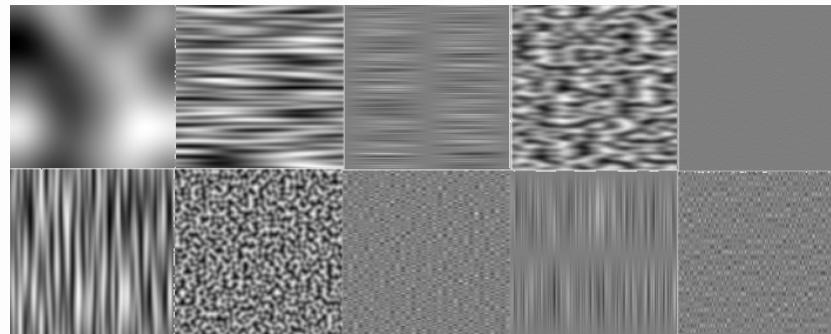
無限のデータを数式から生成
事前学習



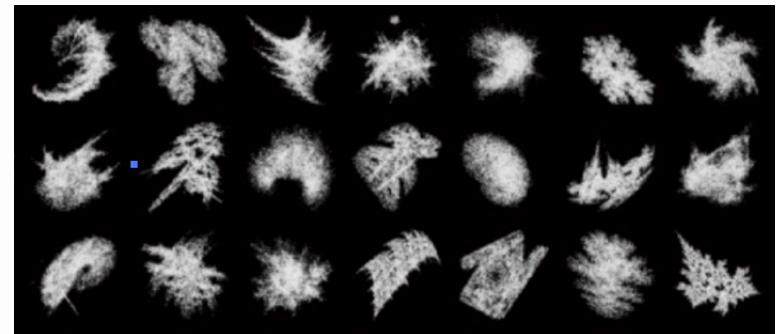
少量ラベルは人手で付与
Fine-Tuning

我々の取り組み

Perlin Noiseカテゴリを学習



フラクタルカテゴリを学習



N. Inoue, et al., Initialization Using Perlin Noise for Training Networks with a Limited Amount of Data, ICPR, 2020.

H. Kataoka, et al., Pre-training without Natural Images, ACCV, 2020. (**Best Paper Honorable Mention Award**)

Pre-training without Natural Images



自然画像を用いない事前学習

生成規則(式)



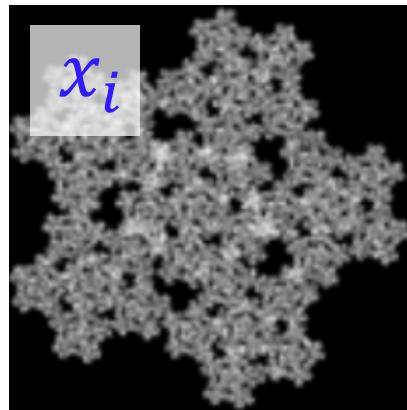
+



Pre-training without Natural Images

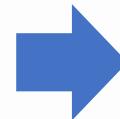


全ての画像にラベルが付与されている状況と同じ

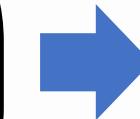


画像

$f(x_i)$

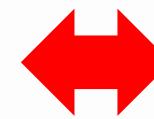


$$\begin{pmatrix} -1.61 \\ 2.90 \\ -0.50 \\ 0.11 \\ 0.52 \\ \vdots \end{pmatrix}$$



$$\begin{pmatrix} 0.01 \\ 0.03 \\ 0.02 \\ 0.91 \\ 0.01 \\ \vdots \end{pmatrix}$$

(確率)



$y_i : 4$

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{pmatrix}$$

Class 1
Class 2
Class 3
Class 4
Class 5

ラベル: フラクタルの種類

Supervised Learning

全ての画像にラベルが付与されている状況と同じ



画像

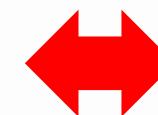
$$f(x_i)$$



$$\begin{pmatrix} -1.61 \\ 2.90 \\ -0.50 \\ 0.11 \\ 0.52 \\ \vdots \end{pmatrix} \rightarrow \begin{pmatrix} 0.01 \\ 0.03 \\ 0.02 \\ 0.91 \\ 0.01 \\ 0.02 \end{pmatrix}$$

画像表現

y_i : motorbike



$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

airplane
bus
car
motorbike
train
truck

(確率)

ラベルの例: 物体名

Pre-training without Natural Images



自然画像を用いない事前学習

生成規則(式)



+



原理



生成規則(式)

ラベルの生成 (ランダムサンプリング)

$$\Theta = \{(\theta_i, p_i)\}_{i=1}^N$$

データの生成 (IFS)

$$\text{IFS} = \{\mathcal{X}; w_1, w_2, \dots, w_N; p_1, p_2, \dots, p_N\}$$

$$w_i(\mathbf{x}; \theta_i) = \begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix} \mathbf{x} + \begin{bmatrix} e_i \\ f_i \end{bmatrix}$$

$$p_i = p(w^* = w_i) \quad \mathbf{x}_{t+1} = w^*(\mathbf{x}_t)$$

基本はこれだけ！

Iterated Function System (IFS)



1枚のフラクタル画像を生成する手順

1. 座標 x_0 をランダムに選択, $t = 0$ とする
2. 座標 x_t に点をプロット
3. 次の座標 $x_{t+1} = w^*(x_t)$ を計算
4. $t < T$ なら $t = t + 1$ として 2 に戻る

この部分でフラクタル図形になる



Iterated Function System (IFS)



1枚のフラクタル画像を生成する手順

3. 次の座標 $x_{t+1} = w^*(x_t)$ を計算

$$\text{IFS} = \{\mathcal{X}; w_1, w_2, \dots, w_N; p_1, p_2, \dots, p_N\}$$

N 個のアフィン変換とそれを選択する確率 (固定)

$$w_i(\mathbf{x}; \theta_i) = \begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix} \mathbf{x} + \begin{bmatrix} e_i \\ f_i \end{bmatrix} \quad \theta_i = (a_i, b_i, c_i, d_i, e_i, f_i)$$

3-1. アフィン変換を1つ選択して w^* とおく

ただし、 $p(w^* = w_i) = p_i$ ————— 生成される画像は

3-2. $x_{t+1} = w^*(x_t)$ を計算



Iterated Function System (IFS)



生成パラメータ $\Theta = \{(\theta_i, p_i)\}_{i=1}^N$ を 1 つ固定すれば
“似たような”画像が無限に得られる
→ これらの画像には同じ“ラベル”を付与する

$$\text{IFS} = \{\mathcal{X}; w_1, w_2, \dots, w_N; p_1, p_2, \dots, p_N\}$$

N 個のアフィン変換とそれを選択する確率 (固定)

$$w_i(\mathbf{x}; \theta_i) = \begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix} \mathbf{x} + \begin{bmatrix} e_i \\ f_i \end{bmatrix}$$

N 個のアフィン変換とそれを選択する確率 (固定)

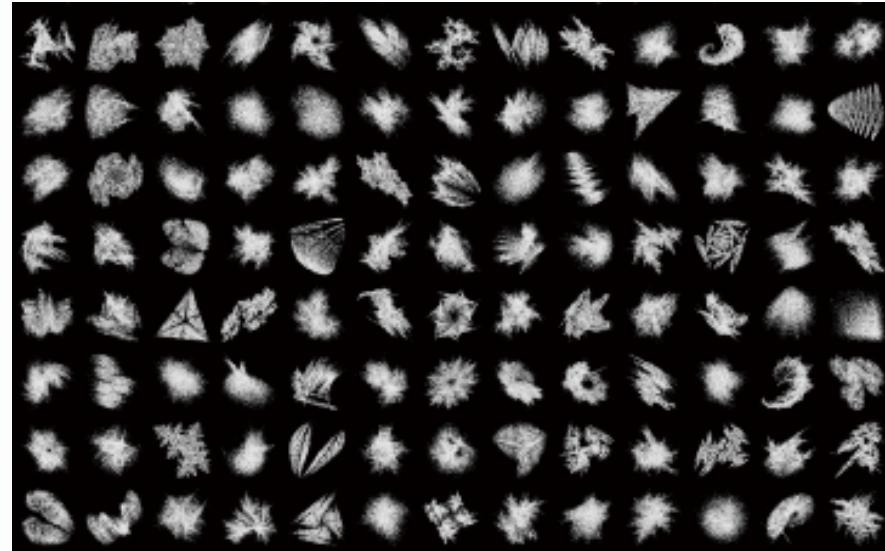


FractalDB-10k

10,000通りの生成パラメータをランダムに生成

1. N を2~8からランダムに選択
2. 各アフィン変換のパラメータを $(-1,1)$ の範囲でランダムに選択し、 $p_i = (\det A_i) / (\sum_{i=1}^N \det A_i)$ とする
回転行列
3. Filling rateが閾値以上なら採用

1万カテゴリの画像データセット
として事前学習に利用



最新の結果 (Vision Transformer)



- Fractalが最も良い、精度ではImageNetに劣る
- 最新のSSLといい勝負ができる場合もある

DeiT-Ti (Resnet18程度の大きさのVision Transformer)

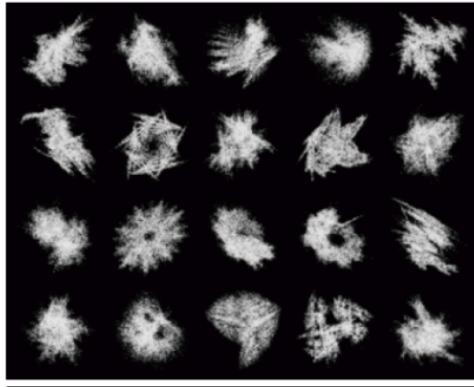
Method	Use Natural Images?	C10	C100	Cars	Flowers	VOC12	P30	Average
Jigsaw	YES	96.4	82.3	55.7	98.2	82.1	80.6	82.5
Rotation	YES	95.8	81.2	70.0	96.8	81.1	79.8	84.1
MoCov2	YES	96.9	83.2	78.0	98.5	85.3	80.8	87.1
SimCLRv2	YES	97.4	<u>84.1</u>	84.9	<u>98.9</u>	86.2	80.0	88.5
FractalDB-10k	NO	97.6	83.5	87.7	98.8	86.9	78.5	88.8

K. Nakashima, et al., Can Vision Transformers Learn without Natural Images?, arXiv2103.13023.

最新の結果 (Vision Transformer)

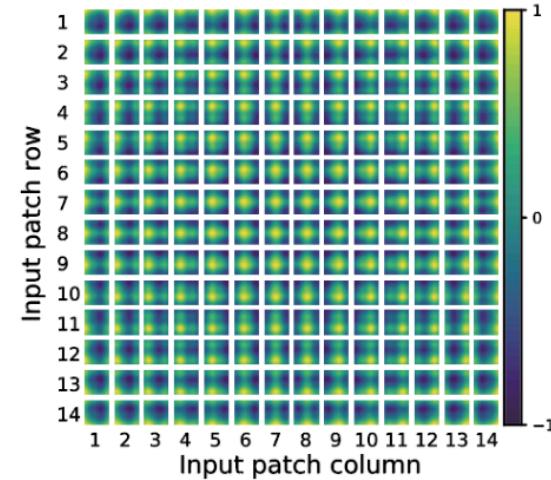
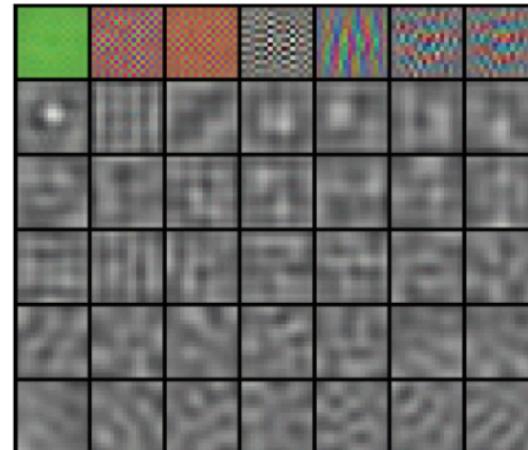


フィルタはある程度学習できていそう？



FractalDB (Generated Images)

→
FDL



最新の結果 (Vision Transformer)

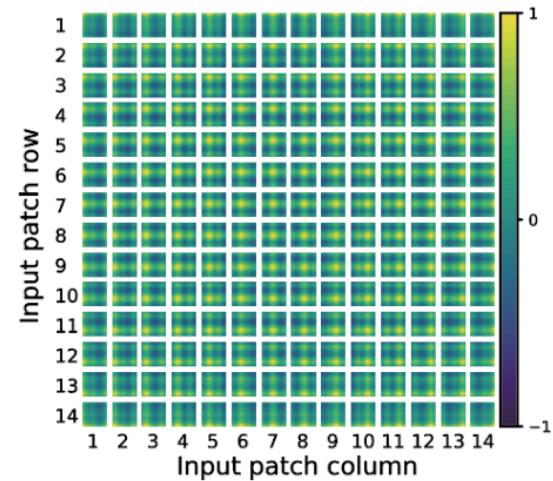
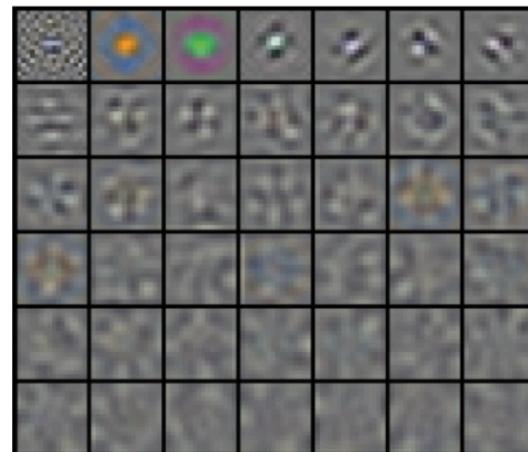


自己教師あり学習(SimCLR)で得られたフィルター



ImageNet (Natural Images)

SSL



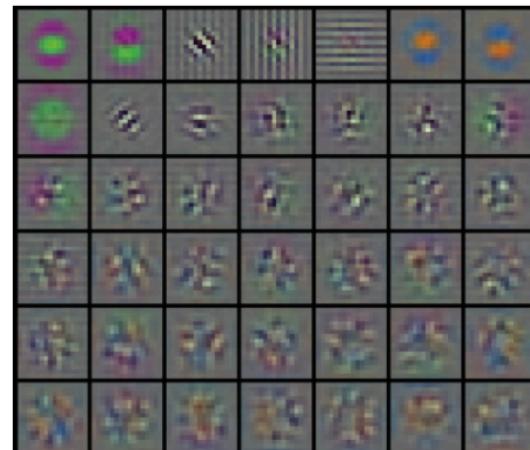
最新の結果 (Vision Transformer)



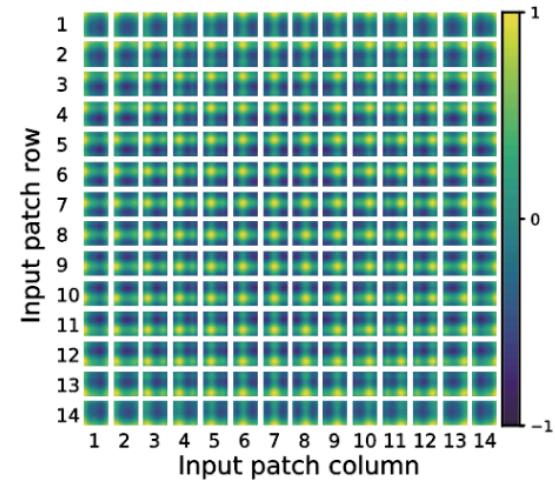
教師あり学習で得られたフィルター



→
SL



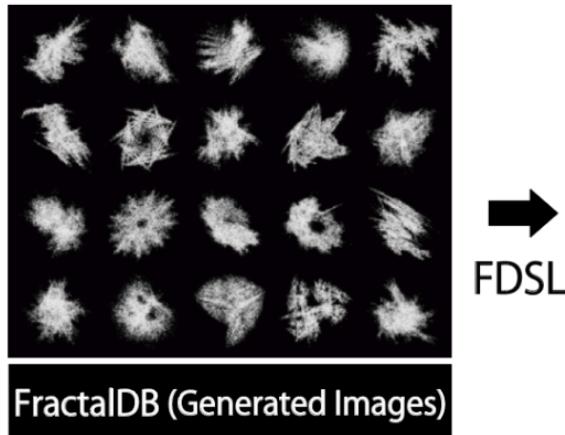
ImageNet (Natural Images)



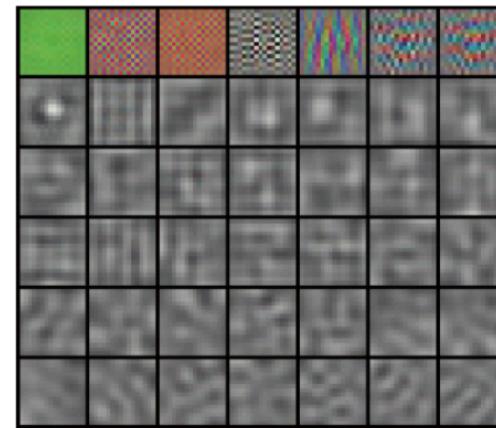
最新の結果 (Vision Transformer)



フィルタはある程度学習できていそう？

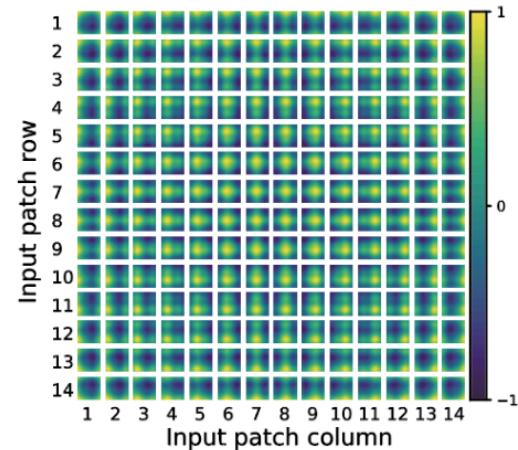


FDSL



色が欲しい

※今はcolor jitteringのデータ拡張で色がついている



見た目はOK



cvpaper.challenge

Pre-training without Natural Images

Asian Conference on Computer Vision (ACCV) 2020

Best Paper Honorable Mention Award

Oral Presentation, The paper got 3 strong accepts

Hirokatsu Kataoka¹ Kazushige Okayasu^{1,2} Asato Matsumoto^{1,3} Eisuke Yamagata⁴

Ryosuke Yamada^{1,2} Nakamasa Inoue⁴ Akio Nakamura² Yutaka Satoh^{1,3}

1: AIST 2: TDU 3: Univ. of Tsukuba 4: TITech

Paper

Code

Dataset

Oral

Poster

Supp. Mat.

Related Work

<https://hirokatsukataoka16.github.io/Pretraining-without-Natural-Images/>

まとめ



第1部 背景と問題設定

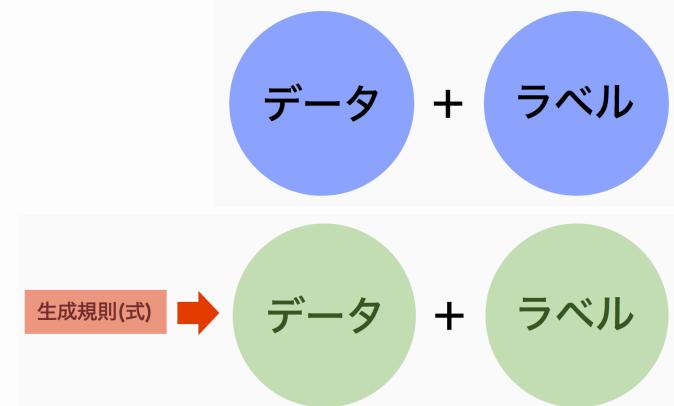
- 様々な学習フレームワークに対する問題設定の整理

第2部 Self-Supervised Learning

- ラベルなしデータでの事前学習

第3部 Formula-driven Approach

- フラクタル画像による事前学習



第1部 学習フレームワーク

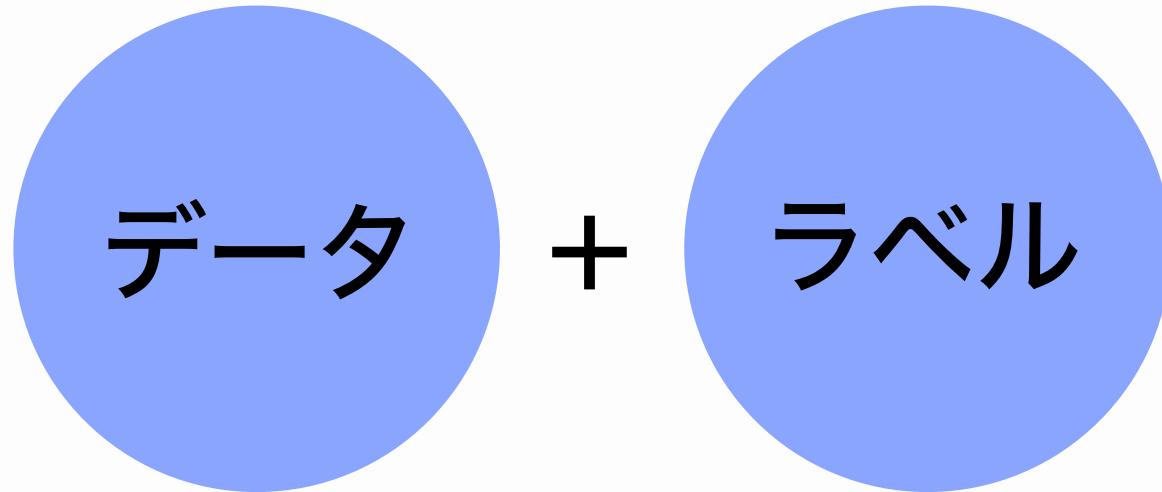
様々な条件下での「学習」が提案されている

ラベルの量	Supervised	Semi-supervised	Unsupervised
データの量	(Many-Shot)	Few-Shot	Zero-Shot
ラベルの完全さ	Strongly-	Weakly-	Randomly-
ラベル付与方式	(Manually-)	Self-supervised	
実データの有無	Data-driven	Formula-driven	

第2部 Self-supervised Learning

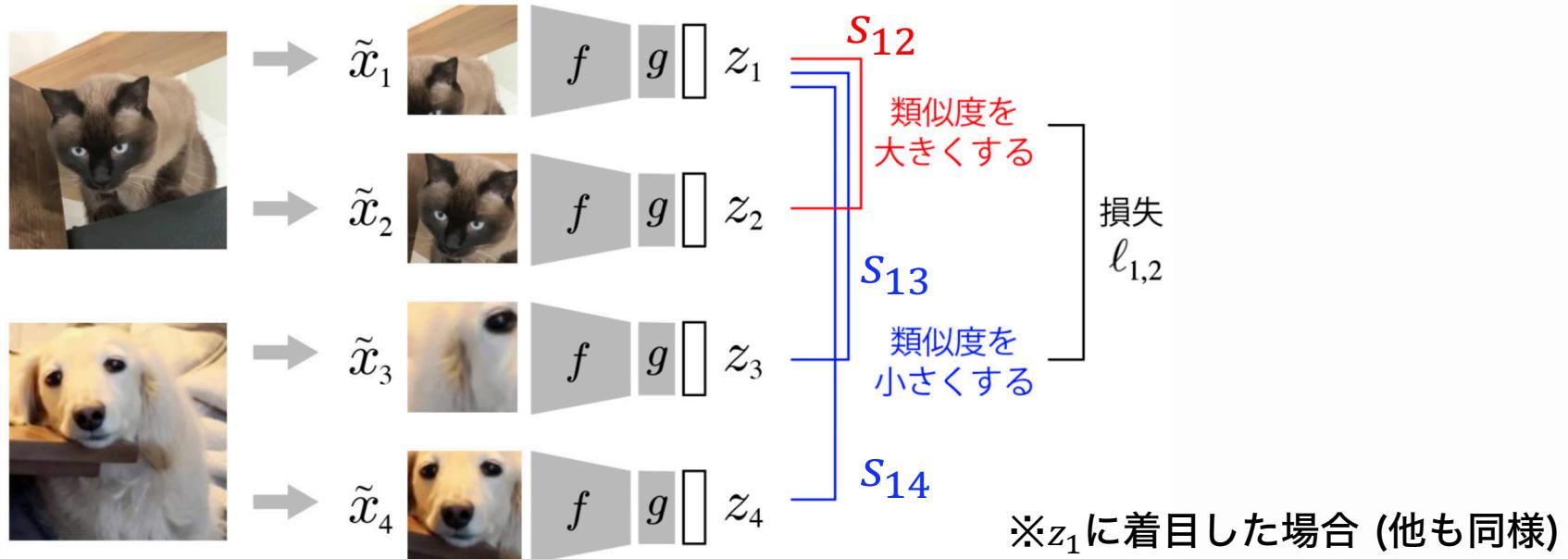


自己的に付与されたラベルを用いた学習



Contrastive Learning

2枚の画像の場合



第3部 Formula-driven Approach



「データ」も「ラベル」も何らかの規則で生成できる？

生成規則(式)



データ

+

ラベル

Pre-training without Natural Images



自然画像を用いない事前学習

生成規則(式)



+



今日お話ししなかった内容

- Few-Shot Learning, Meta Learning
 - 少量データでの学習の工夫はたくさんある (論文も書きやすい)
 - タスクごとに特化した内容であることが多い
- 実社会応用ではどうすべきか？

事前学習が Yes → Fine-Tuning → あと一歩なら工夫も検討
可能か？ No → スクラッチ学習とカーネルSVMを比較検討

まとめ



第1部 背景と問題設定

- 様々な学習フレームワークに対する問題設定の整理

第2部 Self-Supervised Learning

- ラベルなしデータでの事前学習

第3部 Formula-driven Approach

- フラクタル画像による事前学習

