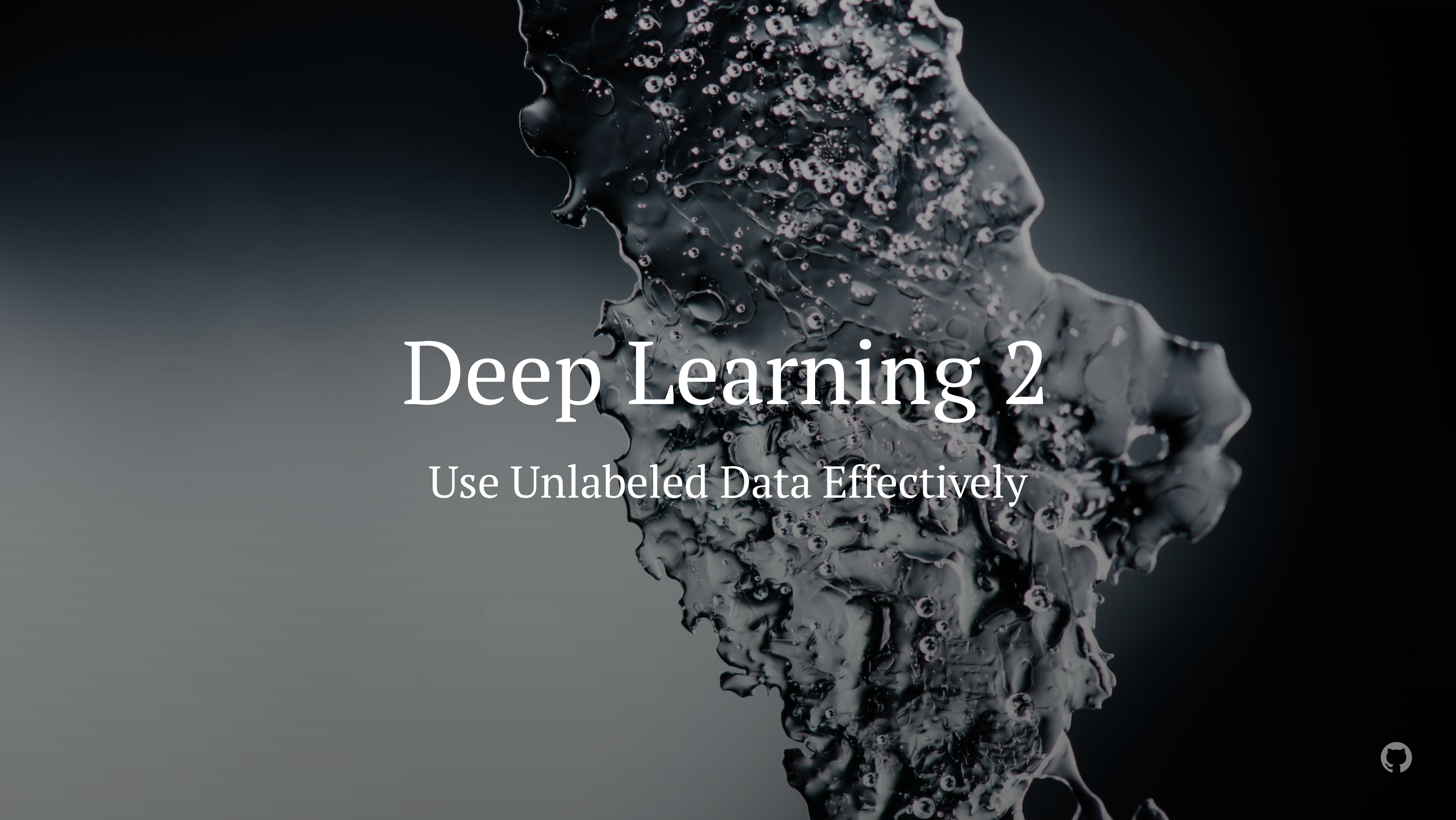
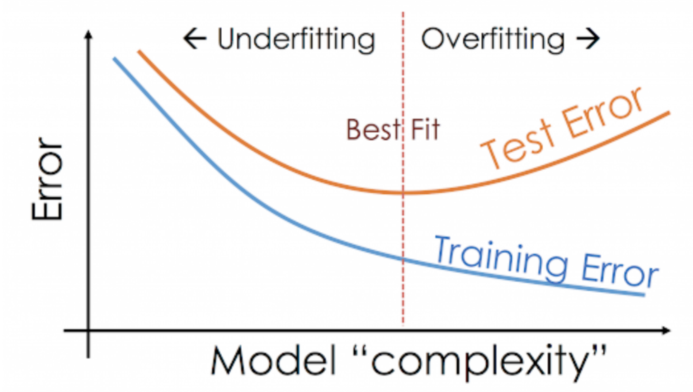


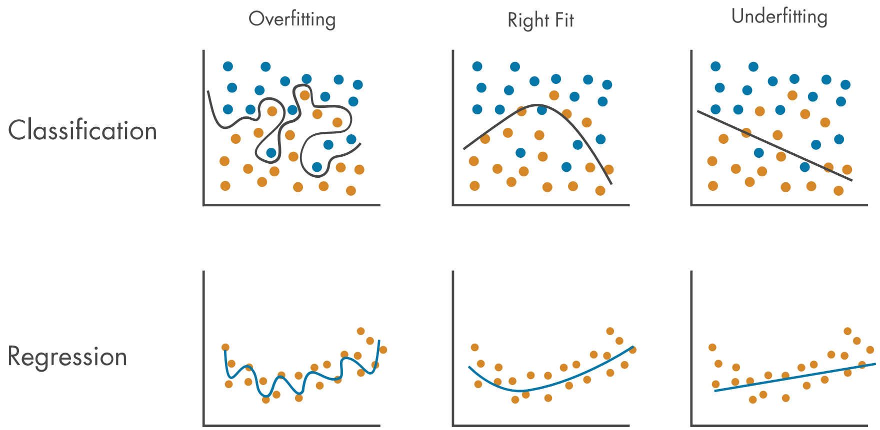
Deep Learning 2

Use Unlabeled Data Effectively

Overfitting, Generalization and Robustness



cite: Overfitting And Underfitting in Machine Learning



cite: What Is Overfitting?

- 當模型複雜度超過訓練資料集複雜度時，就容易發生 Overfitting
- 當訓練資料集複雜度超過模型複雜度時，就容易發生 Underfitting

Overfitting, Generalization and Robustness

避免 Overfitting 的方法

- 使用更多訓練資料
- **資料擴增（Data Augmentation）**
- Dropout
- L1/L2 Regularization
- 降低模型複雜度

避免 Underfitting 的方法

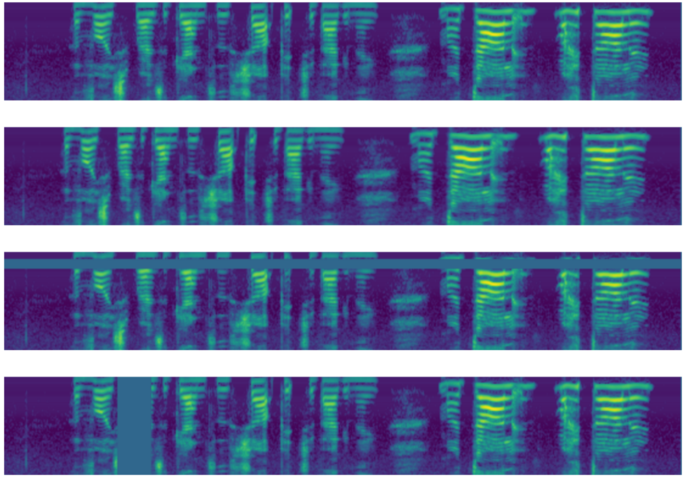
- 大力出奇蹟

  疊更多參數能解決一切
- 錢是萬能的

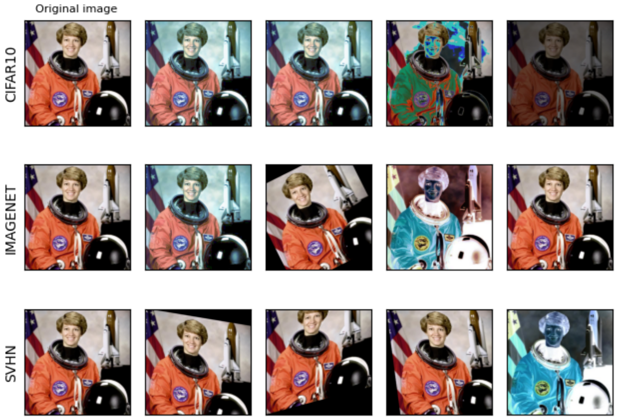
  用更強的硬體訓練更大的模型

Overfitting, Generalization and Robustness

Data Augmentation



cite: SpecAugment



cite: Pytorch: Illustration of transforms

Text Augmentation

- Replacing Words or Phrases with Their Synonyms
- Back Translation
- Text Generation
- Mixing-based Text Augmentation

Overfitting, Generalization and Robustness

Generalization

- 模型應用到其他資料分佈的能力

> The classic approach towards the assessment of any machine learning model revolves around the evaluation of its generalizability i.e. its performance on unseen test scenarios.

Robustness

- 模型對抗干擾的能力

> Evaluating such models on an available non-overlapping test set is popular, yet significantly limited in its ability to explore the model's resilience to outliers and noisy data / labels (i.e. robustness).

cite: Generalizability vs. Robustness: Adversarial Examples for Medical Imaging, Robustness vs Generalization

Garbage In, Garbage Out

cite: 盡信資料，不如無資料

- **資料稀缺**

  使用預訓練（Pretrain）好的模型微調（fine-tune）

- **缺乏標記**

  使用 Self/Semi Supervised Learning

- 標記錯誤 Awesome-Noisy-Labels
- 分佈不平衡

  Classification on imbalanced data

Unlabeled Data

Supervised

- 資料需要有人工標記
- 標記成本高昂、費時
- 無法輕易擴展資料集

Semi-Supervised

訓練資料（同時使用）：

- 少量帶有人工標記的資料
- 大量的未標記資料

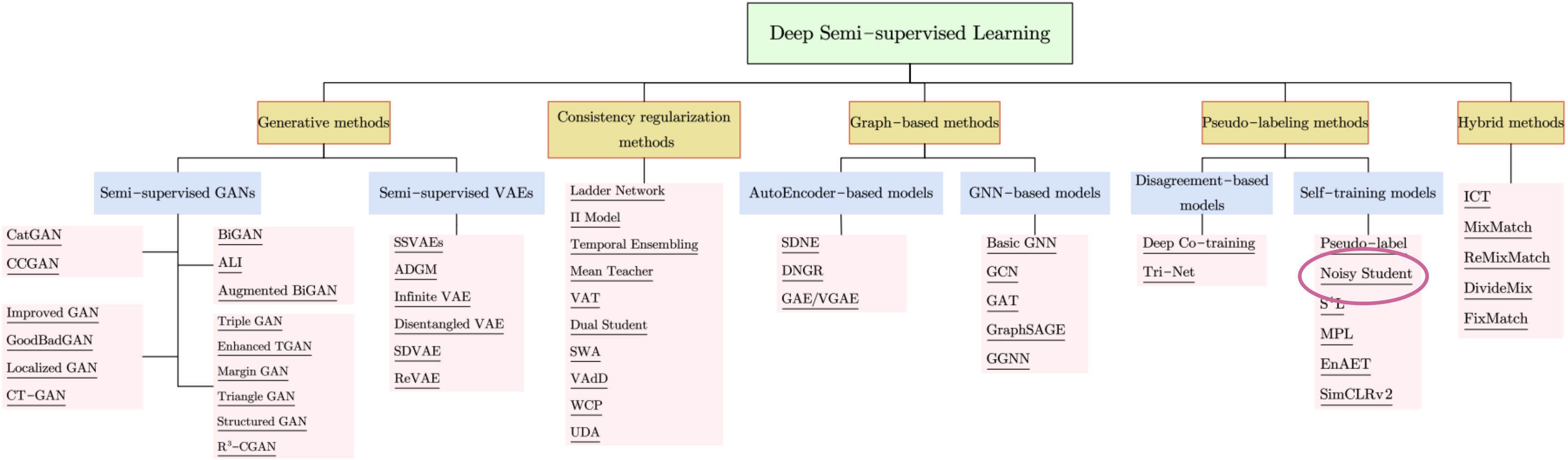
Representation Learning

- 任務目標：抽取良好的資料特徵
- 可以是 Supervised or Unsupervised

Self-Supervised

- 屬於 Unsupervised Learning 的一類
- 從大量的未標記資料中學習特徵
- 訓練好的模型應用在多個下游任務

Semi-Supervised



cite: A Survey on Deep Semi-supervised Learning

Semi-Supervised

Noisy Student Training

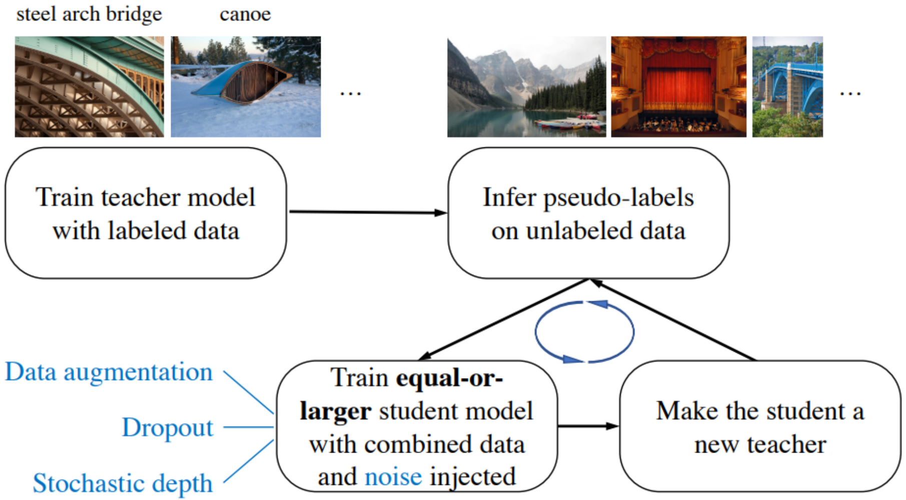


Figure 1: Illustration of the Noisy Student Training. (All shown images are from ImageNet.)

cite: Self-training with Noisy Student improves ImageNet classification

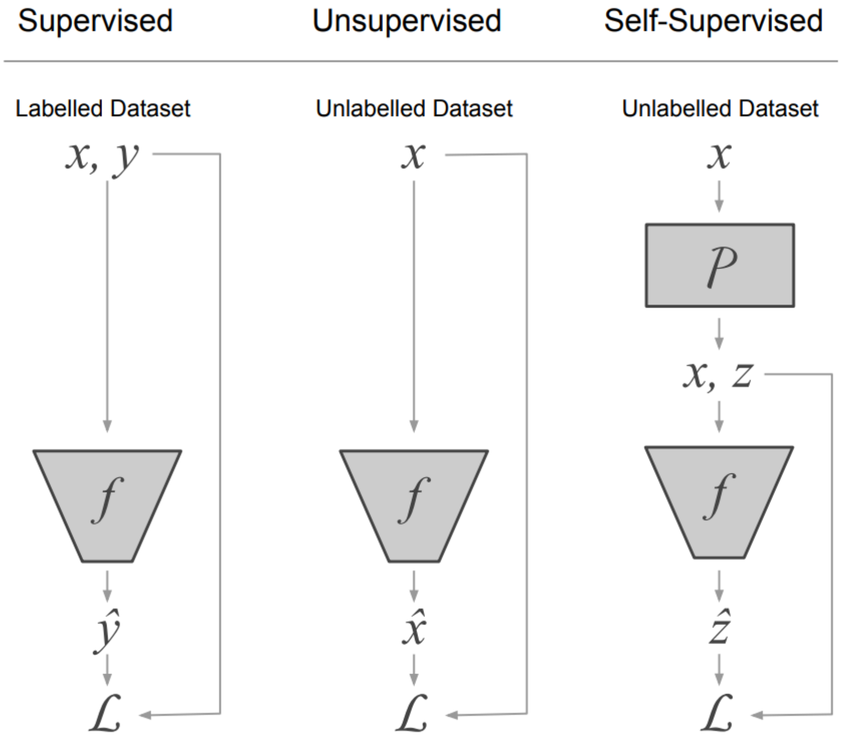
Noisy Student Training

1. 在有人工標記的資料集 $\mathcal{D}$ 上訓練模型（Teacher）。
2. 在大量的未標記資料 $\mathcal{U}$ 上使用 Teacher Model 推斷標籤得到 $\tilde{\mathcal{U}}$。
3. 使用通過資料增強的 $\mathcal{D} \cup \tilde{\mathcal{U}}$ 訓練新的模型（Noisy Student）。
4. 回到步驟 2，並使用 Noisy Student 取代 Teacher。

Model / Unlabeled Set Size	1.3M	130M
EfficientNet-B5	83.3%	84.0%
Noisy Student Training (B5)	**83.9%**	**85.1%**
student w/o Aug	83.6%	84.6%
student w/o Aug, SD, Dropout	83.2%	84.3%
teacher w. Aug, SD, Dropout	83.7%	84.4%

cite: Self-training with Noisy Student improves ImageNet classification

Self-Supervised



cite: Self-Supervised Representation Learning: Introduction, Advances and Challenges

Audio

- CPC
- Wav2Vec/2.0
- HuBERT

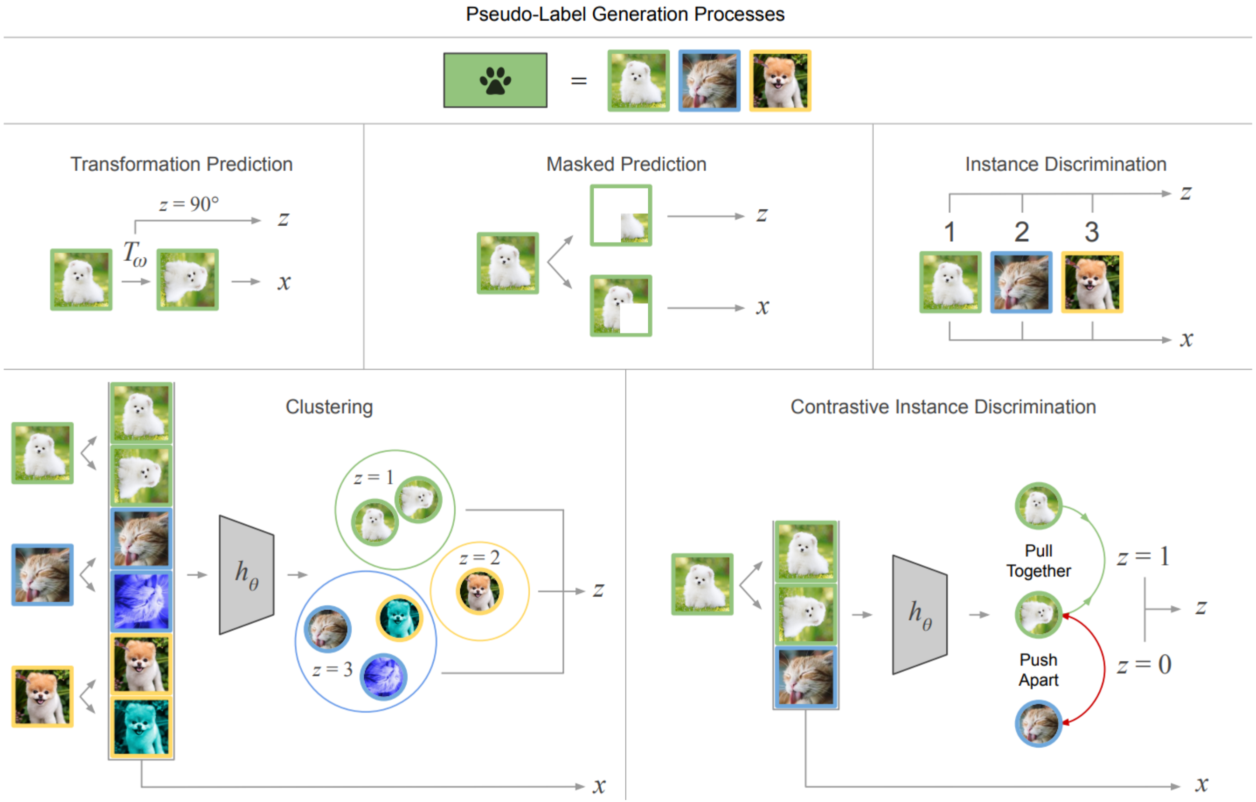
Text

- BERT

Image

- SimCLR
- BYOL
- DINOv1/v2

Self-Supervised



cite: Self-Supervised Representation Learning: Introduction, Advances and Challenges

Self-Supervised

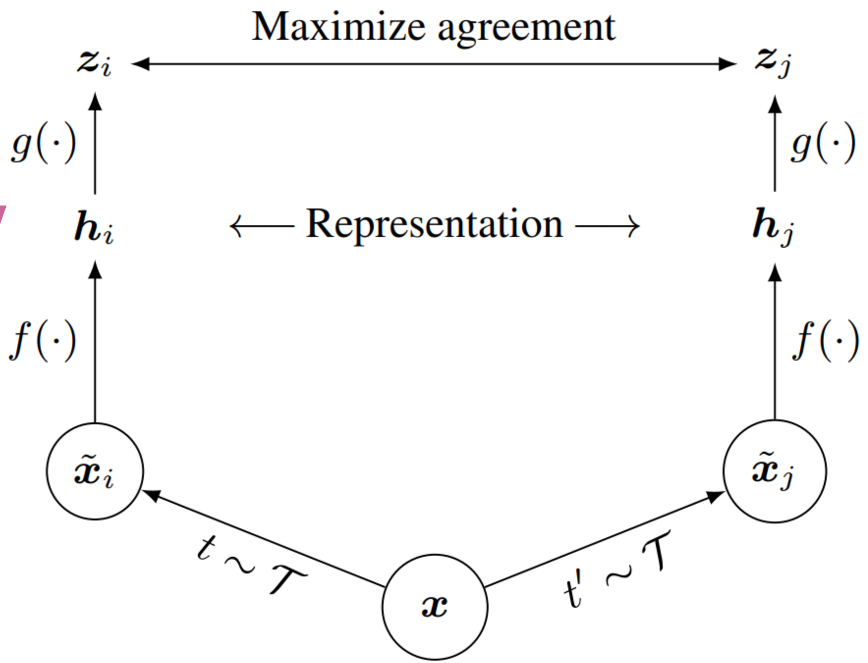
SimCLR

- Contrastive Learning

$$l(i,j) = -log\frac{\exp(s_{i,j}/\tau) \blacktriangle}{\sum_{k=1}^{2N} \mathbf{1}_{k\neq i}\exp(s_{i,k}/\tau) \blacktriangledown}$$

$s_{i,j}$ is the cos similarity between z_i & z_j.

- 正樣本特徵越接近越好
- 負樣本特徵差越多越好
- $g(\cdot)$ 是輔助骨幹網路 $f(\cdot)$ 訓練的 projector，訓練完就會丟棄



cite: A Simple Framework for Contrastive Learning of Visual Representations

Self-Supervised

SimCLR

- 使用資料擴增建構正樣本
- 正樣本特徵**越接近**越好
- 負樣本特徵**差越多**越好
- MLP=$g(\cdot)$，輔助訓練的 projector
- CNN=$f(\cdot)$，骨幹網路

cite：Advancing Self-Supervised and Semi-Supervised Learning with SimCLR

Self-Supervised Contrastive Learning

Issues and Solutions

1. 需要大 Batch Size 才有足夠多的負樣本用來提昇性能
   - 重複使用 Memory Bank 暫存的特徵當成正負樣本目標
   - Momentum-updated Encoder 減少計算正負樣本目標消耗的記憶體
2. 需要從無標記的資料中建構正負樣本
   - 設計無須使用負樣本的方法

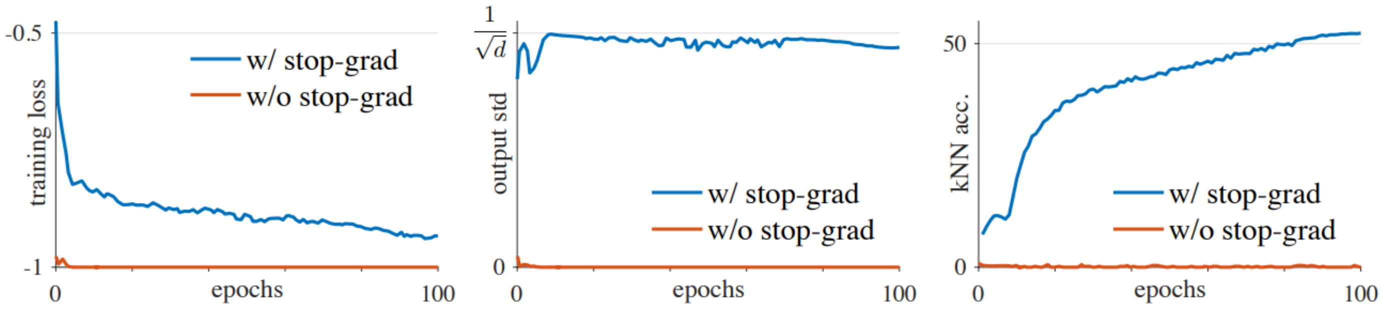
     e.g. BYOL, SimSiam, DINOv1/v2

Contrastive Learning Without Negative Sample

Why Do We Need Negative Samples?

Contrastive Learning = 正樣本特徵越接近越好 + ~~負樣本特徵差越多越好~~

→ 只要模型輸出**常數**就能輕鬆達成，又稱為 Collapse (Trivial Solution)

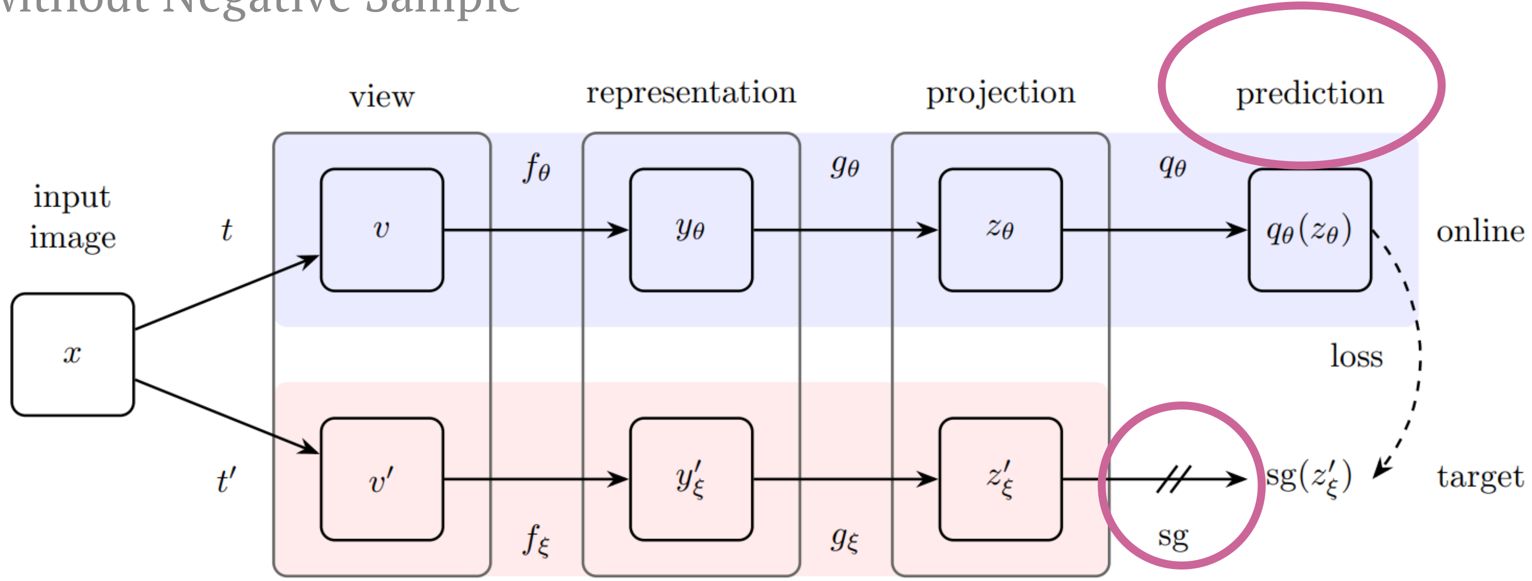


	acc. (%)
w/ stop-grad	67.7±0.1
w/o stop-grad	0.1

cite: Exploring Simple Siamese Representation Learning

Bootstrap Your Own Latent

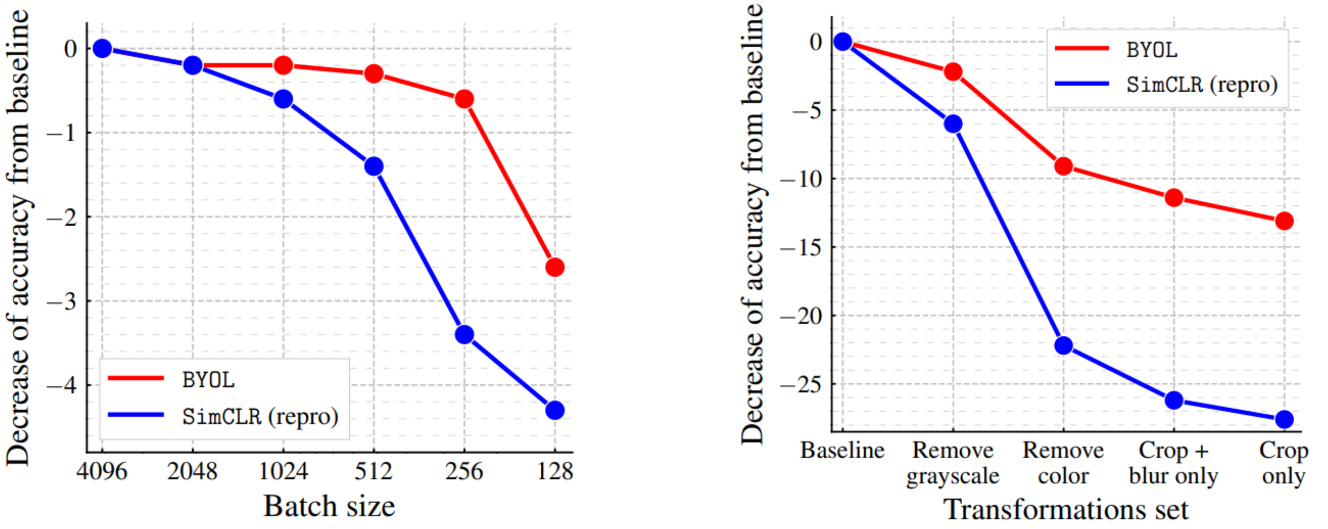
SSL without Negative Sample



cite: Bootstrap Your Own Latent A New Approach to Self-Supervised Learning

Bootstrap Your Own Latent

SSL without Negative Sample



(a) Impact of batch size

(b) Impact of progressively removing transformations

Figure 3: Decrease in top-1 accuracy (in % points) of BYOL and our own reproduction of SimCLR at 300 epochs, under linear evaluation on ImageNet.

cite: Bootstrap Your Own Latent A New Approach to Self-Supervised Learning

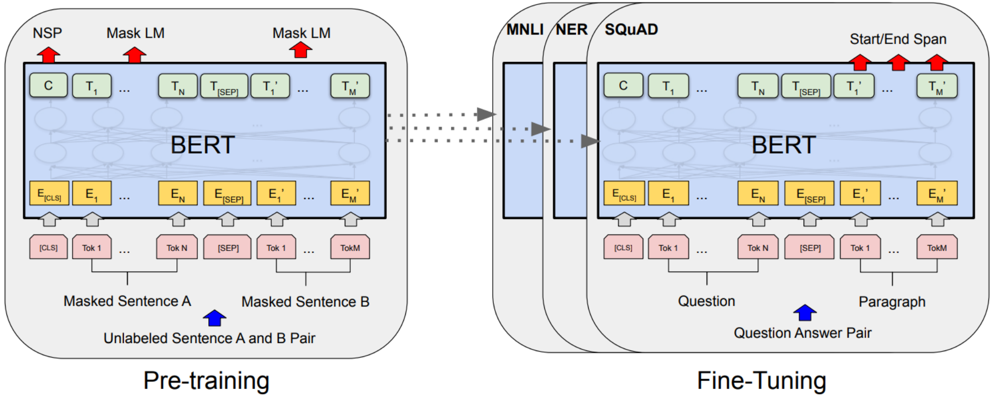
SSL Without Negative Sample

method	batch size	negative pairs	momentum encoder	100 ep	200 ep	400 ep	800 ep
SimCLR (repro.+)	4096	✓		66.5	68.3	69.8	70.4
MoCo v2 (repro.+)	**256**	✓	✓	67.4	69.9	71.0	72.2
BYOL (repro.)	4096		✓	66.5	**70.6**	**73.2**	**74.3**
SwAV (repro.+)	4096			66.5	69.1	70.7	71.8
SimSiam	**256**			**68.1**	70.0	70.8	71.3

cite: Exploring Simple Siamese Representation Learning

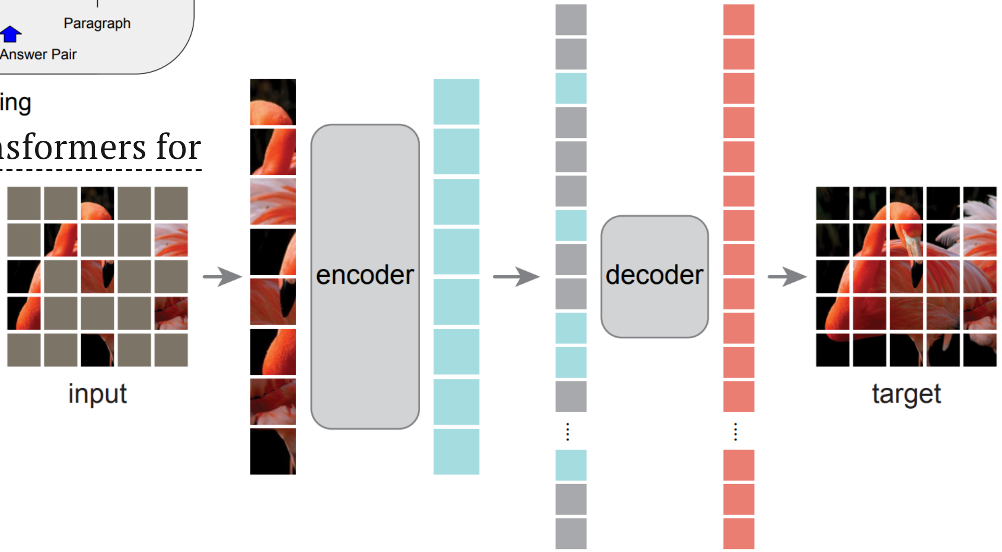
- 實作簡單
- 性能優異
- 無需定義負樣本
- 若超參數沒設好，有發生 collapse 的風險

Masked AutoEncoder



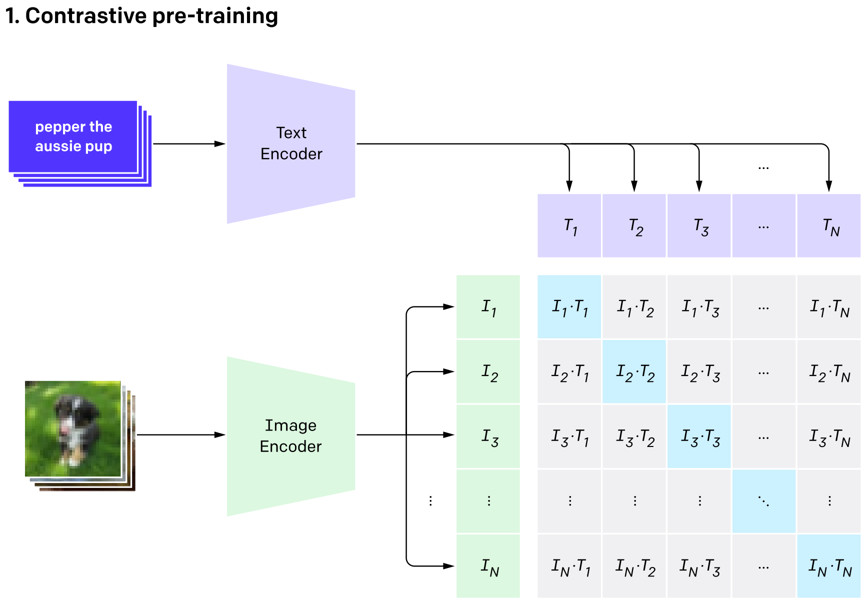
cite: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

- 穩定、容易訓練
- 不會 collapse
- 適合 Transformer 學習全域特徵



cite: Masked Autoencoders Are Scalable Vision Learners

Representation Learning



cite: Learning Transferable Visual Models From Natural Language Supervision

Text & Text

- Sentence-BERT

Text & Image

- CLIP
- CoCa

Multimodal

- Data2Vec
- ImageBind

Survey

Self-Supervised

- Self-Supervised Representation Learning
- Contrastive Representation Learning
- Awesome Self-Supervised Learning
- 自監督式學習 Self-Supervised Learning for Computer Vision 之概述
- A Survey on Contrastive Self-supervised Learning, 2020
- Self-Supervised Representation Learning: Introduction, Advances and Challenges, 2021
- Self-Supervised Learning for Videos: A Survey, 2022
- Survey on Self-Supervised Learning: Auxiliary Pretext Tasks and Contrastive Learning Methods in Imaging, 2022
- Self-Supervised Speech Representation Learning: A Review, 2022
- A Survey of Self-Supervised Learning from Multiple Perspectives: Algorithms, Theory, Applications and Future Trends, 2023

Survey

Semi-Supervised

- An Overview of Deep Semi-Supervised Learning, 2020
- A Survey on Deep Semi-supervised Learning, 2021