CS331. Thị giác máy tính nâng cao

Self-Supervised & Contrastive Learning

Phần A – Giới thiệu & Đặt vấn đề

Self-Supervised & Contrastive Learning

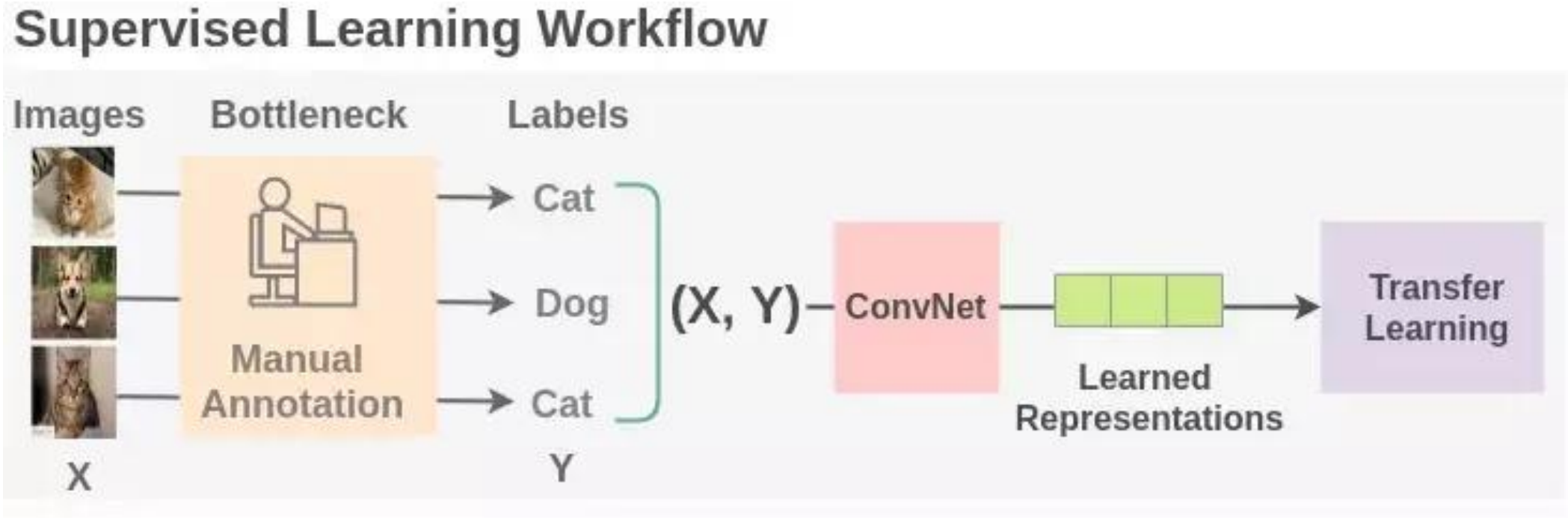
- Self-Supervised Learning (hay còn gọi là SSL)
- Contrastive Learning

Supervised learning

- Supervised learning: cần dữ liệu ảnh + nhãn.
- Ví dụ:
  - ImageNet có 1.2 triệu ảnh, gán nhãn bởi con người. Nhưng:
  - Tốn kém: gán nhãn hàng triệu ảnh rất đắt.
  - Khó mở rộng: nhiều lĩnh vực không có nhãn, ví dụ y tế, video.

→ Chúng ta có thể học từ dữ liệu không nhãn không?

Supervised learning



Tổng quan về Self-supervised representation learning (học tự giám sát)

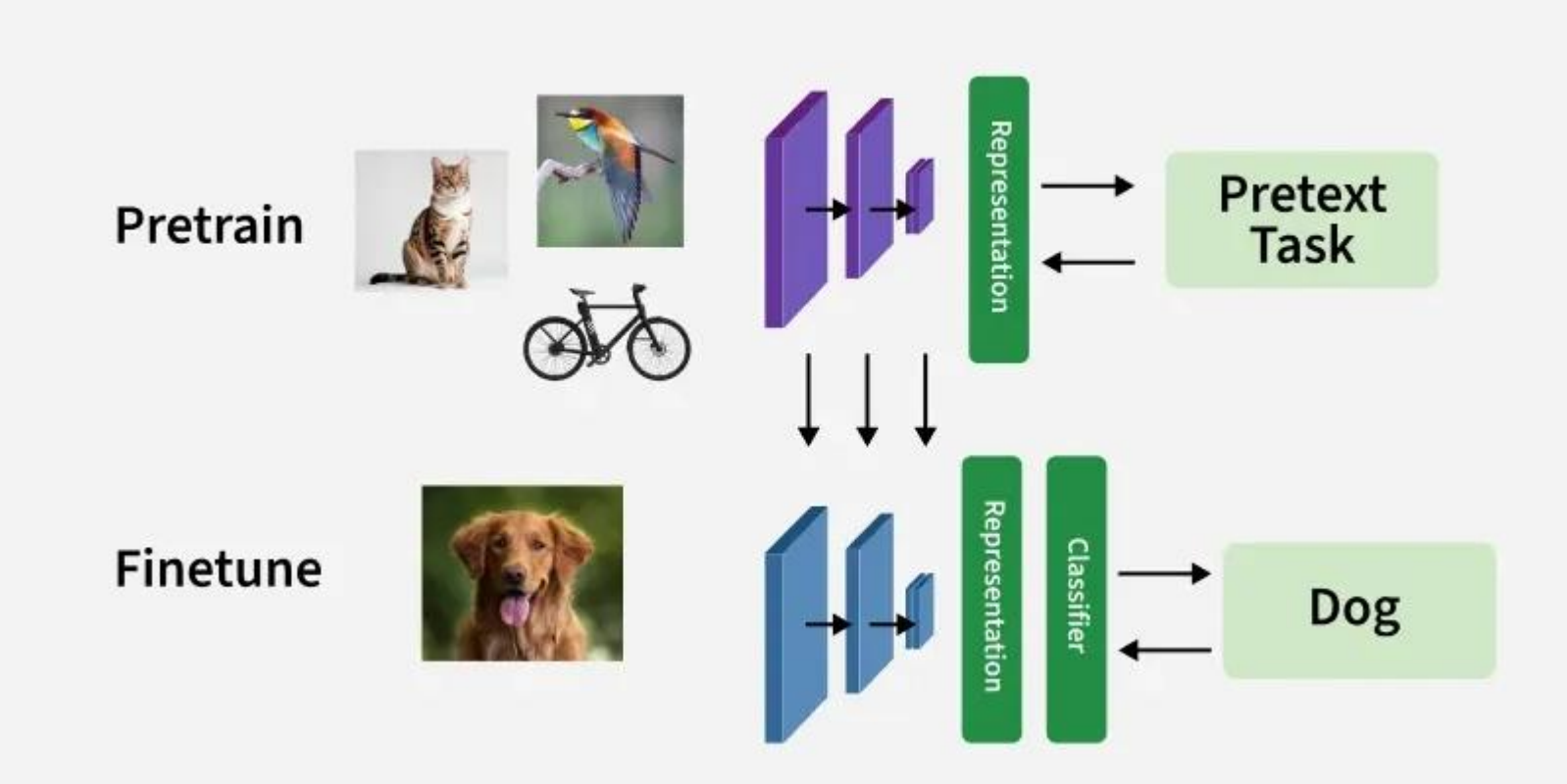
Vấn đề của supervised learning

- Tốn tài nguyên: cần nhiều nhãn.
- Không linh hoạt: chỉ giải quyết task cụ thể.
- Overfitting domain: nếu dataset khác đi, model yếu.

Ý tưởng Self-Supervised Learning (SSL)

- SSL tạo ra nhiệm vụ phụ (một bài toán giả, auxiliary or pretext task) từ dữ liệu không nhãn → để ép mô hình học cách biểu diễn (representation) có ý nghĩa.
- Ví dụ: che một phần ảnh và yêu cầu mô hình đoán phần che.
- Mặc dù không cần nhãn, nhưng mô hình học được cách hiểu cấu trúc ảnh.
- Sau khi huấn luyện xong, ta dùng representation này cho bài toán thực (downstream task) như classification, detection.

Ý tưởng Self-Supervised Learning (SSL)



[Self-Supervised Learning (SSL) - GeeksforGeeks]

Pretext task trong ảnh

- Các pretext task phổ biến:
  - **Colorization**: chuyển ảnh xám thành ảnh màu.
  - **Jigsaw puzzle**: xáo trộn patch ảnh rồi yêu cầu mô hình ghép lại.
  - **Inpainting**: che ô vuông rồi yêu cầu dự đoán pixel.
  - **Rotation prediction**: xoay ảnh 0°, 90°, 180°, 270° và yêu cầu mô hình nhận biết.
- Tất cả đều không cần nhãn.

Pretext task trong video

- Trong video:
  - Dự đoán frame tiếp theo.
  - Xác định thứ tự frame.
  - Khớp audio với hình ảnh.

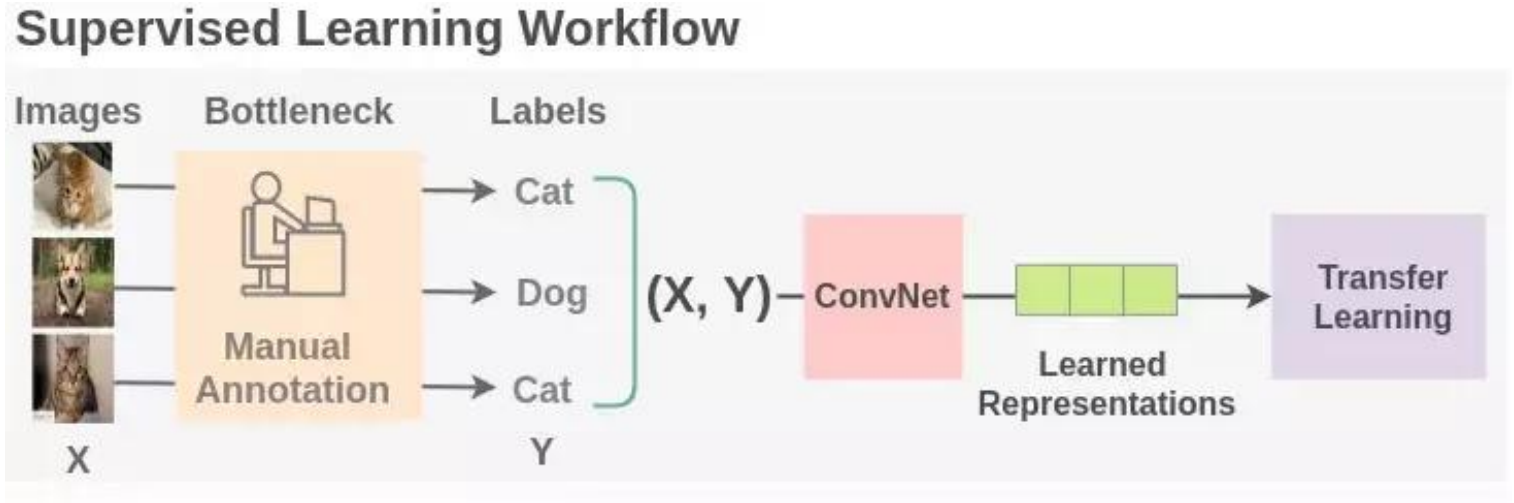
→ Các nhiệm vụ này giúp mô hình học representation temporal và multimodal.

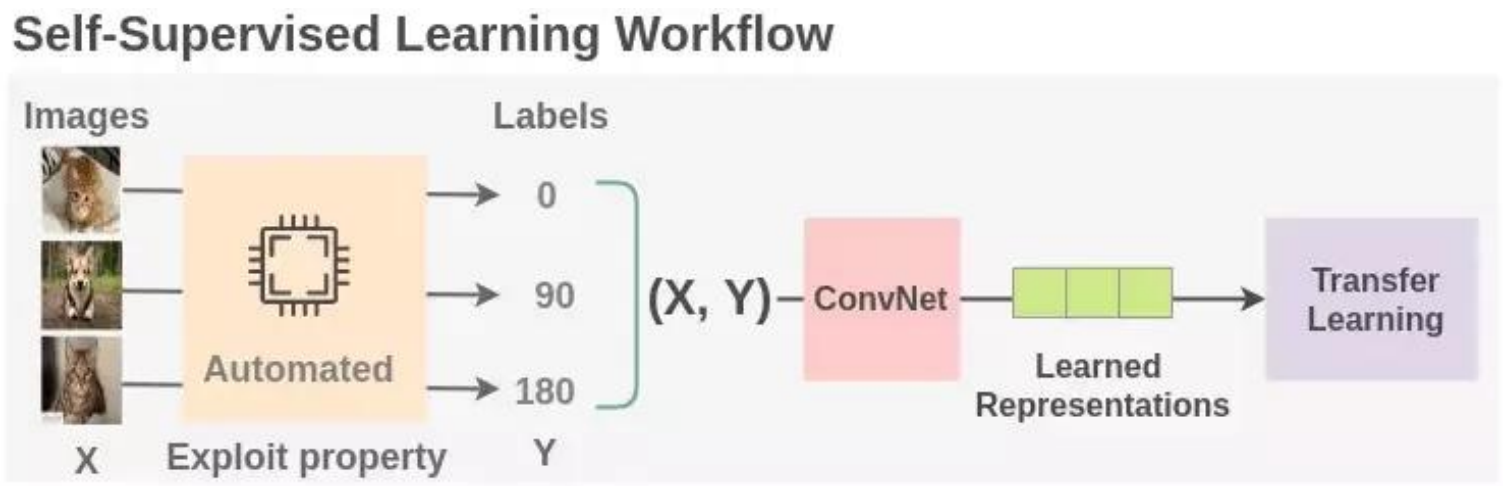
Kết quả của SSL

- Representation học được từ SSL rất mạnh, đôi khi còn mạnh hơn supervised khi transfer sang task mới.

→ Đây là lý do các mô hình foundation hiện nay (như CLIP, DINO, MAE) đều dựa vào SSL ở giai đoạn pretraining.

Supervised learning vs SSL





[Tổng quan về Self-supervised representation learning (học tự giám sát)]

Phần B – Contrastive Learning

Contrastive Learning là gì?

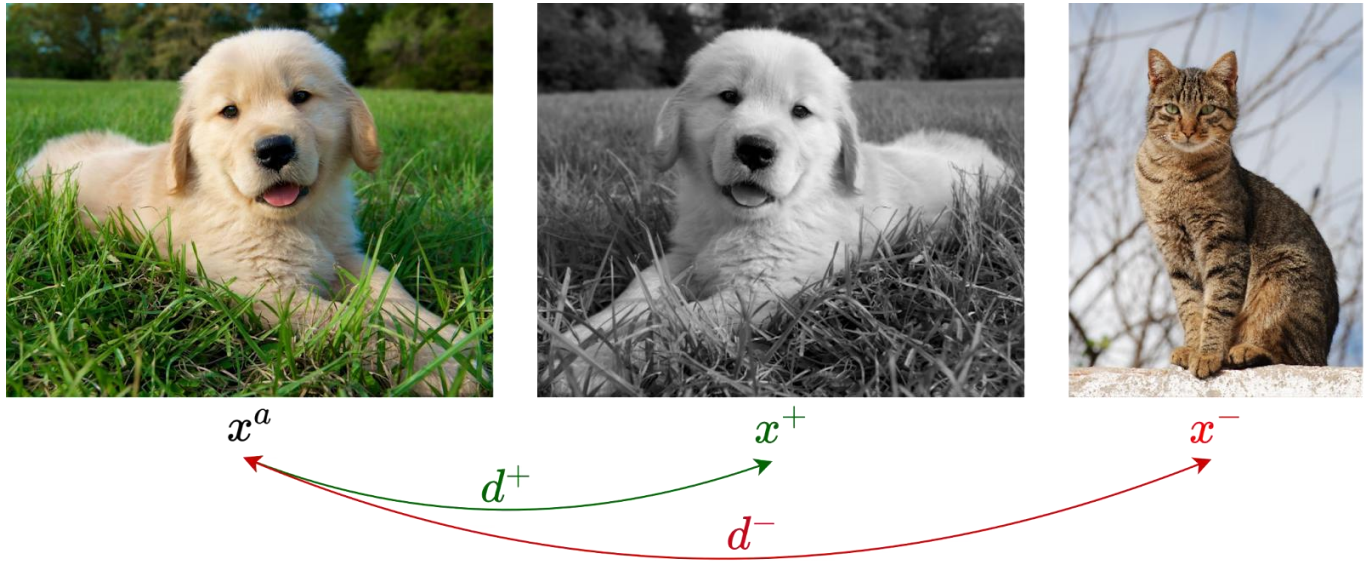
- Contrastive learning là một dạng SSL.
- Ý tưởng:
  - Hai ảnh giống nhau (positive pair) → embedding gần nhau.
  - Hai ảnh khác nhau (negative pair) → embedding xa nhau.
- Công thức loss phổ biến: *InfoNCE loss*.

Positive & Negative pairs

Ví dụ:

- Ảnh con mèo + phiên bản augment (crop, rotate) → positive pair.
- Ảnh con mèo + ảnh con chó → negative pair.

- Mô hình được huấn luyện để kéo gần vector mèo–mèo, đẩy xa mèo–chó.



InfoNCE Loss

- Công thức InfoNCE:

$$L = -\log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_k \exp(sim(z_i, z_k)/\tau)}$$

Trong đó:

- z_i, z_j =embedding của positive pair.
- z_k =embedding của các negative.
- τ= temperature → điều chỉnh độ "sắc nét" của phân phối.

Ý nghĩa: tối đa hóa tương đồng của positive, tối thiểu hóa với tất cả negative

InfoNCE Loss

- Công thức InfoNCE:

$$L = -\log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_k \exp(sim(z_i, z_k)/\tau)}$$

- InfoNCE huấn luyện mô hình sao cho positive có xác suất cao nhất trong softmax.
- InfoNCE Loss giúp học biểu diễn bằng cách tối đa hóa tương đồng của cặp positive và giảm tương đồng với các cặp negative, dựa trên softmax phân biệt.

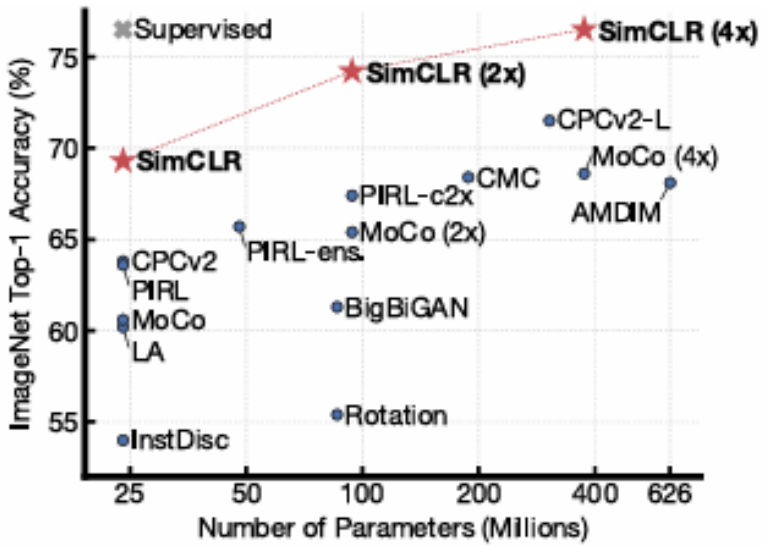
SimCLR (2020)

A Simple Framework for Contrastive Learning of Visual Representations

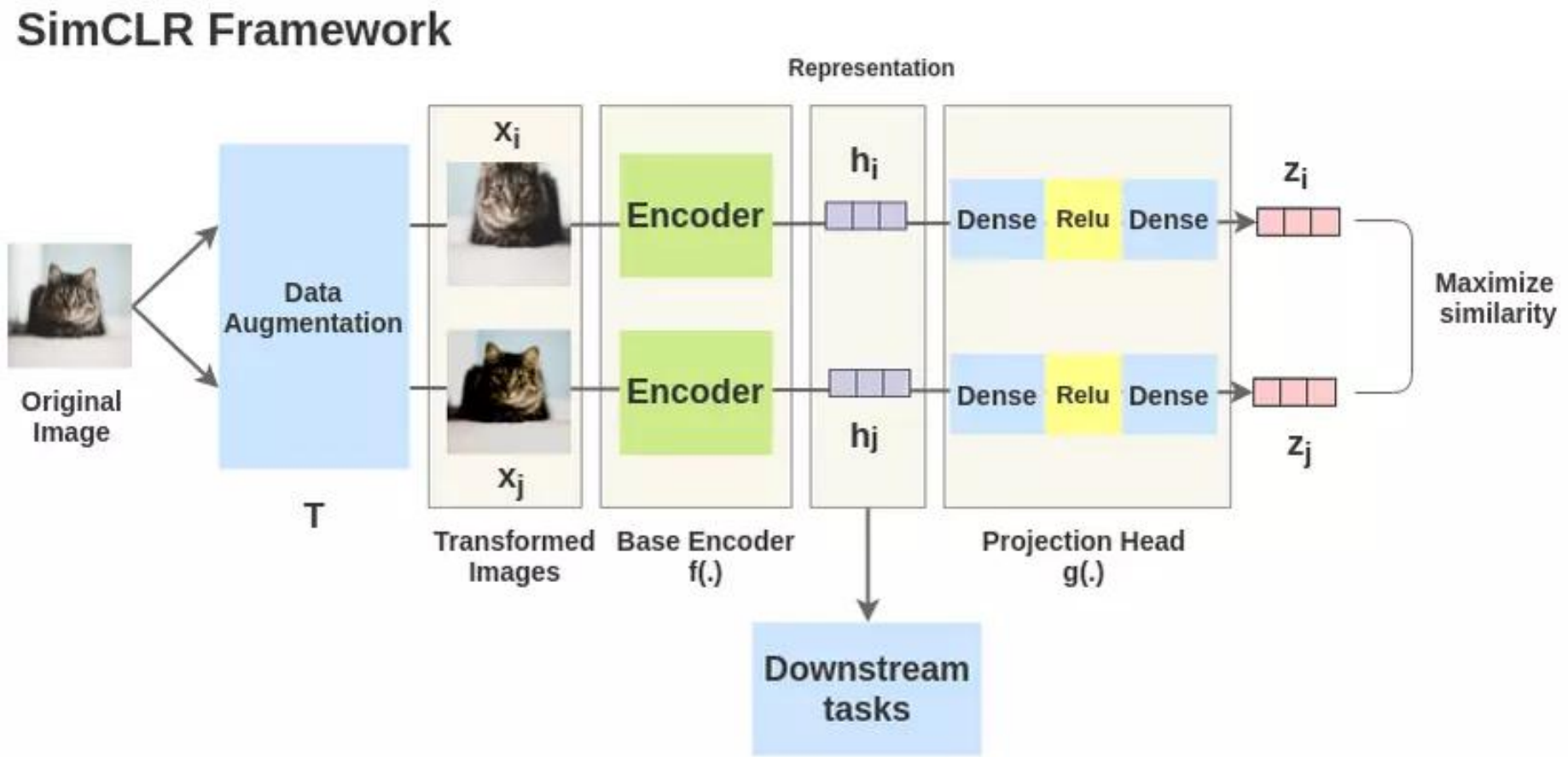
Ting Chen [1] Simon Kornblith [1] Mohammad Norouzi [1] Geoffrey Hinton [1]

Abstract

This paper presents *SimCLR*: a simple framework for contrastive learning of visual representations. We simplify recently proposed contrastive self-supervised learning algorithms without requiring specialized architectures or a memory bank. In order to understand what enables the contrastive prediction tasks to learn useful representations, we systematically study the major components of our framework. We show that (1) composition of data augmentations plays a critical role in defining effective predictive tasks, (2) introducing a learnable nonlinear transformation between the repre-



SimCLR (2020)



https://arxiv.org/abs/2002.05709

SimCLR (2020)

A Simple Framework for Contrastive Learning of Visual Representations

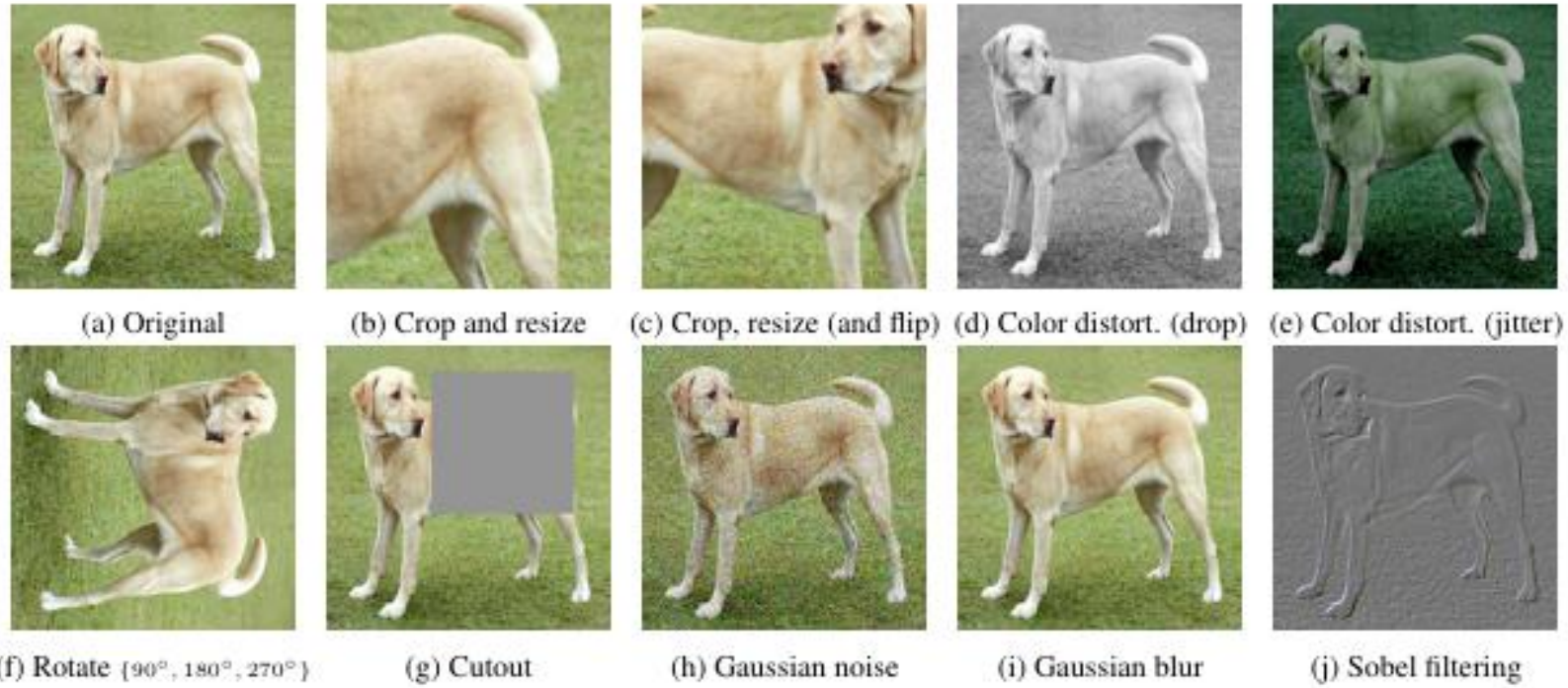


Figure 4. Illustrations of the studied data augmentation operators. Each augmentation can transform data stochastically with some internal parameters (e.g. rotation degree, noise level). Note that we *only* test these operators in ablation, the *augmentation policy used to train our models* only includes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

https://arxiv.org/abs/2002.05709

SimCLR (2020)

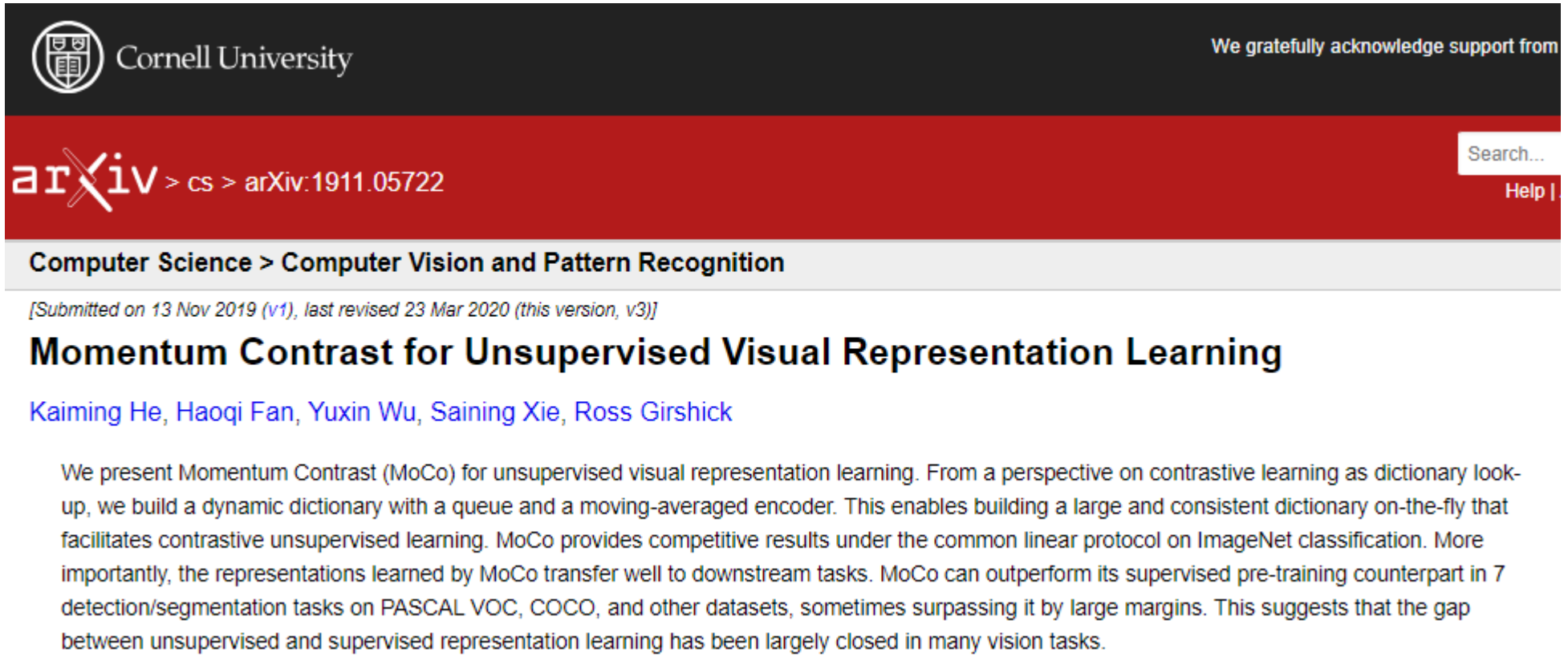
SimCLR là mô hình contrastive nổi tiếng:

- Dùng data augmentation mạnh (crop, color jitter, blur).
- Huấn luyện backbone CNN/ViT bằng contrastive loss.

→ Kết quả: chỉ với dữ liệu chưa nhãn, đạt hiệu năng gần supervised trên ImageNet.

MoCo (Momentum Contrast)

- https://arxiv.org/abs/1911.05722



Cornell University

We gratefully acknowledge support from

arXiv > cs > arXiv:1911.05722

Search...

Help |

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 13 Nov 2019 (v1), last revised 23 Mar 2020 (this version, v3)]

Momentum Contrast for Unsupervised Visual Representation Learning

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Ross Girshick

We present Momentum Contrast (MoCo) for unsupervised visual representation learning. From a perspective on contrastive learning as dictionary look-up, we build a dynamic dictionary with a queue and a moving-averaged encoder. This enables building a large and consistent dictionary on-the-fly that facilitates contrastive unsupervised learning. MoCo provides competitive results under the common linear protocol on ImageNet classification. More importantly, the representations learned by MoCo transfer well to downstream tasks. MoCo can outperform its supervised pre-training counterpart in 7 detection/segmentation tasks on PASCAL VOC, COCO, and other datasets, sometimes surpassing it by large margins. This suggests that the gap between unsupervised and supervised representation learning has been largely closed in many vision tasks.

MoCo (Momentum Contrast)

MoCo (2019–2020): cải tiến SimCLR bằng cách dùng queue lưu negative samples.

- Tránh phải dùng batch cực lớn.
- Sử dụng momentum encoder để duy trì ổn định.
- MoCo trở thành chuẩn mực trong SSL.

- Bài báo: https://arxiv.org/abs/1911.05722
- https://github.com/facebookresearch/moco
- Video: https://www.youtube.com/watch?v=LvHwBQF14zs

BYOL (Bootstrap Your Own Latent)

- https://arxiv.org/abs/2006.07733

Cornell University

We gratefully acknowledge support from the

arXiv > cs > arXiv:2006.07733

Search...

Help | Ad

Computer Science > Machine Learning

[Submitted on 13 Jun 2020 (v1), last revised 10 Sep 2020 (this version, v3)]

Bootstrap your own latent: A new approach to self-supervised Learning

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, Michal Valko

We introduce Bootstrap Your Own Latent (BYOL), a new approach to self-supervised image representation learning. BYOL relies on two neural networks, referred to as online and target networks, that interact and learn from each other. From an augmented view of an image, we train the online network to predict the target network representation of the same image under a different augmented view. At the same time, we update the target network with a slow-moving average of the online network. While state-of-the art methods rely on negative pairs, BYOL achieves a new state of the art without them. BYOL reaches 74.3% top-1 classification accuracy on ImageNet using a linear evaluation with a ResNet-50 architecture and 79.6% with a larger ResNet. We show that BYOL performs on par or better than the current state of the art on both transfer and semi-supervised benchmarks. Our implementation and pretrained models are given on GitHub.

Thực hi

BYOL (Bootstrap Your Own Latent)

- YOL: không cần negative samples.
- Dùng hai network: online và target.
- Online học để dự đoán representation của target.
- Kết quả: representation rất tốt, chứng minh **negative pair không bắt buộc.**

- https://github.com/google-deepmind/deepmind-research/tree/master/byol
- https://www.youtube.com/watch?v=YPfUiOMYOEE

DINO (2021)

DINO (2021)

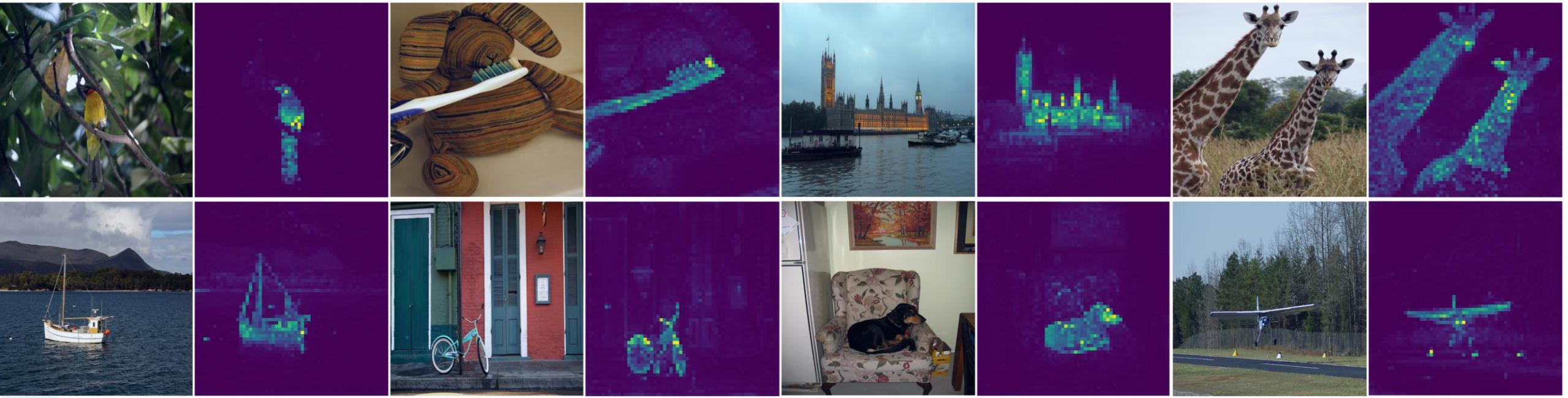
DINO sử dụng ViT làm backbone.

- Không cần nhãn.
- Representation học được có khả năng tự phân cụm theo object.

- https://github.com/facebookresearch/dino

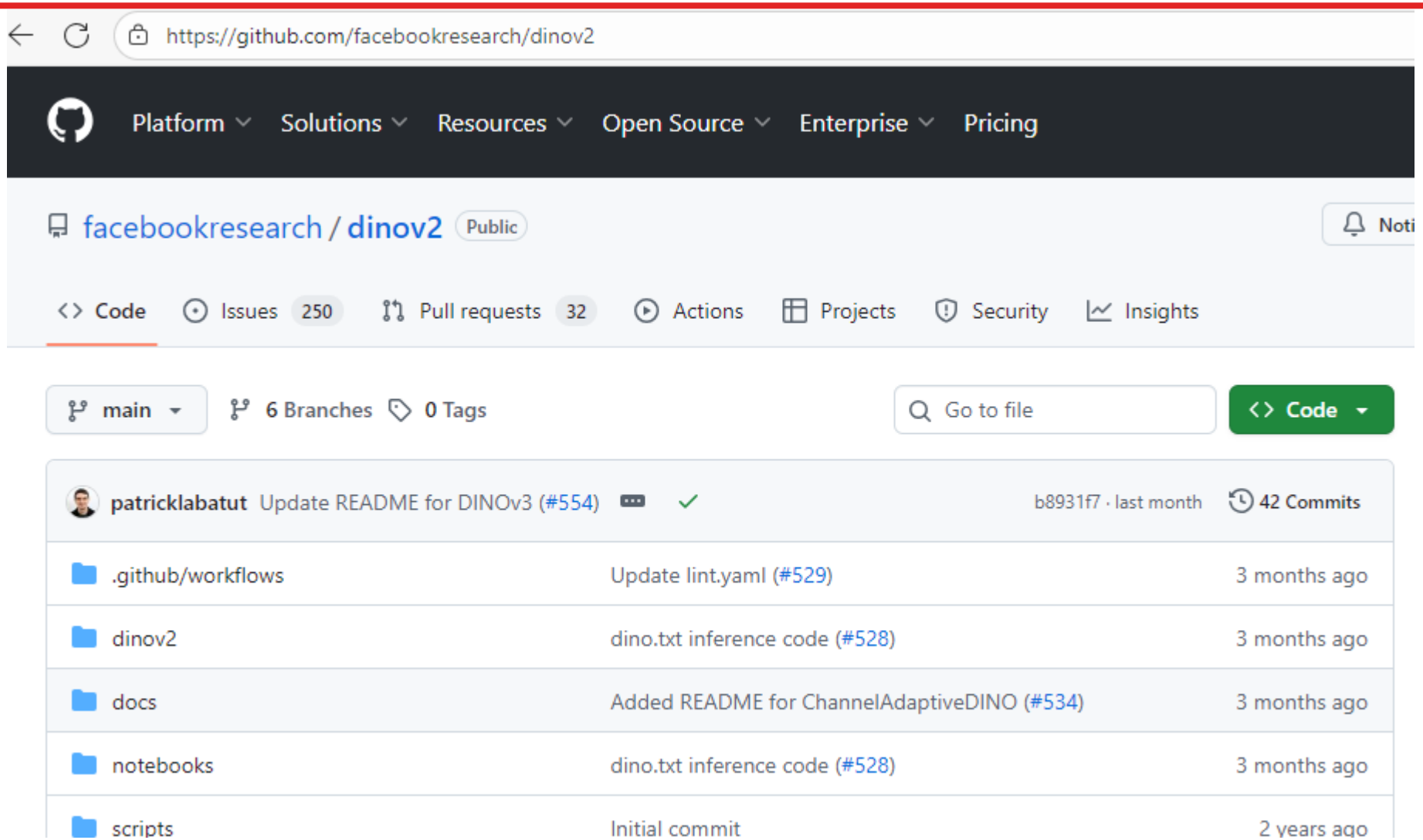
DINO (2021)

- Khi visualization attention của DINO cho thấy mô hình tự ‘highlight’ object chính trong ảnh



https://github.com/facebookresearch/dino

DINOv2


https://github.com/facebookresearch/dinov2
Platform
Solutions
Resources
Open Source
Enterprise
Pricing
facebookresearch / dinov2
Public
Noti
Code
Issues 250
Pull requests 32
Actions
Projects
Security
Insights
main
6 Branches
0 Tags
Go to file
Code
patricklabatut Update README for DINOv3 (#554)
b8931f7 · last month
42 Commits
.github/workflows
Update lint.yaml (#529)
3 months ago
dinov2
dino.txt inference code (#528)
3 months ago
docs
Added README for ChannelAdaptiveDINO (#534)
3 months ago
notebooks
dino.txt inference code (#528)
3 months ago
scripts
Initial commit
2 years ago

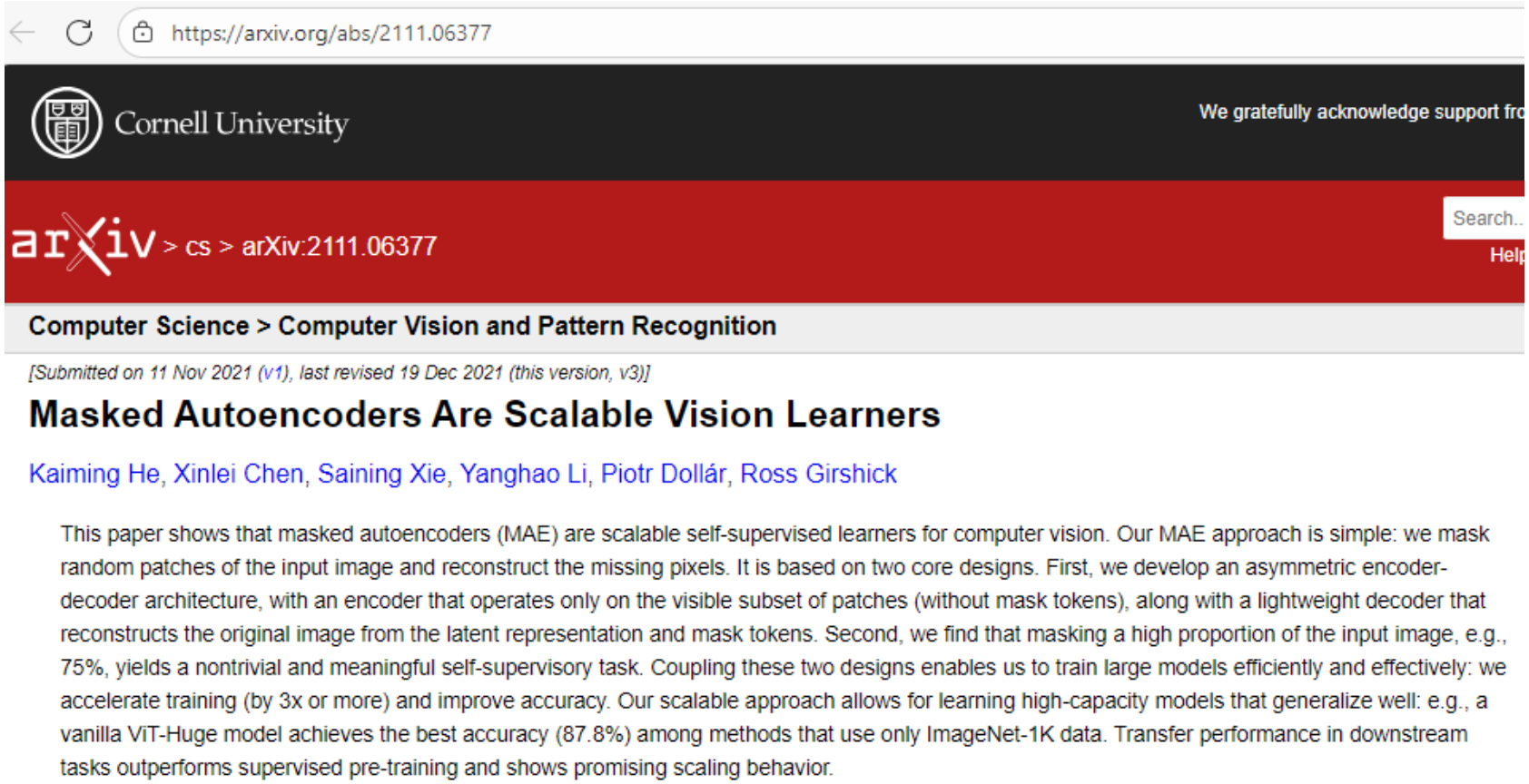
DINOv2

https://github.com/facebookresearch/dinov2

https://ai.meta.com/blog/dino-paws-computer-vision-with-self-supervised-transformers-and-10x-more-efficient-training/

MAE (Masked Autoencoder)

- https://arxiv.org/abs/2111.06377

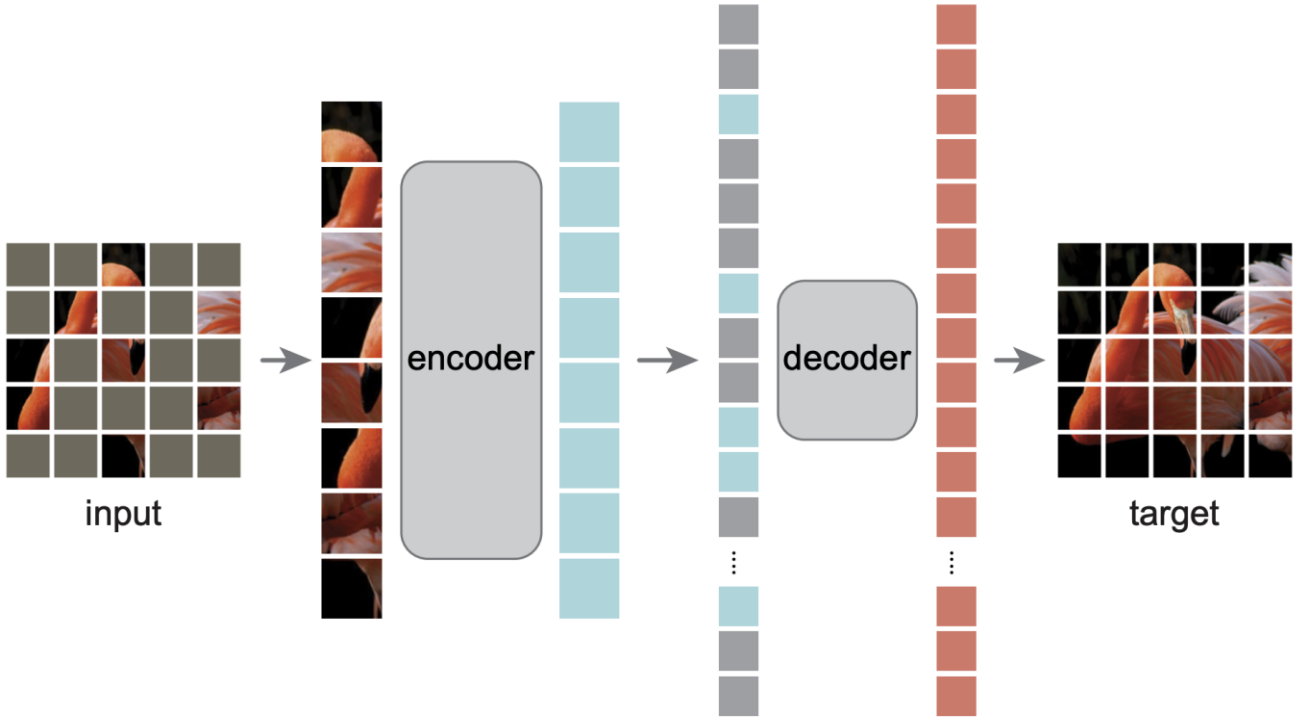


MAE (Masked Autoencoder)

- MAE là phương pháp nổi bật gần đây.
- Che 75% patch của ảnh.
- Encoder chỉ xử lý phần còn lại.
- Decoder tái dựng lại ảnh.

→ Kết quả: học representation hiệu quả, giảm chi phí huấn luyện

MAE (Masked Autoencoder)



- https://github.com/facebookresearch/mae
- https://colab.research.google.com/github/facebookresearch/mae/blob/main/demo/mae_visualize.ipynb
- https://www.geeksforgeeks.org/artificial-intelligence/masked-autoencoders-in-deep-learning/

Các yếu tố quan trọng trong contrastive learning

- **Data augmentation:** càng đa dạng, representation càng mạnh.
- **Batch size/negative samples**: nhiều negative → tốt hơn.
- **Architecture:** CNN, ViT, hybrid.

Ứng dụng contrastive learning

- Pretraining cho classification, detection, segmentation.
- Video understanding: frame contrastive learning.
- Multimodal: CLIP là một dạng contrastive giữa ảnh và text.

CLIP (OpenAI 2021)

- https://openai.com/index/clip/

CLIP (OpenAI 2021)

- https://openai.com/index/clip/
- https://arxiv.org/abs/2103.00020

- Positive = (ảnh, caption đúng).
- Negative = (ảnh, caption sai).
- CLIP map ảnh và văn bản vào cùng embedding space.

CLIP (OpenAI 2021)

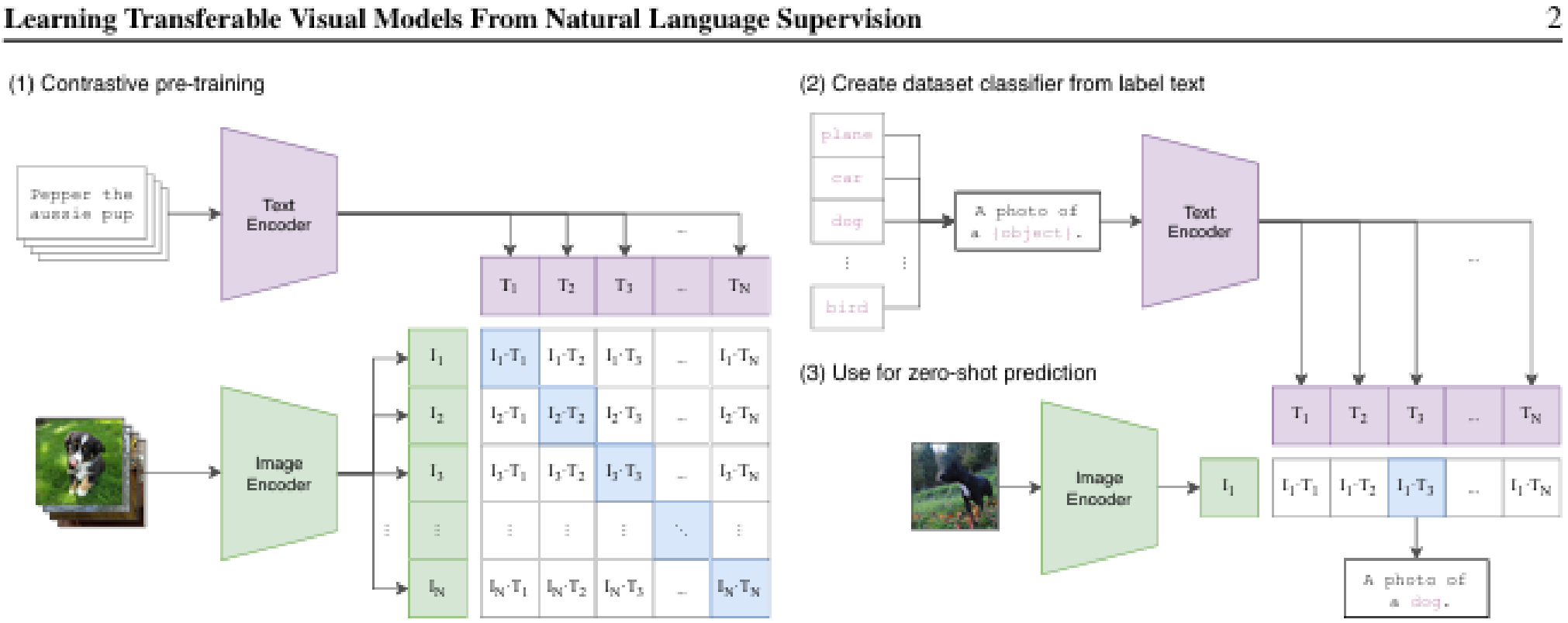


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

Kết quả CLIP

- CLIP chứng minh: chỉ cần huấn luyện contrastive trên dữ liệu Internet, mô hình có thể nhận diện hầu hết đối tượng mà không cần nhãn cụ thể.
- Hiệu quả với các bài toán zero-shot

→ Đây là bước đệm cho foundation models sau này.

Phần C – Case Studies từ CVPR

SSL trong video (CVPR 2024)

- Paper CVPR 2024: sử dụng self-supervised để pretrain video backbone.
- Ý tưởng: contrastive giữa frame video và motion clip.
- Kết quả: giảm 80% nhãn nhưng vẫn đạt SOTA action recognition.

DINOv2 (Meta AI 2023, CVPR 2024)

- DINOv2 là phiên bản nâng cấp:
  - ViT backbone lớn.
  - SSL trên dataset hàng tỷ ảnh.
  - Representation rất mạnh, trở thành nền tảng cho nhiều task.

MAE trong video (VideoMAE)

- VideoMAE: áp dụng masked autoencoder vào video.
- Che 90% patch spatio-temporal, yêu cầu mô hình tái dựng.
- Hiệu quả: representation video cực kỳ tốt, benchmark trên Kinetics-400

CLIP & Multimodal (CVPR 2023/24)

- Rất nhiều paper CVPR khai thác CLIP:
  - Image captioning.
  - Cross-modal retrieval.
  - Video–text alignment.
- Đây là minh chứng contrastive learning mở rộng sang multimodal.

Segment Anything + SSL

- Có paper kết hợp SAM với SSL: dùng SSL để refine mask, hoặc để học representation mạnh hơn từ segmentation.
- Đây là cách foundation models liên kết với SSL

Visual Prompting (CVPR 2024)

- Xu hướng mới: thay vì fine-tuning, chỉ dùng prompt.
- SSL representation mạnh đến mức chỉ cần một prompt đơn giản là đủ để áp dụng cho các bài toán mới

So sánh SSL và supervised

- Supervised: cần nhãn nhưng chính xác.
- SSL: không cần nhãn, nhưng cần dữ liệu lớn và training phức tạp

Kết nối tới foundation models

- Tất cả foundation models – CLIP, SAM, DINOv2 – đều dựa trên SSL.

→ **Đây là mảnh ghép nền tảng để xây dựng AI thị giác hiện đại**

Phần D – Ứng dụng & Demo

Ứng dụng trong y tế

- SSL hữu ích trong y tế vì dữ liệu nhãn rất hiếm.
- Ví dụ: MRI, CT scan.
- Pretrain bằng SSL → fine-tune cho segmentation khối u.
- Giảm 70% công sức gán nhãn

Ứng dụng trong video surveillance

- Video giám sát thường không có nhãn.
- SSL cho phép mô hình học representation người đi bộ, xe cộ mà không cần annotate.

- Ứng dụng: anomaly detection.

Ứng dụng trong thương mại điện tử

- Tìm kiếm sản phẩm bằng ảnh.
- SSL học representation tốt cho retrieval.
- Ví dụ: chụp áo thun → hệ thống gợi ý sản phẩm tương tự

Ứng dụng trong robotics

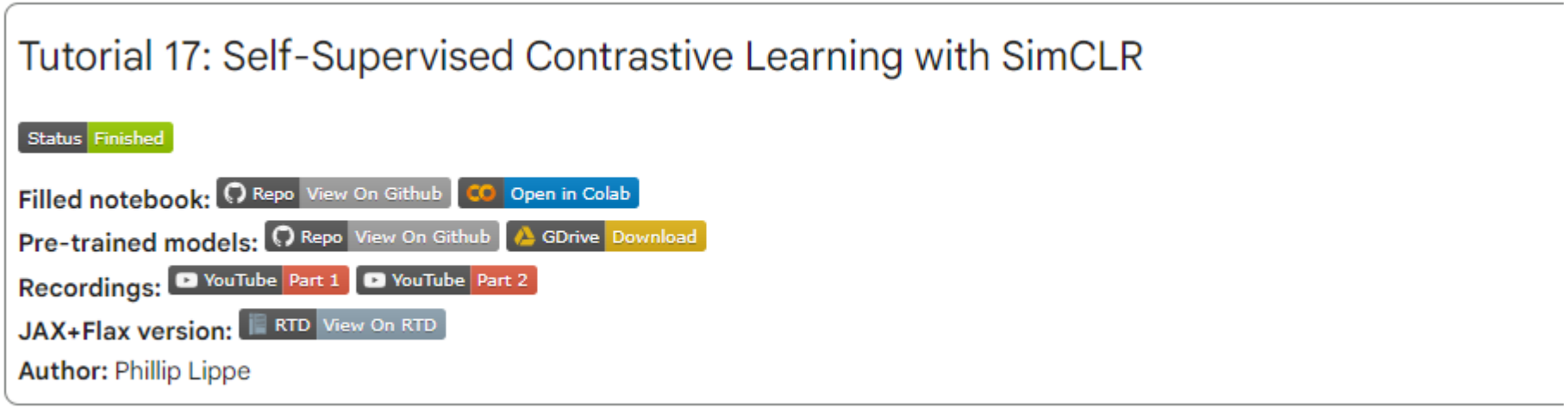
- Robot quan sát môi trường nhưng không có nhãn.
- SSL giúp robot học representation để định hướng, nắm bắt vật thể.
- Đây là bước quan trọng cho embodied AI

Ứng dụng trong multimodal AI

- Kết hợp ảnh + text + audio bằng contrastive learning.
- Ví dụ: video có tiếng chim → hệ thống tự nhận ra loài chim.

Demo: SimCLR toy example

https://colab.research.google.com/github/phlippe/uvadlc_notebooks/blob/master/docs/tutorial_notebooks/tutorial17/SimCLR.ipynb



Demo: CLIP zero-shot

- Nhập text prompt ‘a photo of a cat’, mô hình nhận diện được ảnh mèo mà không cần huấn luyện riêng.

- → **CLIP for Zero Shot Image Classification**

- medium.com

So sánh kết quả SSL vs supervised

- Với dữ liệu nhỏ: supervised mạnh hơn.
- Với dữ liệu lớn chưa nhãn: SSL vượt trội.

Phần E – Tổng kết & Bài tập

Tổng kết kiến thức

- Self-supervised learning là gì, pretext task ra sao.
- Contrastive learning: SimCLR, MoCo, BYOL, DINO, MAE.
- Ứng dụng: y tế, video, multimodal.
- Foundation models đều dựa trên SSL.

Điểm quan trọng

- SSL = học representation từ dữ liệu chưa nhãn.
- Contrastive learning = kéo gần positive, đẩy xa negative.
- CLIP = ví dụ thành công nhất của contrastive multimodal.

Bài tập cá nhân

1. Tham khảo và thực hiện lại theo hướng dẫn tại đường link: https://www.geeksforgeeks.org/machine-learning/self-supervised-learning-ssl/
2. Áp dụng cho bài toán phân loại các loài Hoa theo tập dataset được cung cấp.

Tài liệu tham khảo

- https://en.wikipedia.org/wiki/Self-supervised_learning
- https://www.geeksforgeeks.org/machine-learning/self-supervised-learning-ssl/
- https://learnopencv.com/contrastive-learning-simclr-and-byol-with-code-example/
- https://viblo.asia/p/tong-quan-ve-self-supervised-representation-learning-hoc-tu-giam-sat-Eb85oArkZ2G