NGHIÊN CỨU PHƯƠNG PHÁP MASKED AUTOENCODERS (MAE) TRONG HỌC TỰ GIÁM SÁT CHO THỊ GIÁC MÁY TÍNH.

Trịnh Tuấn Nam - 250101046

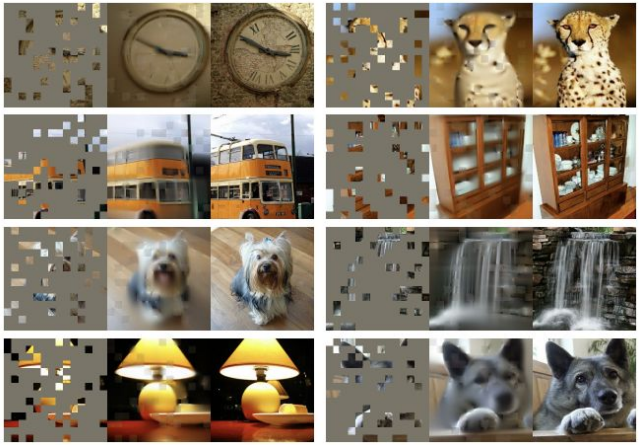
Tóm tắt

- Lớp: CS2205.SEP2025
- Link Github của nhóm: https://github.com/namt9/CS2205.CH201
- Link YouTube video: https://www.youtube.com/watch?v=NLoJBAL9TvY
- Ảnh + Họ và Tên của các thành viên
- Tổng số slides không vượt quá 10

- Họ và Tên: Trịnh Tuấn Nam
- MSSV: 250101046

Giới thiệu

- **NLP:** Rất thành công với Masked Autoencoding (BERT, GPT)
- **Computer Vision:** Tụt hậu hơn
- **Nguyên nhân chính:**
  - Hình ảnh có độ dư thừa không gian
  - Che ít -> dễ đoán -> không học được ngữ nghĩa

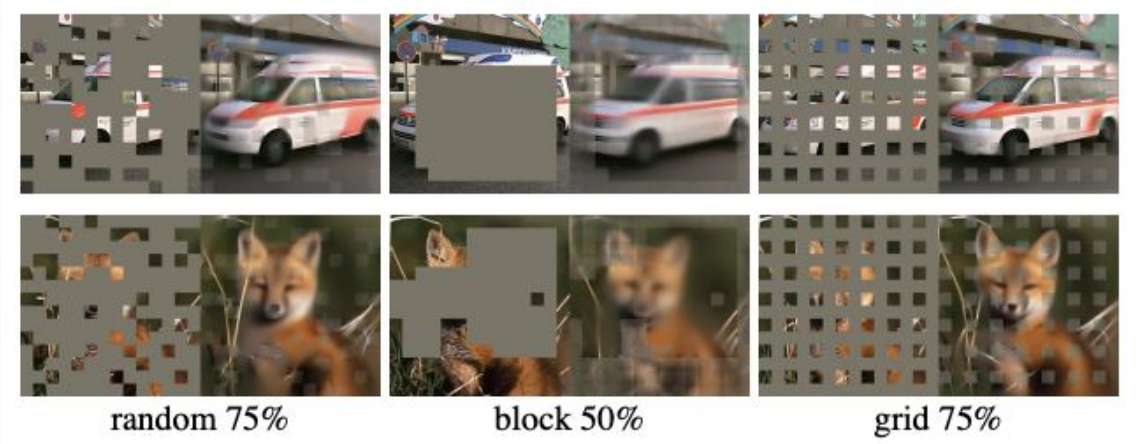


Mục tiêu

- **Scalable (Mở rộng):** Huấn luyện hiệu quả các mô hình lớn
- **High-Performance (Hiệu năng cao):** Đạt độ chính xác SOTA trên ImageNet-1K
- **Efficient (Hiệu quả):** Tăng tốc độ huấn luyện gấp 3 lần

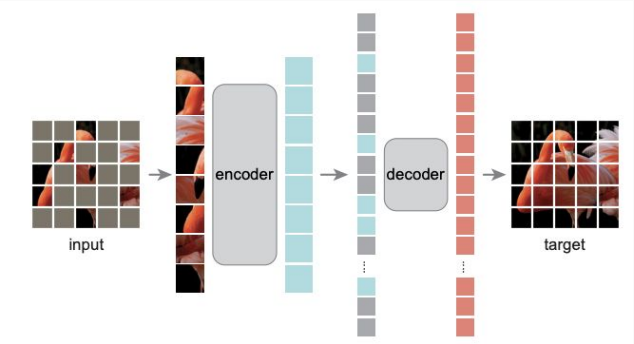
Phương pháp - Chiến lược Masking

- Random Sampling: Lấy mẫu ngẫu nhiên theo phân phối
- Tỷ lệ che 75%:
  - Loại bỏ dư thừa
  - Buộc mô hình học biểu diễn toàn diện



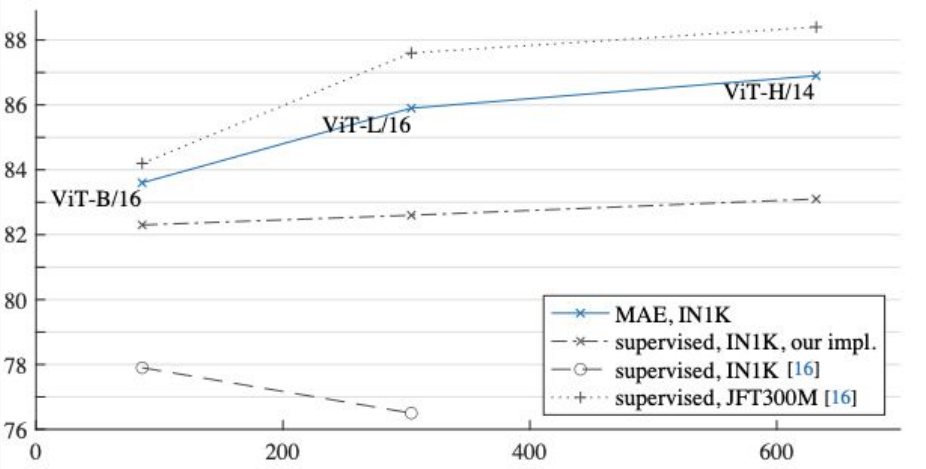
Phương pháp - Kiến trúc Bất đối xứng

- **Encoder (ViT):**
  - Chỉ xử lý **Visible Patches** (~25%).
  - Không dùng Mask tokens -> **Tiết kiệm tính toán.**
- **Decoder (Lightweight):**
  - Tái tạo điểm ảnh (Pixels) từ Latent + Mask tokens
  - Nhẹ hơn Encoder (< 10% computation)



Kết quả dự kiến

- **Tốc độ:** Tăng tốc **3x-4x**
- **Độ chính xác (ImageNet-1K):**
  - **ViT-Huge: 87.8%** (Tốt nhất dùng dữ liệu IN-1K)
  - Vượt trội so với Tiền huấn luyện có giám sát (Supervised)



Kết quả dự kiến

- **Object Detection (COCO):** Tốt hơn Supervised pre-training (**+4.0 AP** với ViT-L)
- **Segmentation (ADE20K):** Cải thiện đáng kể **(+3.7 mIoU)**
- **Chất lượng tái tạo:** Khôi phục cấu trúc hợp lý dù mất 75% thông tin

Tài liệu tham khảo

[MAE] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, Ross Girshick: **Masked Autoencoders Are Scalable Vision Learners.** *CVPR 2022* (arXiv:2111.06377).

[ViT] Alexey Dosovitskiy et al.: **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.** *ICLR 2021*.

[BERT] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.** *NAACL-HLT 2019*.

[GPT-3] Tom B. Brown et al.: **Language Models are Few-Shot Learners.** *NeurIPS 2020*.