THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): *https://youtu.be/NLoJBAL9TvY*
- Link slides (dạng .pdf đặt trên Github của nhóm): *https://github.com/namt9/CS2205.CH201/blob/master/Nam-Trinh-Tuan-CS2205.SEP2025.FinalReport.Slide.pdf*
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*
- *Lớp Cao học, mỗi nhóm một thành viên*

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

- Lớp: CS2205.RM
- Tự đánh giá (điểm tổng kết môn): 7.5/10
- Số buổi vắng: 1
- Số câu hỏi QT cá nhân: 3
- Số câu hỏi QT của cả nhóm: 0
- Link Github: *https://github.com/namt9/CS2205.CH201*

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

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

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

MASKED AUTOENCODERS ARE SCALABLE VISION LEARNERS

TÓM TẮT *(Tối đa 400 từ)*

Trong bối cảnh các mô hình học sâu (Deep Learning) ngày càng đòi hỏi lượng lớn dữ liệu, học tự giám sát (self-supervised learning) đã đạt được những thành tựu lớn trong xử lý ngôn ngữ tự nhiên nhưng vẫn gặp thách thức khi áp dụng sang thị giác máy tính. Đề tài này nghiên cứu phương pháp **Masked Autoencoders (MAE)**, một hướng tiếp cận có khả năng mở rộng cao cho các mô hình thị giác. Phương pháp luận của nghiên cứu dựa trên hai thiết kế cốt lõi: thứ nhất là kiến trúc encoder-decoder bất đối xứng, nơi encoder chỉ xử lý các phần hình ảnh hiển thị (không bị che) và decoder nhẹ tái tạo lại ảnh gốc; thứ hai là chiến lược che (masking) tỷ lệ lớn lên tới 75% diện tích ảnh đầu vào. Kết quả thực nghiệm cho thấy sự kết hợp này không chỉ tăng tốc độ huấn luyện lên gấp 3 lần mà còn cải thiện đáng kể độ chính xác và khả năng tổng quát hóa của mô hình. Cụ thể, mô hình ViT-Huge sử dụng MAE đạt độ chính xác 87.8% trên tập ImageNet-1K, vượt qua các phương pháp tiền huấn luyện có giám sát truyền thống.

GIỚI THIỆU *(Tối đa 1 trang A4)*

1. Bối cảnh và Đặt vấn đề Sự bùng nổ của các kiến trúc học sâu (Deep Learning) gần đây đã cho phép các mô hình đạt dung lượng khổng lồ, tuy nhiên chúng thường đòi hỏi hàng trăm triệu dữ liệu gán nhãn – một nguồn tài nguyên tốn kém và khó tiếp cận. Trong khi lĩnh vực Xử lý ngôn ngữ tự nhiên (NLP) đã giải quyết vấn đề này hiệu quả nhờ phương pháp học tự giám sát (self-supervised learning) như BERT hay GPT , thì thị giác máy tính vẫn chưa đạt được thành công tương xứng.

Nguyên nhân cốt lõi nằm ở sự khác biệt về **mật độ thông tin**: ngôn ngữ mang tính ngữ nghĩa cao, trong khi hình ảnh có độ dư thừa không gian (spatial redundancy) lớn. Một mảng ảnh bị thiếu có thể dễ dàng được suy luận từ các mảng lân cận mà không cần hiểu biết sâu về đối tượng, khiến các phương pháp masking truyền thống kém hiệu quả trên dữ liệu hình ảnh.

2. Phương pháp tiếp cận Để khắc phục hạn chế trên, nghiên cứu này đề xuất **Masked Autoencoders (MAE)** – một phương pháp học tự giám sát đơn giản nhưng có khả năng mở rộng (scalable) cho thị giác máy tính. Giải pháp dựa trên hai thiết kế chính:

- **Chiến lược che tỷ lệ cao (High masking ratio):** Che đi tới 75% diện tích ảnh đầu vào để loại bỏ sự dư thừa, buộc mô hình phải học các biểu diễn toàn diện thay vì chỉ ngoại suy đơn giản.
- **Kiến trúc bất đối xứng (Asymmetric architecture):** Sử dụng encoder chỉ hoạt động trên tập hợp con các mảng hiển thị (không có mask token) và một decoder hạng nhẹ để tái tạo ảnh gốc.

3. Đóng góp của đề tài Nghiên cứu chứng minh tính hiệu quả vượt trội của MAE về cả tốc độ lẫn hiệu suất. Nhờ encoder chỉ cần xử lý khoảng 25% dữ liệu đầu vào, thời gian huấn luyện giảm xuống **3 lần hoặc hơn** đồng thời tiết kiệm bộ nhớ. Về độ chính xác, mô hình ViT-Huge sử dụng MAE đạt **87.8%** trên ImageNet-1K, vượt qua các

phương pháp tiền huấn luyện có giám sát và thể hiện khả năng chuyển giao (transfer learning) xuất sắc. Kết quả này mở ra hướng đi mới giúp các mô hình thị giác mở rộng quy mô tương tự như thành công đã đạt được trong NLP.

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

1. Phát triển phương pháp học tự giám sát (Self-supervised learning) có khả năng mở rộng cho thị giác máy tính Mục tiêu đầu tiên là chứng minh rằng cơ chế "masked autoencoding" (tự mã hóa có che) – vốn đã rất thành công trong xử lý ngôn ngữ tự nhiên (NLP) như BERT – cũng có thể trở thành một giải pháp học tự giám sát hiệu quả và có khả năng mở rộng (scalable) cho thị giác máy tính. Nghiên cứu nhằm thu hẹp khoảng cách giữa sự phát triển của các phương pháp tự giám sát trong NLP và thị giác máy tính bằng cách giải quyết các khác biệt về kiến trúc và mật độ thông tin giữa hai lĩnh vực này.

2. Tối ưu hóa hiệu suất huấn luyện thông qua kiến trúc bất đối xứng và tỷ lệ che cao Mục tiêu thứ hai là thiết kế một kiến trúc **encoder-decoder bất đối xứng** kết hợp với chiến lược che (masking) tỷ lệ lớn (khoảng 75%). Cách tiếp cận này nhằm mục đích giảm thiểu đáng kể chi phí tính toán và bộ nhớ bằng cách chỉ cho encoder xử lý các phần hình ảnh hiển thị (khoảng 25%), từ đó tăng tốc độ huấn luyện lên gấp 3 lần hoặc hơn so với các phương pháp thông thường mà vẫn đảm bảo hiệu quả.

3. Đạt độ chính xác vượt trội và khả năng tổng quát hóa trên các mô hình dung lượng lớn Mục tiêu cuối cùng là chứng minh phương pháp MAE cho phép huấn luyện các mô hình có dung lượng rất lớn (high-capacity models) như ViT-Large hoặc ViT-Huge đạt khả năng tổng quát hóa tốt. Cụ thể, nghiên cứu hướng tới việc đạt độ chính xác cao trên các tập chuẩn (benchmark) như ImageNet-1K và vượt qua các phương pháp tiền huấn luyện có giám sát (supervised pre-training) trong các tác vụ chuyển giao (transfer learning) như phát hiện đối tượng hay phân đoạn ảnh.

NỘI DUNG VÀ PHƯƠNG PHÁP

1. Tổng quan phương pháp tiếp cận

Nghiên cứu đề xuất sử dụng kiến trúc **Masked Autoencoder (MAE)** – một mô hình học tự giám sát (self-supervised learning) có khả năng mở rộng cho thị giác máy tính. Cách tiếp cận này dựa trên ý tưởng đơn giản: che đi các phần ngẫu nhiên của ảnh đầu vào và tái tạo lại các điểm ảnh (pixels) bị thiếu.

Khác với các phương pháp Autoencoder truyền thống, MAE sử dụng thiết kế **bất đối xứng (asymmetric)** giữa encoder và decoder để tối ưu hóa hiệu quả tính toán và khả năng biểu diễn của mô hình.

2. Thiết kế kiến trúc mô hình (Model Architecture)

Hệ thống được thiết kế gồm hai thành phần chính:

- **Encoder:**
    - Sử dụng kiến trúc Vision Transformer (ViT) tiêu chuẩn nhưng chỉ hoạt động trên các mảng ảnh hiển thị (visible patches).
    - Các token bị che (mask tokens) **không** được đưa vào encoder. Điều này giúp giảm đáng kể chi phí tính toán và bộ nhớ, cho phép huấn luyện các mô hình rất lớn (như ViT-Large/Huge) với chỉ một phần nhỏ dữ liệu đầu vào (khoảng 25%).
- **Decoder:**
    - Là một kiến trúc nhẹ (lightweight), nhỏ hơn và nông hơn so với encoder (ví dụ: chi phí tính toán mỗi token nhỏ hơn 10% so với encoder).
    - Đầu vào của decoder là tập hợp đầy đủ các token: bao gồm các token đã được mã hóa từ encoder và các mask token (đại diện cho các phần bị thiếu cần dự đoán).
    - Decoder chỉ được sử dụng trong giai đoạn tiền huấn luyện (pre-training)

để tái tạo ảnh và được loại bỏ khi thực hiện các tác vụ nhận diện sau này.

3. Quy trình thực hiện (Methodology)

Quy trình huấn luyện và hoạt động của MAE được thực hiện qua 4 bước chính:

- **Phân chia và Che ảnh (Masking):**
  - Chia ảnh đầu vào thành các mảng (patches) không chồng lấn.
  - Áp dụng chiến lược lấy mẫu ngẫu nhiên (random sampling) theo phân phối đều để chọn ra tập con các mảng sẽ bị che.
  - **Tỷ lệ che (Masking Ratio):** Sử dụng tỷ lệ che rất cao, lên tới **75%** diện tích ảnh. Tỷ lệ này giúp loại bỏ sự dư thừa thông tin không gian, buộc mô hình phải học các biểu diễn ngữ nghĩa toàn diện thay vì chỉ ngoại suy từ các chi tiết lân cận.
- **Mã hóa (Encoding):**
  - Chỉ các mảng hiển thị (không bị che) được nhúng (embedded) và đưa qua các khối Transformer của Encoder.
- **Tái tạo (Reconstruction):**
  - Encoder output được ghép nối với các mask token (có thêm positional embeddings) để khôi phục lại thứ tự không gian ban đầu.
  - Decoder xử lý toàn bộ chuỗi này để dự đoán giá trị pixel của các mảng bị che.
- **Hàm mất mát (Loss Function):**
  - Sử dụng Mean Squared Error (MSE) giữa ảnh tái tạo và ảnh gốc.
  - Loss chỉ được tính trên các mảng bị che (masked patches), tương tự như phương pháp BERT trong xử lý ngôn ngữ.
  - Sử dụng kỹ thuật chuẩn hóa pixel (pixel normalization) trên từng mảng để cải thiện chất lượng biểu diễn.

4. Kế hoạch thực nghiệm và Đánh giá

- **Dữ liệu:** Sử dụng tập dữ liệu ImageNet-1K để tiền huấn luyện mô hình.
- **Mô hình nền tảng:** ViT-Large và ViT-Huge.
- **Phương pháp đánh giá:**
    - **End-to-end Fine-tuning:** Tinh chỉnh toàn bộ mô hình trên các tác vụ có giám sát để đánh giá độ chính xác.
    - **Linear Probing:** Đánh giá chất lượng của biểu diễn (representation quality) bằng cách huấn luyện một bộ phân loại tuyến tính trên đặc trưng đã đóng băng.
    - **Transfer Learning:** Đánh giá khả năng chuyển giao tri thức sang các tác vụ khác như phát hiện đối tượng (trên bộ dữ liệu COCO) và phân đoạn ngữ nghĩa (trên bộ dữ liệu ADE20K).

KẾT QUẢ MONG ĐỢI

1. Tối ưu hóa hiệu suất huấn luyện

- **Tốc độ huấn luyện:** Dự kiến tăng tốc độ huấn luyện lên **3 lần hoặc hơn** so với các phương pháp tiêu chuẩn. Kết quả này đạt được nhờ thiết kế encoder chỉ xử lý tập con các mảng hiển thị (khoảng 25% dữ liệu ảnh), giúp giảm thiểu khối lượng tính toán.
- **Tiêu thụ bộ nhớ:** Giảm đáng kể dung lượng bộ nhớ yêu cầu trong quá trình tiền huấn luyện, cho phép mở rộng quy mô mô hình lên mức rất lớn hoặc tăng kích thước batch size.

2. Độ chính xác trên tập dữ liệu chuẩn (ImageNet-1K)

- **Khả năng mở rộng (Scalability):** Phương pháp dự kiến thể hiện khả năng mở rộng tốt; các mô hình càng lớn sẽ cho kết quả càng cao mà không bị bão hòa

nhanh chóng như phương pháp có giám sát.

- **Kết quả cụ thể:**
    - Mô hình **ViT-Huge** được huấn luyện bằng MAE dự kiến đạt độ chính xác top-1 là **87.8%** khi fine-tune trên ImageNet-1K.
    - Kết quả này vượt qua tất cả các phương pháp trước đó chỉ sử dụng dữ liệu ImageNet-1K (bao gồm cả các phương pháp học có giám sát và tự giám sát khác).

3. Hiệu quả trong các tác vụ chuyển giao (Transfer Learning) Nghiên cứu dự kiến chứng minh rằng các đặc trưng (features) học được từ MAE có khả năng tổng quát hóa tốt sang các bài toán thị giác khác:

- **Phát hiện đối tượng (Object Detection):** Trên tập dữ liệu COCO, mô hình sử dụng MAE pre-training dự kiến vượt trội hơn so với tiền huấn luyện có giám sát (supervised pre-training). Ví dụ, với backbone ViT-Large, MAE có thể cao hơn tới **4.0 điểm AP**.
- **Phân đoạn ngữ nghĩa (Semantic Segmentation):** Trên tập dữ liệu ADE20K, phương pháp này dự kiến cải thiện đáng kể kết quả, ví dụ tăng **3.7 điểm** cho mô hình ViT-Large so với các đối thủ cạnh tranh.

4. Chất lượng tái tạo hình ảnh (Qualitative Results)

- Mặc dù tỷ lệ che lên tới 75% (chỉ để lại 25% thông tin hình ảnh), mô hình dự kiến vẫn có khả năng tái tạo lại các cấu trúc toàn diện và hợp lý về mặt ngữ nghĩa của các đối tượng và khung cảnh, thay vì chỉ đơn thuần là làm mờ hay ngoại suy các đường nét cơ bản.
- Điều này minh chứng rằng mô hình đã học được các khái niệm thị giác (visual concepts) cấp cao và hiểu được bối cảnh tổng thể của bức ảnh.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," *arXiv preprint arXiv:2111.06377*, 2021.

[2] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[4] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and Composing Robust Features with Denoising Autoencoders," in *International Conference on Machine Learning (ICML)*, 2008.

[5] H. Bao, L. Dong, and F. Wei, "BEiT: BERT Pre-Training of Image Transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[6] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning" (MoCo), in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[7] M. Caron *et al.*, "Emerging Properties in Self-Supervised Vision Transformers" (DINO), in *International Conference on Computer Vision (ICCV)*, 2021.