THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): *https://www.youtube.com/watch?v=MLMj3KqrkpE*
- Link slides (dạng .pdf đặt trên Github của nhóm): *https://github.com/LamThanhNgan/CS2205.CH201/blob/master/Ng%C3%A2n%20L%C3%A2m%20Thanh%20-%20CS2205.SEP2025.DeCuong.FinalReport.Template.Slide.pdf*

- Họ và Tên: Lâm Thanh Ngân
- MSSV: 250202015

- Lớp: CS2205.CH201
- Tự đánh giá (điểm tổng kết môn): 9.5/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 3
- Số câu hỏi QT của cả nhóm: 0
- Link Github: https://github.com/LamThanhNgan/CS2205.CH201

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

HỌC ĐA PHƯƠNG THỨC VIDEO-NGÔN NGỮ BẰNG PHƯƠNG PHÁP LẤY MẪU THƯA: TIẾP CẬN END-TO-END TỪ PIXEL VIDEO VÀ TOKEN VĂN BẢN

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

VIDEO-AND-LANGUAGE LEARNING VIA SPARSE SAMPLING: AN END-TO-END APPROACH FROM RAW VIDEO PIXELS AND TEXT TOKENS

TÓM TẮT *(Tối đa 400 từ)*

Trong những năm gần đây, các tác vụ học đa phương thức kết hợp video và ngôn ngữ như truy vấn video bằng văn bản (text-to-video retrieval), trả lời câu hỏi dựa trên video (video question answering), và mô tả video tự động (video captioning) đã thu hút sự quan tâm lớn từ cộng đồng nghiên cứu trí tuệ nhân tạo. Tuy nhiên, các phương pháp hiện tại chủ yếu dựa vào việc trích xuất đặc trưng video dày đặc (dense features) từ các mô hình thị giác được huấn luyện trước một cách offline, dẫn đến hai hạn chế chính: (i) sự ngắt kết nối giữa pre-training và downstream tasks do các đặc trưng được tối ưu cho mục tiêu khác; và (ii) chi phí tính toán và bộ nhớ lớn khi xử lý đặc trưng từ toàn bộ khung hình video.

Để khắc phục những thách thức trên, nghiên cứu này tìm hiểu và đánh giá ClipBERT – một khung làm việc học đa phương thức video-ngôn ngữ theo hướng end-to-end với chiến lược lấy mẫu thưa (sparse sampling). Thay vì xử lý toàn bộ video, phương pháp chỉ lấy mẫu ngẫu nhiên một hoặc vài đoạn clip ngắn tại mỗi bước huấn luyện, dựa trên giả thuyết rằng các clip thưa đã chứa đủ thông tin ngữ nghĩa quan trọng. Kiến trúc bao gồm ba thành phần: (i) Bộ mã hóa thị giác 2D CNN (ResNet-50) trích xuất đặc trưng không gian-thời gian; (ii) Bộ mã hóa ngôn ngữ từ

token văn bản; và (iii) Transformer đa phương thức mô hình hóa quan hệ video-văn bản. Toàn bộ mô hình được huấn luyện end-to-end với chiến lược "sparse-training then dense-inference" để cân bằng hiệu quả và độ chính xác.

Nghiên cứu đánh giá phương pháp trên nhiều benchmark chuẩn: MSR-VTT, DiDeMo, ActivityNet Captions cho tác vụ text-to-video retrieval, và TGIF-QA, MSRVTT-QA cho tác vụ video question answering. Kết quả mong đợi cho thấy phương pháp đạt hiệu suất cạnh tranh hoặc vượt trội so với các baseline sử dụng đặc trưng offline dày đặc, đồng thời giảm đáng kể chi phí bộ nhớ và thời gian tính toán. Các thí nghiệm ablation cũng được thực hiện để phân tích ảnh hưởng của số lượng clips, số frames, và chiến lược pre-training đến hiệu suất mô hình.

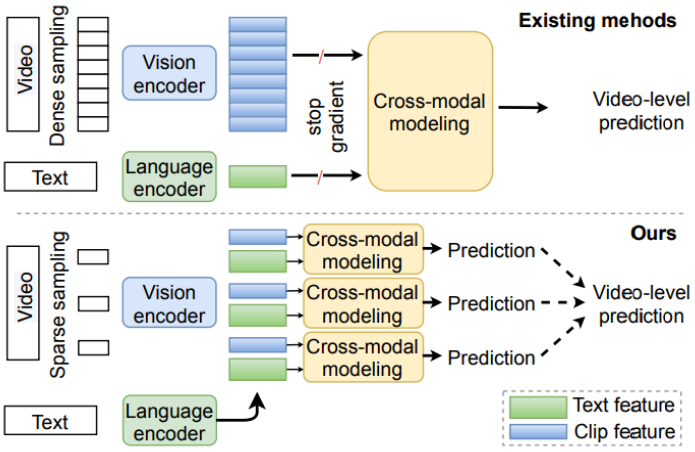
GIỚI THIỆU *(Tối đa 1 trang A4)*

Trong thời đại bùng nổ nội dung đa phương thức, khả năng hiểu và kết nối thông tin từ video và ngôn ngữ tự nhiên đã trở thành yêu cầu thiết yếu cho các hệ thống trí tuệ nhân tạo hiện đại. Các ứng dụng như tìm kiếm video thông minh, hệ thống hỏi đáp tự động dựa trên video, và phụ đề tự động cho người khiếm thính đang có nhu cầu ngày càng tăng cao khi lượng video trực tuyến phát triển theo cấp số nhân. Vì vậy, việc nghiên cứu các giải pháp công nghệ cho bài toán học đa phương thức video-ngôn ngữ có ý nghĩa xã hội và thực tiễn rõ rệt.

Các phương pháp học đa phương thức video-ngôn ngữ hiện tại chủ yếu dựa trên paradigm truyền thống: trích xuất đặc trưng video dày đặc (dense features) từ các mô hình thị giác đã được huấn luyện trước một cách offline (ví dụ: I3D, C3D cho action recognition), sau đó kết hợp với đặc trưng ngôn ngữ qua các mô-đun fusion để thực hiện các tác vụ cụ thể. Tuy nhiên, cách tiếp cận này tồn tại ba hạn chế nghiêm trọng: (i) Sự ngắt kết nối giữa pre-training và downstream tasks – các đặc trưng được tối ưu cho mục tiêu ban đầu (phân loại hành động) nên có thể bỏ lỡ thông tin quan trọng cho tác vụ video-ngôn ngữ; (ii) Chi phí tính toán và bộ nhớ khổng lồ khi trích xuất đặc trưng từ toàn bộ khung hình video (25-30 frames/giây), khiến việc huấn luyện

end-to-end trở nên không khả thi; và (iii) Khả năng tổng quát hóa kém khi chuyển đổi sang các miền video khác biệt.

Để khắc phục những hạn chế trên, nghiên cứu này tìm hiểu ClipBERT – một khung làm việc học đa phương thức theo hướng end-to-end với chiến lược lấy mẫu thưa (sparse sampling). Ý tưởng cốt lõi dựa trên quan sát "Less is More": chỉ cần lấy mẫu ngẫu nhiên một vài đoạn clip ngắn từ video đã chứa đủ thông tin thị giác và ngữ nghĩa quan trọng để mô hình học được biểu diễn hiệu quả. Thay vì xử lý toàn bộ video với hàng trăm khung hình, phương pháp chỉ lấy mẫu ngẫu nhiên 1-2 đoạn clip (mỗi clip chứa 2-16 khung hình) tại mỗi bước huấn luyện, sử dụng 2D CNN nhẹ (ResNet-50) kết hợp với Transformer đa phương thức. Chiến lược này giúp giảm đáng kể chi phí tính toán, cho phép huấn luyện end-to-end trực tiếp từ pixel video, đồng thời tận dụng được image-text pre-training để khởi tạo trọng số tốt hơn. Hình 1 minh họa sự khác biệt giữa kiến trúc đề xuất và các phương pháp truyền thống.



Hình 1. So sánh kiến trúc giữa phương pháp truyền thống (trên) sử dụng dense sampling với stop gradient và phương pháp ClipBERT (dưới) sử dụng sparse sampling cho phép huấn luyện end-to-end.

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

Nghiên cứu hướng tới ba mục tiêu cụ thể sau:

1. Tìm hiểu sâu về cơ chế lấy mẫu thưa (sparse sampling) cho video và kiến trúc end-to-end kết hợp 2D CNN với Transformer đa phương thức. Phân tích giả thuyết "less is more" và chiến lược "sparse-training then dense-inference".
2. Triển khai pipeline hoàn chỉnh với quy trình hai giai đoạn: pre-training trên dữ liệu ảnh-văn bản (COCO Captions, Visual Genome) và fine-tuning trên dữ liệu video-văn bản cho các tác vụ text-to-video retrieval và video question answering.
3. Đánh giá định lượng với các chỉ số Recall@K, Median Rank, và Accuracy trên các benchmark chuẩn (MSR-VTT, DiDeMo, ActivityNet Captions, TGIF-QA, MSRVTT-QA). Thực hiện ablation study phân tích ảnh hưởng của số clips, frames và pre-training.

NỘI DUNG VÀ PHƯƠNG PHÁP

1. Nội dung

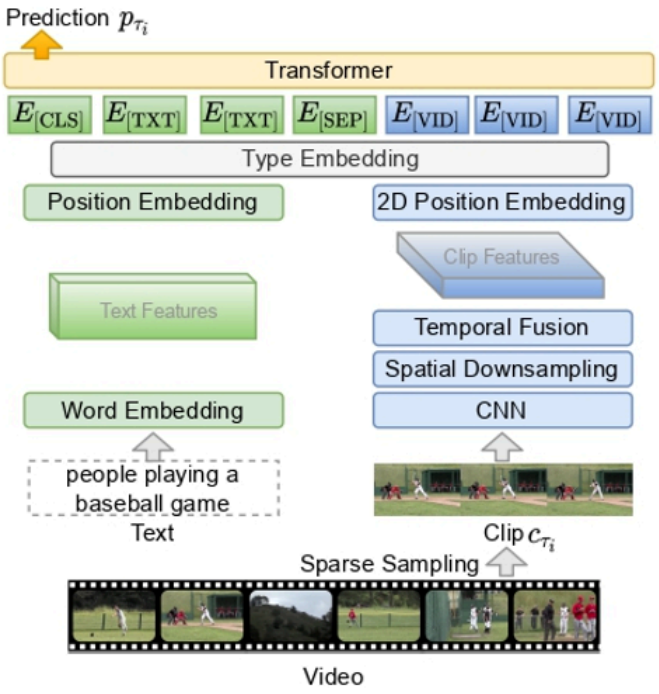
Nghiên cứu tập trung xây dựng khung làm việc học đa phương thức video-ngôn ngữ theo hướng end-to-end với chiến lược lấy mẫu thưa. Thay vì trích xuất đặc trưng dày đặc từ toàn bộ video như các phương pháp truyền thống, chúng tôi chỉ lấy mẫu ngẫu nhiên một hoặc vài đoạn clip ngắn tại mỗi bước huấn luyện. Giả thuyết nghiên cứu là các clip thưa này đã chứa đủ thông tin thị giác và ngữ nghĩa quan trọng từ video liên tục, do đó có thể huấn luyện mô hình hiệu quả mà không cần xử lý toàn bộ video dày đặc.

Kiến trúc mô hình bao gồm ba thành phần chính: (i) Bộ mã hóa thị giác 2D CNN (ResNet-50) trích xuất đặc trưng không gian-thời gian từ các frame được lấy mẫu; (ii) Bộ mã hóa ngôn ngữ học được từ token văn bản thô; và (iii) Transformer đa phương

thức 12 lớp mô hình hóa quan hệ giữa video và văn bản. Toàn bộ mô hình được huấn luyện end-to-end, cho phép lan truyền gradient từ tác vụ cuối về các bộ mã hóa.

Nghiên cứu được triển khai và đánh giá trên các benchmark chuẩn với độ dài video đa dạng từ 3 giây (TGIF-QA) đến 180 giây (ActivityNet Captions), bao gồm hai nhóm tác vụ chính: text-to-video retrieval và video question answering.

2. Phương pháp



Hình 2. Pipeline tổng thể của ClipBERT: từ video thô và văn bản đầu vào, qua các bước lấy mẫu thưa, mã hóa đặc trưng đa phương thức bằng 2D CNN và Transformer, đến dự đoán cuối cùng cho các tác vụ video-ngôn ngữ.

Phương pháp nghiên cứu được thực hiện theo pipeline tổng thể như Hình 2, trong đó mỗi bước xử lý đảm nhiệm một vai trò chức năng cụ thể:

Bước 1: Lấy mẫu thưa và tiền xử lý

Đầu vào gồm video V và văn bản S. Video được chia thành segments, sau đó lấy mẫu ngẫu nhiên N_train clips (N_train=1 hoặc 2). Mỗi clip được lấy mẫu T frames đều đặn ($T \in \{2,4,8,16\}$) và resize về kích thước L×L pixels. Văn bản được tokenize

với giới hạn tối đa 20 tokens.

Bước 2: Mã hóa đặc trưng đa phương thức

Mỗi clip được xử lý qua ResNet-50 (lấy output từ conv5) để tạo spatial feature maps, sau đó áp dụng temporal fusion (mean pooling hoặc Conv2D) để tổng hợp thông tin từ T frames. Đặc trưng video và text được concatenate cùng với type embeddings, sau đó đưa vào Transformer 12 lớp. Output từ token [CLS] được sử dụng làm biểu diễn cho cặp clip-text.

Bước 3: Huấn luyện hai giai đoạn

(i) Pre-training trên COCO và Visual Genome (5.6M image-text pairs) với image-text matching và masked language modeling; (ii) Fine-tuning end-to-end cho từng tác vụ video-văn bản.

Bước 4: Dự đoán và aggregation

Tùy theo tác vụ, thêm MLP classifier cho QA hoặc similarity scoring cho retrieval. Trong inference, tăng số clips lên N_test=16 để cải thiện độ chính xác (chiến lược "sparse-training then dense-inference"). Dự đoán từ nhiều clips được tổng hợp qua mean pooling hoặc LogSumExp.

Bước 5: Đánh giá và phân tích:

Mô hình được đánh giá với Recall@K và Median Rank cho retrieval, Accuracy cho QA. Các thí nghiệm ablation phân tích ảnh hưởng của số clips, frames, temporal fusion, và so sánh chi phí tính toán giữa sparse và dense sampling.

KẾT QUẢ MONG ĐỢI

1. Luận văn dự kiến tìm hiểu và đánh giá thành công cơ chế lấy mẫu thưa (sparse sampling) trong học đa phương thức video-ngôn ngữ. Kết quả nghiên cứu nhằm chứng minh giả thuyết "less is more": chỉ với một vài clip ngắn lấy mẫu ngẫu nhiên, mô hình vẫn có thể học được biểu diễn hiệu quả mà không cần xử lý toàn bộ khung hình video. Đây là kết quả nền tảng phục vụ cho việc phát triển

các hệ thống video-ngôn ngữ hiệu quả hơn trong tương lai.

2. Hệ thống ClipBERT được kỳ vọng đạt hiệu suất cạnh tranh hoặc vượt trội so với các phương pháp sử dụng đặc trưng offline dày đặc trên các tác vụ text-to-video retrieval và video question answering. Mô hình dự kiến có khả năng hiểu tốt nội dung video và trả lời chính xác các câu hỏi phức tạp, đồng thời cho thấy khả năng tổng quát hóa tốt trên các miền video có độ dài đa dạng. Tính khả thi của hướng tiếp cận được thể hiện qua khả năng vận hành theo pipeline hoàn chỉnh từ đầu vào đến đầu ra.
3. Về đánh giá, luận văn dự kiến thực hiện các thí nghiệm ablation để phân tích đóng góp của từng thành phần trong kiến trúc: số clips, số frames, phương pháp temporal fusion, và vai trò của image-text pre-training. Kết quả đánh giá sẽ làm rõ ưu điểm về chi phí tính toán và bộ nhớ của sparse sampling so với dense sampling, mở ra hướng nghiên cứu mới về tối ưu hóa hiệu quả cho các tác vụ đa phương thức. Trên cơ sở đó, luận văn cung cấp bằng chứng thực nghiệm cho hiệu quả thiết kế và định hướng các cải tiến tiếp theo.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, Jingjing Liu:
Less Is More: ClipBERT for Video-and-Language Learning via Sparse Sampling. CVPR 2021: 7331-7341.

[2]. Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, Cordelia Schmid: VideoBERT: A Joint Model for Video and Language Representation Learning. ICCV 2019: 7463-7472.

[3]. Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, Jingjing Liu: HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training. EMNLP 2020: 2046-2065.

[4]. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova:
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019: 4171-4186.

[5]. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun:
Deep Residual Learning for Image Recognition. CVPR 2016: 770-778.

[6]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin:
Attention is All you Need. NeurIPS 2017: 5998-6008.

[7]. Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C. Lawrence Zitnick:
Microsoft COCO: Common Objects in Context. ECCV 2014: 740-755.