HỌC ĐA PHƯƠNG THỨC VIDEO-NGÔN NGỮ BẰNG PHƯƠNG PHÁP LẤY MẪU THƯA

Lâm Thanh Ngân - 250202015

Tóm tắt

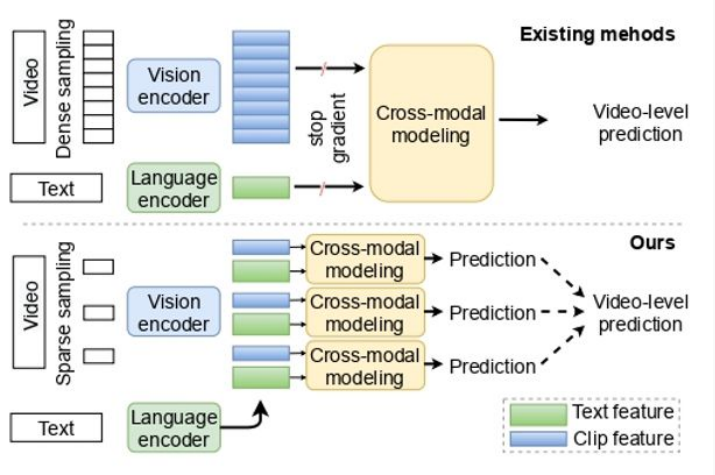
- Lớp: CS2205.CH201
- Link Github: https://github.com/LamThanhNgan/CS2205.CH201
- Link YouTube video: https://www.youtube.com/watch?v=MLMj3KqrkpE
- Họ và Tên: Lâm Thanh Ngân
- MSHV: 250202015

Giới thiệu

- **Bối cảnh:**
  - Video-language learning là bài toán quan trọng trong AI.
  - Ứng dụng: tìm kiếm video, trả lời câu hỏi về video, mô tả video tự động.
- **Vấn đề nghiên cứu:**
  - Phương pháp hiện tại dùng dense video features.

  → Chi phí tính toán và bộ nhớ cao.

  → Features không được tối ưu cho downstream tasks.
- **Input/Output:**
  - Input: Video thô (raw pixels) + Văn bản (text tokens).
  - Output: Dự đoán cho retrieval/QA tasks.



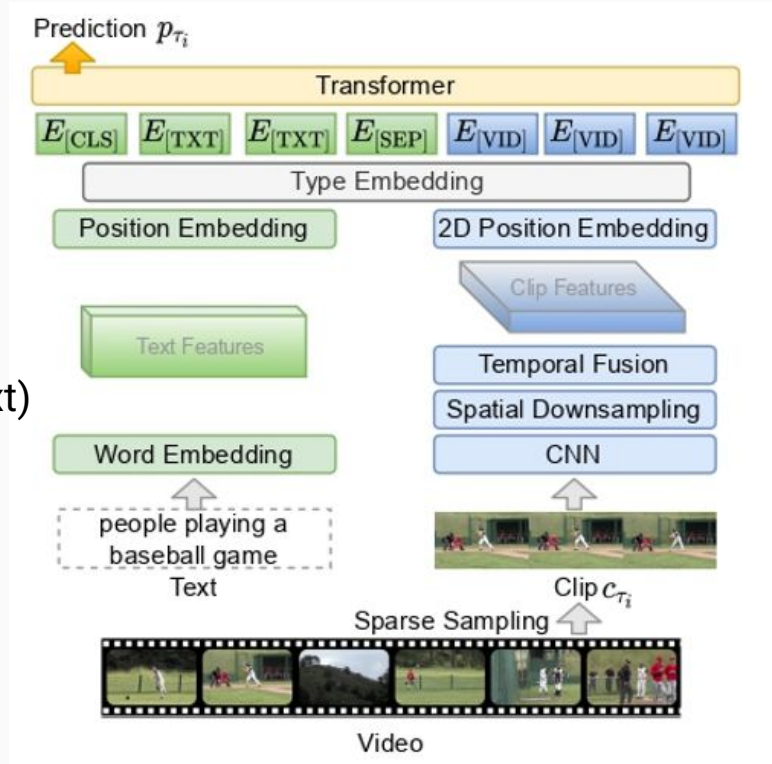
Mục tiêu

1. **Tìm hiểu cơ chế lấy mẫu thưa (Sparse Sampling)**
   - Kiến trúc end-to-end: 2D CNN + Transformer đa phương thức.
   - Phân tích giả thuyết "Less is More".
   - Chiến lược "sparse-training then dense-inference".
2. **Triển khai pipeline huấn luyện hai giai đoạn**
   - Pre-training trên dữ liệu ảnh-văn bản (COCO, Visual Genome).
   - Fine-tuning cho các tác vụ video-ngôn ngữ:
     - Text-to-video retrieval
     - Video question answering
3. **Đánh giá định lượng trên benchmark chuẩn**
   - Datasets: MSR-VTT, DiDeMo, TGIF-QA, MSRVTT-QA
   - Metrics: Recall@K, Median Rank, Accuracy
   - Ablation study: số clips, frames, pre-training

Nội dung và Phương pháp

- **Kiến trúc ClipBERT:**
  - Sparse Sampling: Chỉ lấy 1-2 clips ngẫu nhiên thay vì toàn bộ video
  - Vision Encoder: 2D CNN (ResNet-50)

  → nhẹ hơn 3D CNN
  - Language Encoder: BERT tokenizer + embeddings
  - Cross-modal Fusion: Transformer 12 lớp
  - Training: Pre-train (ảnh-text) → Fine-tune (video-text)
  - Inference: Aggregate từ 16 clips → prediction



Kết quả dự kiến

1. **Về giả thuyết nghiên cứu:**
   - Chứng minh "Less is More": chỉ vài clip ngắn đủ để học biểu diễn hiệu quả
   - Không cần xử lý toàn bộ khung hình video
2. **Về hiệu suất mô hình:**
   - Kỳ vọng đạt hiệu suất cạnh tranh/vượt trội so với các phương pháp dùng dense features
   - Khả năng tổng quát hóa tốt trên video độ dài đa dạng
3. **Về phân tích thực nghiệm (Ablation):**
   - Đánh giá đóng góp: số clips, frames, temporal fusion, image-text pre-training
   - So sánh chi phí tính toán: sparse vs dense sampling

Tài liệu tham khảo

[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.