

PROPOSAL: TÓM TẮT VIDEO BÀI GIẢNG DỰA TRÊN CẢM XÚC

Tên đề tài (Tiếng Việt): Nghiên cứu và Phát triển Mô hình Tóm tắt Video Bài giảng Tự động dựa trên Đường cong Cảm xúc (Affective Curve) và Mạng nơ-ron Chú ý (Attention-based Networks).

Tên đề tài (Tiếng Anh - Suggested): *Affective Video Summarization in E-Learning: An Unsupervised Framework using Deep Reinforcement Learning and Temporal Attention Mechanisms.*

1. Đặt vấn đề (Problem Statement)

- Thực trạng:** Trong kỷ nguyên EdTech, khối lượng video bài giảng được lưu trữ là khổng lồ. Tuy nhiên, việc xem lại (review) toàn bộ video để đánh giá chất lượng dạy và học là bất khả thi về mặt thời gian.
- Hạn chế công nghệ:** Các kỹ thuật tóm tắt video hiện tại (Video Summarization - VS) thường tập trung vào sự thay đổi hình ảnh (Visual diversity) hoặc chuyển động (Motion). Trong giáo dục, một sinh viên ngồi im nghe giảng (ít chuyển động) có thể đang rất tập trung, trong khi một sinh viên quay ngang ngửa (nhiều chuyển động) lại mất tập trung. Các thuật toán VS truyền thống sẽ chọn sai “highlight”.
- Câu hỏi nghiên cứu (Research Question):** Làm thế nào để định nghĩa “Độ quan trọng” (Importance Score) của một khung hình dựa trên trạng thái cảm xúc của người học thay vì các đặc trưng thị giác cấp thấp, và tự động tạo ra một video tóm tắt giữ lại được nội dung ngữ nghĩa quan trọng nhất?

2. Mục tiêu nghiên cứu (Research Objectives)

- Mô hình hóa đường cong cảm xúc (Affective Curve Modeling):** Xây dựng thuật toán chuyển đổi chuỗi cảm xúc rời rạc của người học thành một tín hiệu liên tục (continuous signal) biểu diễn mức độ tương tác (Engagement Level) theo thời gian.
- Phát triển mạng chọn lọc khung hình (Frame Selection Network):** Ứng dụng cơ chế Attention hoặc Reinforcement Learning để lựa chọn các đoạn video (segments) có mật độ cảm xúc cao (Emotional Density) và loại bỏ các đoạn thừa (Redundancy).
- Đảm bảo tính mạch lạc (Temporal Coherence):** Video tóm tắt không được cắt vụn vặt gây khó chịu. Cần thuật toán tối ưu hóa (như Knapsack Problem) để chọn các đoạn video có độ dài phù hợp.

3. Tổng quan tài liệu & Khoảng trống nghiên cứu (Literature Review)

3.1. Tóm tắt video truyền thống (Unsupervised VS):

- Các phương pháp kinh điển như **Clustering (K-Means)** hoặc **Dictionary Learning** cố gắng chọn ra các frame đại diện cho các cụm hình ảnh khác nhau. Tuy nhiên, chúng không hiểu ngữ nghĩa cảm xúc [1].

3.2. Tóm tắt video dựa trên học sâu (Deep Video Summarization):

- **LSTM & Bi-LSTM (2018-2020)**: Sử dụng mạng hồi quy để mô hình hóa sự phụ thuộc giữa các frame.
- **Self-Attention & Transformer (2022-2024)**: Các mô hình như VASNet hay mô hình dựa trên Transformer đang là SOTA. Chúng cho phép mô hình nhìn toàn cục video để quyết định đoạn nào quan trọng nhất [2].
- **Reinforcement Learning (RL)**: Coi việc chọn frame là hành động của một Agent để tối đa hóa phần thưởng (Reward) là độ đại diện của video tóm tắt [3].

3.3. Khoảng trống nghiên cứu (The Gap):

- Hầu hết các nghiên cứu VS hiện nay tập trung vào datasets như *SumMe* hay *TVSum* (video thể thao, du lịch, sự kiện).
- **Thiếu hụt nghiên cứu về “Affective Summarization” trong Giáo dục**: Rất ít công trình kết hợp kết quả nhận diện cảm xúc (FER) làm đầu vào cho bài toán tóm tắt. Đây là điểm đề tài có thể tạo ra đóng góp mới: **“Affective-driven Summarization Agent”**.

4. Tài liệu tham khảo minh chứng (References)

Tập trung vào các bài báo về Video Summarization và Affective Computing (2022-2024):

1. Về kỹ thuật Video Summarization (SOTA):

- *Paper*: “Dilated Temporal Graph Reasoning for Video Summarization” (AAAI 2024). *Bài báo này đề xuất dùng Graph để liên kết các frame xa nhau, rất hay.*
- *Paper*: “Self-Supervised Video Summarization via Contrastive Learning” (CVPR 2023).

2. Về phân tích cảm xúc trong video:

- *Paper*: “Multimodal Affective Analysis for Video Summarization” (IEEE Transactions on Multimedia, 2023). *Tài liệu tham khảo chính yếu cho đề tài này.*
- *Paper*: “Highlight Detection in Educational Videos using Student Engagement Signals” (International Conference on AI in Education, 2023).

5. Phương pháp nghiên cứu & Kiến trúc đề xuất (Methodology)

Đề tài cần xây dựng một hệ thống gồm 3 khối chức năng (Block):

Block 1: Trích xuất đặc trưng Cảm xúc - Thị giác (Visual-Affective Feature Extraction)

Thay vì chỉ đưa ảnh vào model, chúng ta đưa vào một vector lai ghép:

- **Visual Features (F_v)**: Sử dụng **GoogleNet** hoặc **ResNet-101** (pre-trained trên ImageNet) để trích xuất đặc trưng hình ảnh của khung hình (bảng đen, slide, giáo viên).
- **Affective Features (F_a)**: Sử dụng một mô hình FER (như ở Đề tài 3) để trích xuất vector xác suất cảm xúc (ví dụ: [0.1,0.8,0.1] cho [Buồn, Vui, Giận]).
- **Fusion**: Nối (Concatenate) hai vector này lại: $X_t = [F_v, F_a]$.

Block 2: Mạng đánh giá tầm quan trọng (Importance Evaluation Network)

Đây là “bộ não” của hệ thống, quyết định frame nào được giữ lại.

- **Kiến trúc**: Sử dụng **Bi-directional LSTM (Bi-LSTM)** kết hợp với **Self-Attention**.
 - *Input*: Chuỗi vector $X = \{x_1, x_2, \dots, x_T\}$.
 - *Mechanism*: Bi-LSTM quét video theo 2 chiều (quá khứ và tương lai) để hiểu ngữ cảnh. Attention layer sẽ tính toán trọng số α_t (Importance Score) cho từng frame t .
 - *Loss Function (Điểm nhấn)*: Thay vì nhân thủ công (rất tốn kém), đề xuất sử dụng **Unsupervised Reinforcement Learning**.
 - *Agent*: Bộ chọn frame.
 - *Reward*: $R = R_{diversity} + R_{representativeness}$. (Tóm tắt phải đa dạng và đại diện tốt cho video gốc).

Block 3: Tạo sinh tóm tắt (Summary Generation)

- Đầu ra của Block 2 là một chuỗi điểm số $S = \{s_1, s_2, \dots, s_T\}$ (giá trị từ 0 đến 1).
- Áp dụng thuật toán **Knapsack Problem** (Bài toán cái túi) để chọn các đoạn video sao cho tổng điểm quan trọng là lớn nhất, nhưng tổng thời gian không vượt quá 15 thời lượng video gốc.
- **Làm mượt (Smoothing)**: Sử dụng các bộ lọc hình thái học để tránh việc video bị cắt quá vụn (ví dụ: giữ ít nhất 3 giây liên tục).

6. Kế hoạch thực nghiệm (Implementation Plan)

6.1. Dữ liệu (Datasets)

Bài toán này khó về dữ liệu vì cần video dài.

1. **TVSum / SumMe:** Dữ liệu chuẩn để test thuật toán Summarization (không phải giáo dục, nhưng cần để so sánh baseline).
2. **Educational Dataset (Tự thu thập):**
 - Thu thập 20 video bài giảng (zoom recording).
 - **Gán nhãn (Annotation):** Yêu cầu 3 sinh viên giỏi xem lại video và đánh dấu các đoạn “quan trọng”. Sự đồng thuận của 3 người này sẽ là Ground Truth.

6.2. Metrics đánh giá

- **F-measure:** So sánh độ chồng lấp (Overlap) giữa video máy tạo ra và video người gán nhãn.
- **Rank Correlation Coefficients (Kendall’s τ , Spearman’s ρ):** So sánh thứ tự xếp hạng các đoạn quan trọng.
- **User Study (Quan trọng cho ứng dụng):** Mời 10 giáo viên xem tóm tắt và đánh giá theo thang Likert 5 điểm về mức độ hữu ích.

7. Tài nguyên & Công cụ (Resources)

A. GitHub Repositories (Nền tảng):

1. **Deep-Video-Summarization (Comprehensive List):** github.com/flyywh/Video-Summarization-Pytorch
 - Repo này cài đặt lại hầu hết các thuật toán SOTA như VASNet, DSNet. Đây là điểm khởi đầu tuyệt vời.
2. **VASNet (Visual Attention Summary Network):** github.com/Jyl1999/VASNet
 - Mô hình đơn giản nhưng hiệu quả cao, sử dụng cơ chế Attention. Rất dễ để sửa đổi (modify) thêm đầu vào cảm xúc.
3. **Pyscenedetect:** github.com/Breakthrough/PySceneDetect
 - Thư viện giúp phát hiện ranh giới các cảnh (Shot detection). Cần dùng cho bước tiền xử lý.

B. Sách & Lý thuyết:

- *Video Summarization* - Springer Briefs in Computer Science.
- *Reinforcement Learning: An Introduction* - Sutton & Barto (Để hiểu cách thiết kế Reward function cho Block 2).

Lời khuyên:

1. **Đừng tóm tắt Frame, hãy tóm tắt Shot:** Frame là quá nhỏ (1/30 giây). Hãy chia video thành các Shot (cảnh quay) dài 2-5 giây trước khi đưa vào xử lý. Điều này giảm khối lượng tính toán xuống 100 lần.
2. **Kết hợp Audio (Multimodal):** Trong bài giảng, giọng nói giảng viên to lên hoặc dồn dập thường là lúc quan trọng.

- *Gợi ý:* Nếu sinh viên tích hợp thêm đặc trưng âm thanh (Audio Energy / Pitch) vào mô hình, điểm đề tài sẽ tăng đáng kể vì tính đa phương thức (Multimodality).