Cross Interaction Network for Natural Language Guided Video Moment Retrieval

Thành viên:

20521609 – Nguyễn Hoàng Minh

Link paper: CI-MHA

Nội dung

01 Introduction

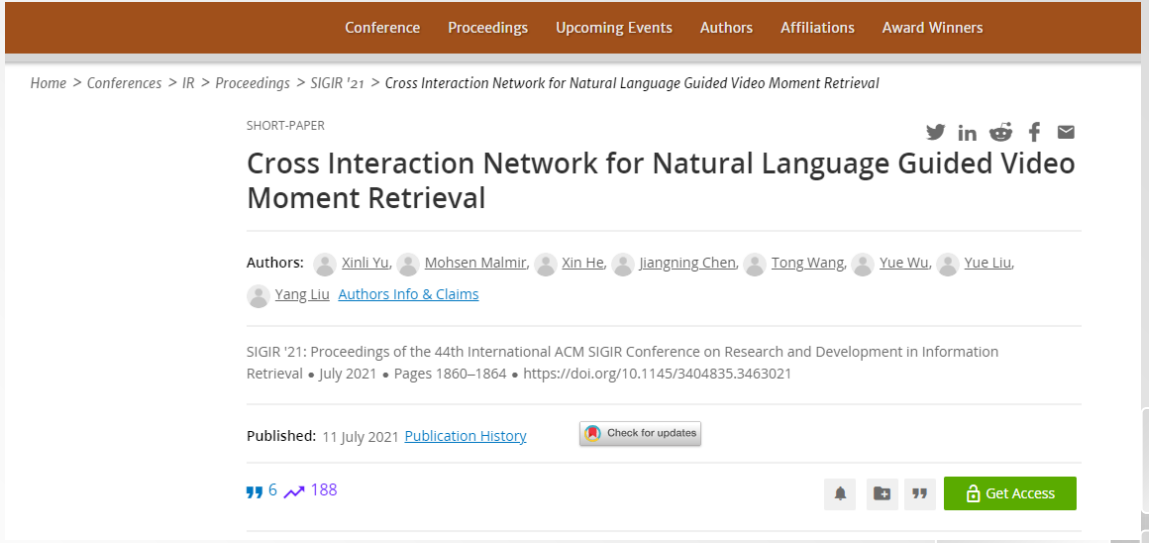
02 Task description

03 Background

04 Cross Interaction Network

05 Experiments

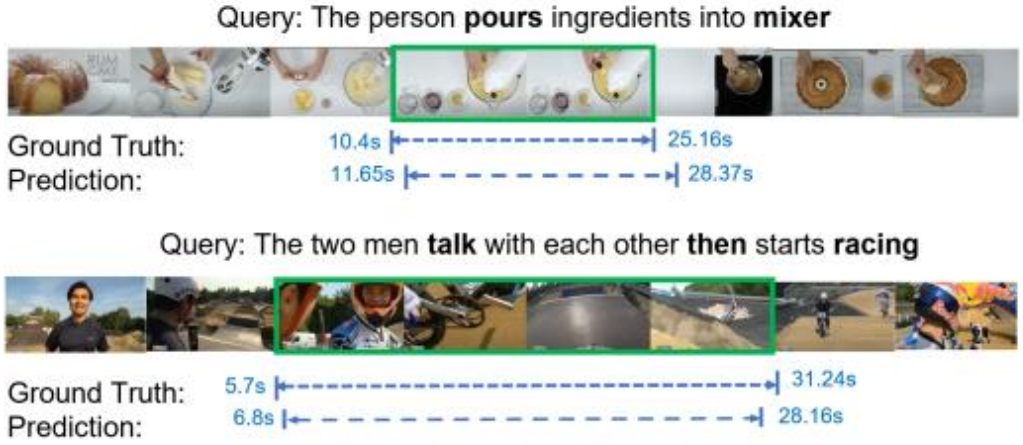
1. Introduction



Đóng góp của nghiên cứu:

- Đề xuất **the cross interaction multi-head attention mechanism** theo thể hợp nhất nhằm liên kết các đặt trưng từ video và đặt trưng từ truy vấn ngôn ngữ theo hai chiều.
- Đề xuất **a multi-task training objective** bao gồm: **1) start/end prediction task**, **2) moment segmentation task**.

2. Task description



Input: Cho video và 1 câu query

Output: Thời điểm (bắt đầu/kết thúc) trong video

3. Background

- **Convolutional 3D (C3D)**: Video được chia thành một chuỗi gồm các segment (16-frame). C3D trích xuất features từ các segment. Đầu ra là ma trận ($N \times D$), với $\boldsymbol{D}$ **= 500** và $\boldsymbol{N}$ **=** $\boldsymbol{M}$**/16** với $\boldsymbol{M}$ là số lượng frames trong video chưa cắt.
- **Two-Stream Inflated 3D ConvNets (I3D)**: Video được chia thành một chuỗi gồm các segment (24-frame per second). I3D lấy 64 frames liên tiếp làm đầu vào. Đầu ra là a snippet-level feature vector.
- **Position Embedding**

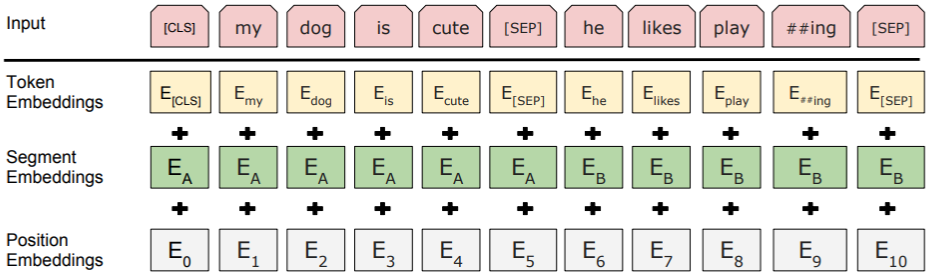


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

4. Cross Interaction Network

Model Architecture

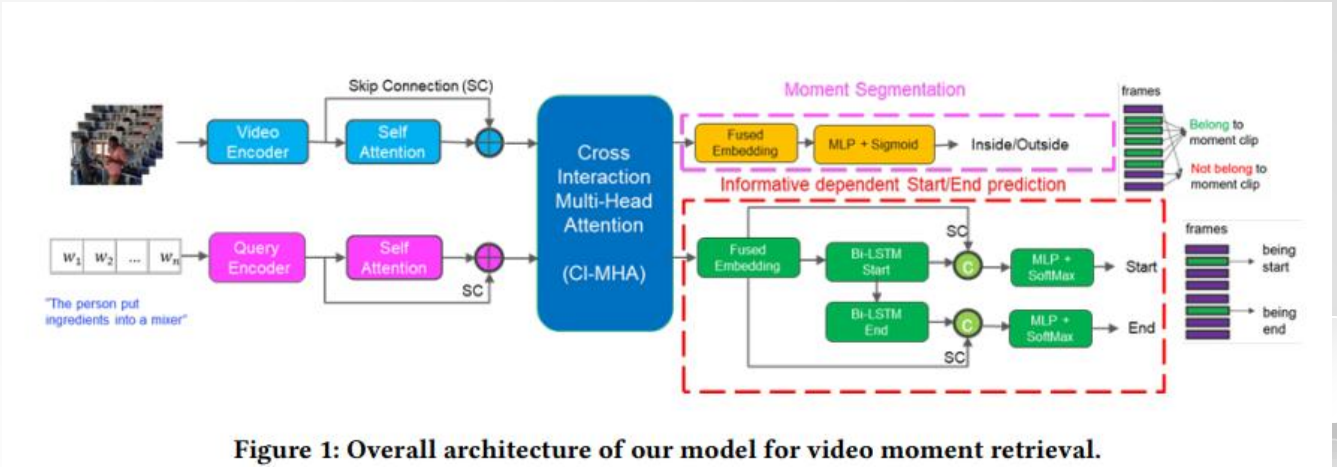


Figure 1: Overall architecture of our model for video moment retrieval.

- **Mỗi lần dự đoán:**
  - Input: Thời điểm *t* (Frame và Query)
  - Output: Label (Inside/Outside) + Label ([..0,0,1,0,..,0,1,0,0..])
- **Query và video** được encode thành **features vector**.
- **Self-attention** dùng để trích xuất **video embeddings** và **query embeddings**.

4. Cross Interaction Network

Model Architecture

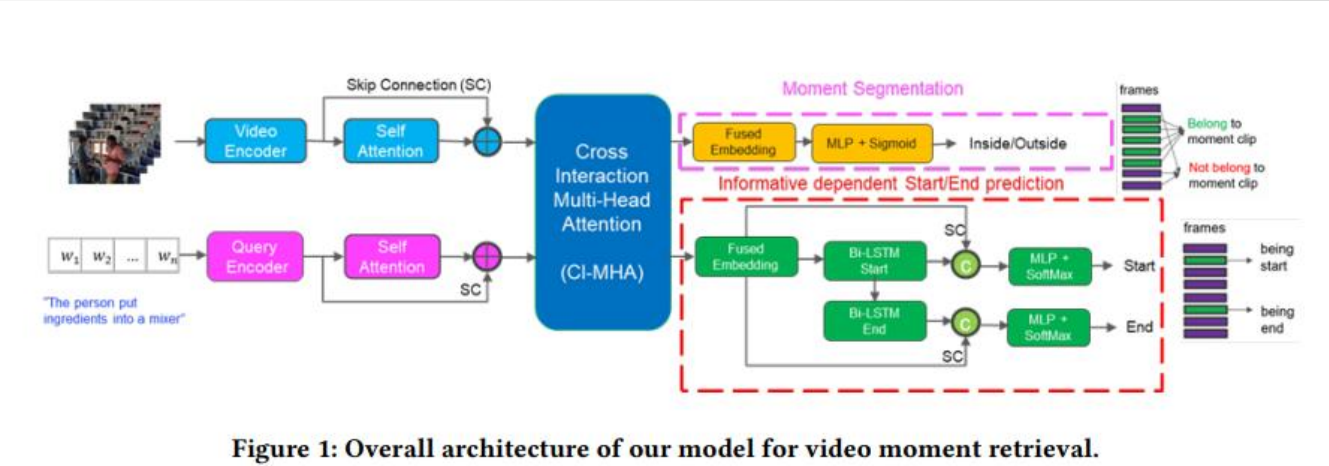


Figure 1: Overall architecture of our model for video moment retrieval.

- **2 embeddings** sẽ được hợp nhất bởi **CI-MHA** để tìm mối quan hệ giữa **video representation với query context** và **ngược lại.**
- **2 enriched representations** sẽ được concatenate và đưa qua **multi-task training module** để dự đoán thời điểm trong video.

4. Cross Interaction Network

Video Encoders

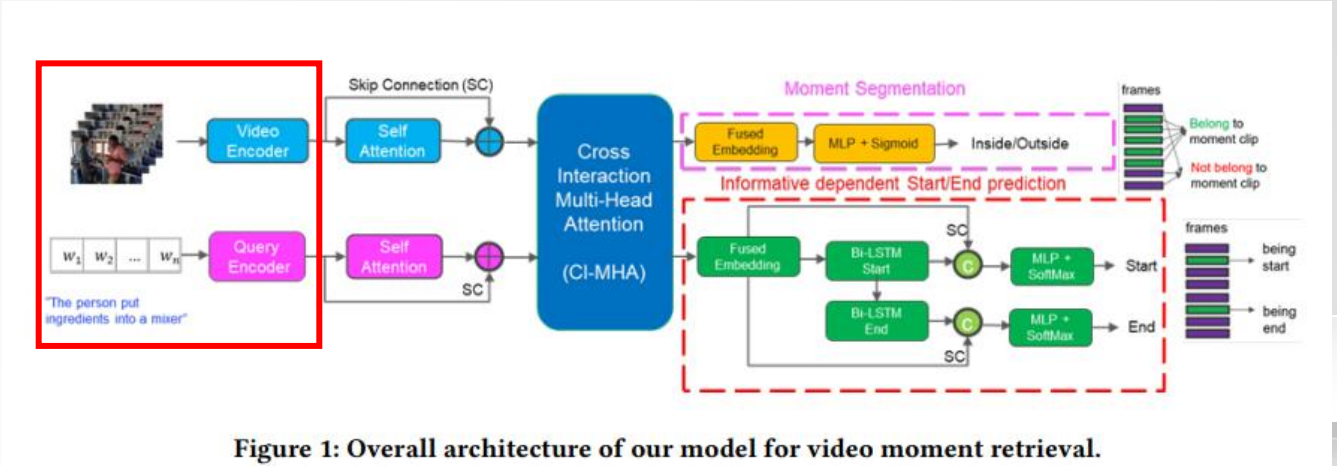


Figure 1: Overall architecture of our model for video moment retrieval.

- **C3D network** pre-trained trên sport1M và **I3D network** pre-trained trên Kinetics được sử dụng để **video feature encoder.**
- **Positional Encoding:**
  - **A temporal positional embedding** Được thêm vào các **video segment feature** tương ứng nhằm cung cấp thông tin về vị trí của từng frame với mục đích cải thiện độ chính xác.
  - **The positional encoding** được xây dựng giống như **Position Embeddings** ở **BERT**.

4. Cross Interaction Network

Video Encoders

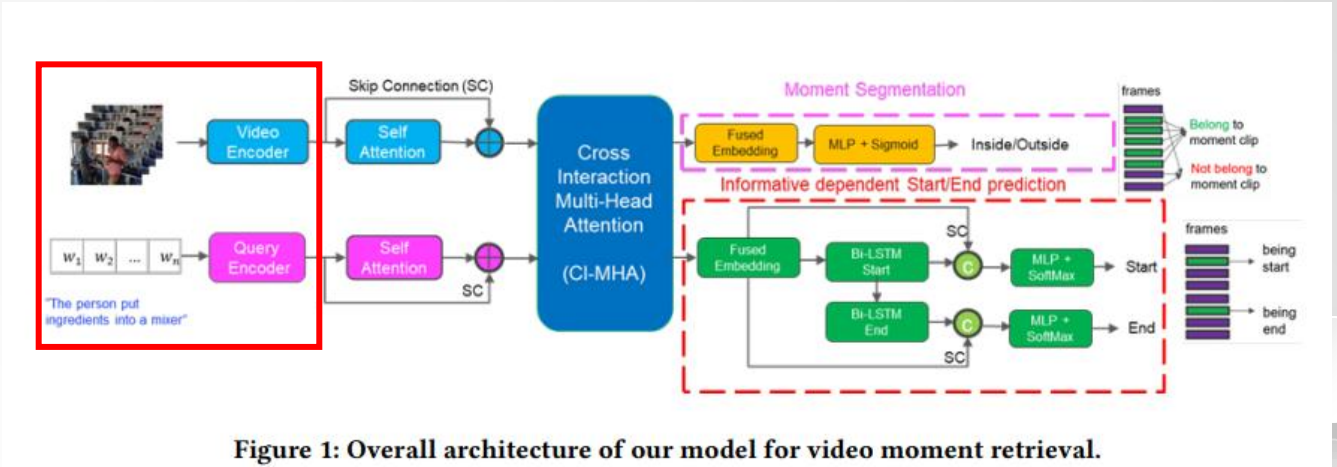


Figure 1: Overall architecture of our model for video moment retrieval.

- **Video encoder embedding output**

$$\mathbf{v}_0 = f_{\text{encoder}}(V), \quad (1)$$

  - $\mathbf{v}_0$ có kích thước (N, Dv); với V là video đầu vào, N là độ dài video, Dv là độ dài video embedding.
- **Query embedding output**

$$\mathbf{q}_0 = f_{\text{encoder}}(W), \quad (2)$$

  - $\mathbf{q}_0$ có kích thước (L, Dq); với W và video đầu vào, L là độ dài query, Dq là độ dài query embedding.

4. Cross Interaction Network

Video/Query Self Interaction

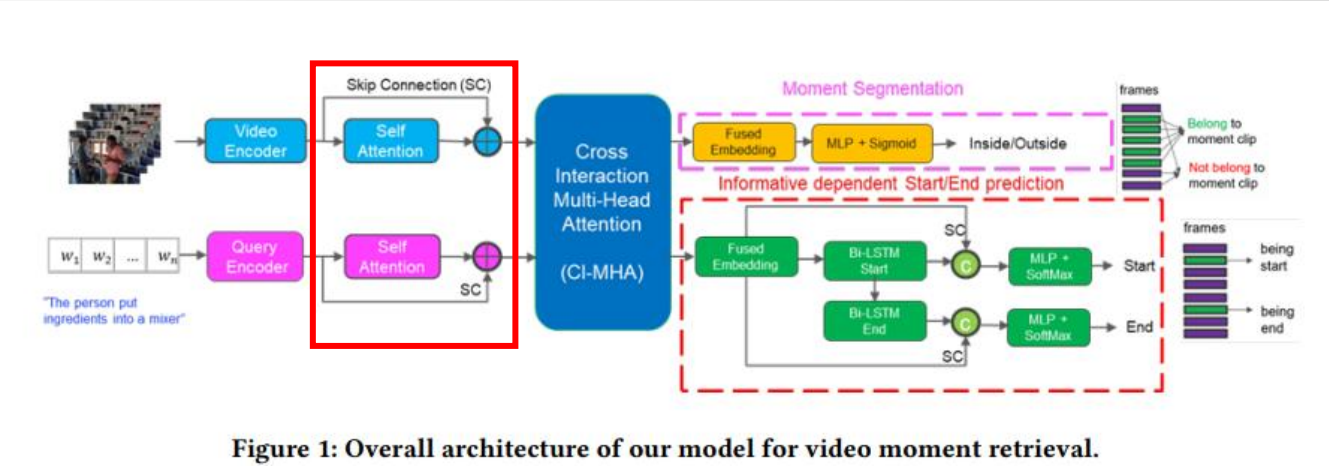


Figure 1: Overall architecture of our model for video moment retrieval.

- Họ sử dụng **Self Attention** trước khi sử dụng CI-MHA module. **Video self-attention** học các mối liên hệ giữa các **frame**, **Query self-attention** học các mối liên hệ giữa các **từ**.
- **Skip-connection (SC)** được thêm vào để tránh mất thông tin.

$$\mathbf{q} = \mathrm{MHA}(Q = \mathbf{q}_0, K = V = \mathbf{q}_0) + \mathbf{q}_0, \quad (3)$$

$$\mathbf{v} = \mathrm{MHA}(Q = \mathbf{v}_0, K = V = \mathbf{v}_0) + \mathbf{v}_0. \quad (4)$$

**MHA : Multi-head Attention*

4. Cross Interaction Network

Visual-Language Fusion

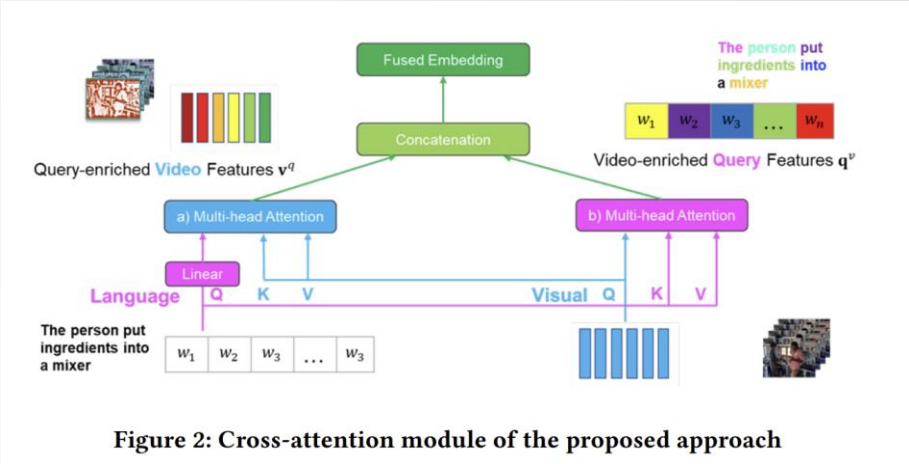


Figure 2: Cross-attention module of the proposed approach

- Họ đề xuất ***Cross interaction multi-head attention*** **(CI-MHA)** cho Visual-Language fusion.
- **Feedforward layer** được sử dụng cho **video encoder embedding v** và **query encoder embedding q**, sau đó **Positional Encoding** sẽ được thêm vào **video encoder embedding.**

$$\hat{\mathbf{v}} = \text{FeedForward}(\mathbf{v}) + \text{Positional Encoding}(\mathbf{v}), \quad (5)$$

$$\hat{\mathbf{q}} = \text{FeedForward}(\mathbf{q}). \quad (6)$$

  - $\hat{v}$ và $\hat{q}$ có kích thước (N, D); với N là độ dài video, D là độ dà embedding.

4. Cross Interaction Network

Visual-Language Fusion

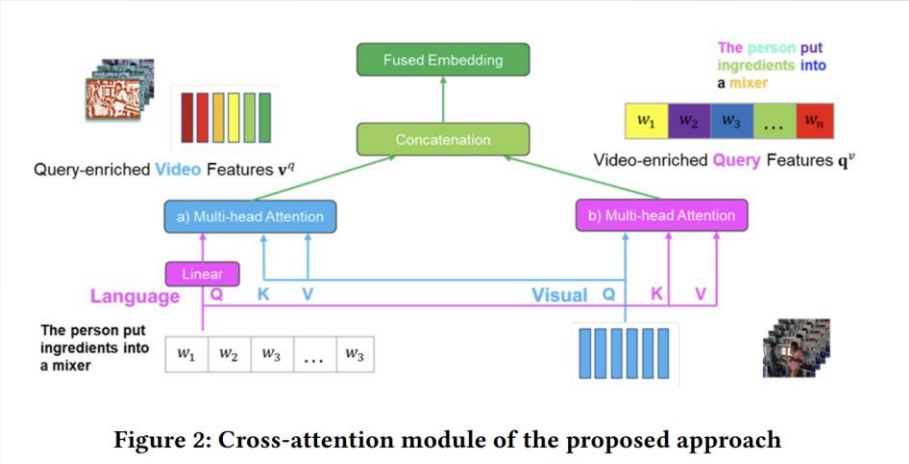


Figure 2: Cross-attention module of the proposed approach

- Ở bài *(https://aclanthology.org/2020.acl-main.585.pdf)* họ **chỉ** tập trung vào mối quan hệ **query-to-video** thông qua **Query-enriched Video Feature embedding**. Tác giả trong báo cáo này cho rằng, mối quan hệ giữa **video-to-query** cũng quan trọng.

4. Cross Interaction Network

Visual-Language Fusion

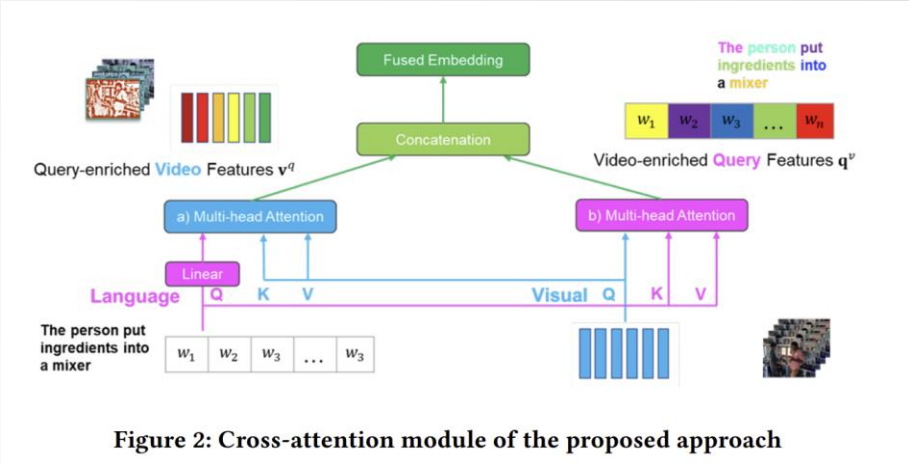


Figure 2: Cross-attention module of the proposed approach

- **CI-MHA** quan tâm tới cả 2 vấn đề **query-to-video** và **video-to-query**.

$$\mathbf{v}^q = \mathrm{MHA}(Q = \hat{\mathbf{q}}, K = V = \hat{\mathbf{v}}), \quad (7)$$

$$\mathbf{q}^v = \mathrm{MHA}(Q = \hat{\mathbf{v}}, K = V = \mathbf{q}). \quad (8)$$

- $\mathbf{v}^q$ và $\mathbf{q}^v$ sẽ được concatenation để làm đầu vào cho **prediction task. Fused embeddings** có chiều như sau:

$$\mathbf{s}^{v,q} = [\mathbf{v}^q, \mathbf{q}^v]. \quad (9)$$

4. Cross Interaction Network

Multi-task Training Objectives (2 Task): *Informative Dependent Start/End*

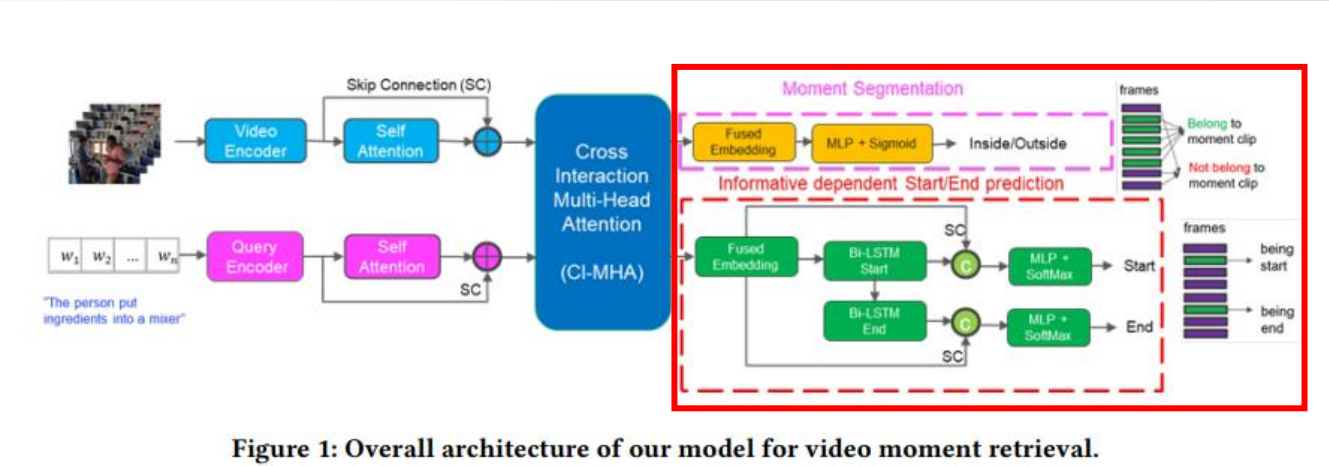


Figure 1: Overall architecture of our model for video moment retrieval.

- Dự đoán thời điểm (bắt đầu/kết thúc) trong video
- Cách tiếp cận giống **ExCL** *(https://arxiv.org/pdf/1904.02755.pdf)*.

$$\mathbf{h}_t^{\text{start}} = \text{Bi-LSTM}_{\text{start}}(\mathbf{s}_t^{v,q}, \mathbf{h}_{t-1}^{\text{start}}), \quad (10)$$

$$\mathbf{h}_t^{\text{end}} = \text{Bi-LSTM}_{\text{end}}(\mathbf{h}_{t-1}^{\text{start}}, \mathbf{h}_{t-1}^{\text{end}}), \quad (11)$$

  - Với $\mathbf{h}_0^{\text{start}}, \mathbf{h}_0^{\text{end}}$ sẽ được khởi tạo ngẫu nhiên cho LSTM layer đầu tiên. *(Tham số đầu là context đầu vào, Tham số thứ 2 là context đầu ra)*

4. Cross Interaction Network

Multi-task Training Objectives (2 Task): *Informative Dependent Start/End*

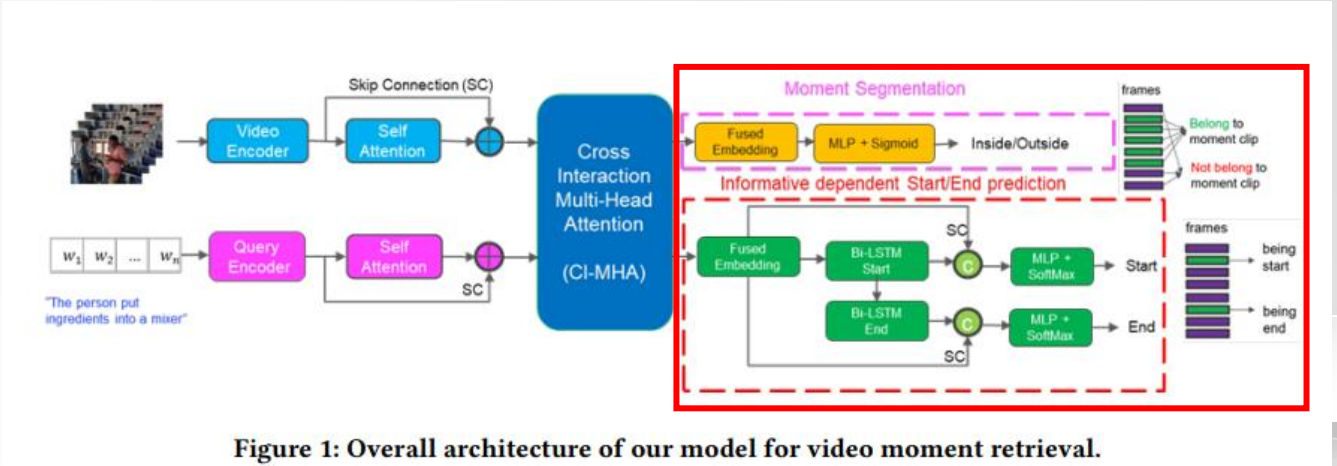


Figure 1: Overall architecture of our model for video moment retrieval.

- **Multi-Layer Perceptron (MLP) + Softmax layers**: Tạo ra phân phối Start/End

$$c_t^{\text{start}} = \text{MLP}_{\text{start}}([\mathbf{h}_t^{\text{start}}, \mathbf{s}_t^{v,q}]), \quad (12)$$

$$c_t^{\text{end}} = \text{MLP}_{\text{end}}([\mathbf{h}_t^{\text{end}} \mathbf{s}_t^{v,q}]). \quad (13)$$

- **Skip connections** thông qua concatenation (là $[\mathbf{h_t^{start}}, \mathbf{h_t^{end}}]$): Giúp giúp ngăn chặn tình trạng mất thông tin

4. Cross Interaction Network

Multi-task Training Objectives (2 Task): *Informative Dependent Start/End*

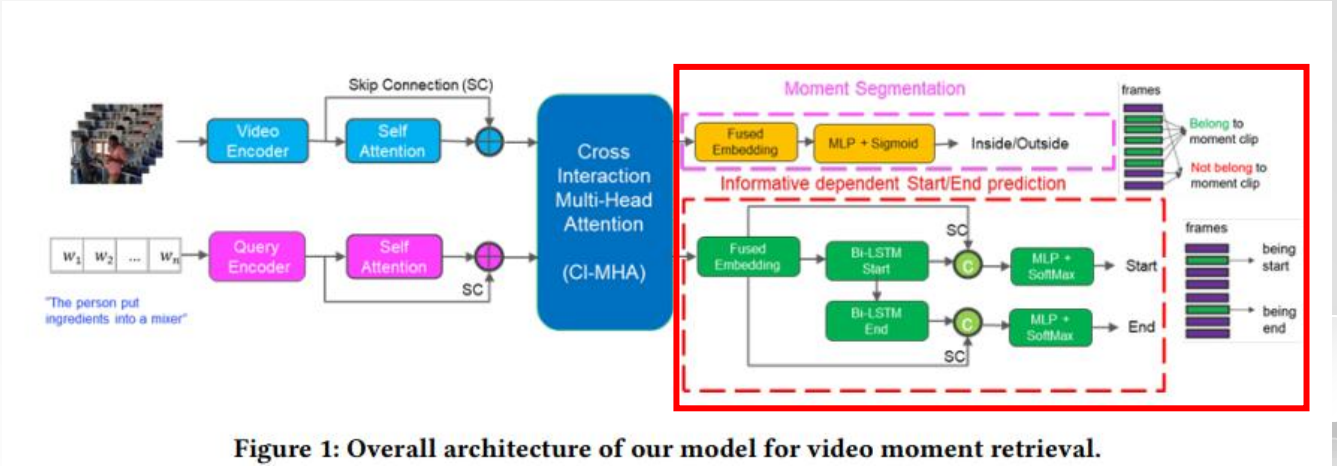


Figure 1: Overall architecture of our model for video moment retrieval.

- Chuẩn hóa lại dự đoán (theo chiều thời gian)

$$P_{\text{start}} = \text{SoftMax}(\text{C}^{\text{start}}), \quad P_{\text{end}} = \text{SoftMax}(\text{C}^{\text{end}}), \qquad (14)$$

  - Với $\boldsymbol{C^{start}} = [\boldsymbol{c_0^{start}}, \boldsymbol{c_1^{start}}, \ldots, \boldsymbol{c_T^{start}}]$, và tương tự cho $\boldsymbol{C^{end}}$.
- **Loss function**: negative log-likelihood.
- **Ground-truth labels:** two sparse one-hot vectors (start/end time point).

4. Cross Interaction Network

Multi-task Training Objectives (2 Task): *Moment Segmentation (MS)*

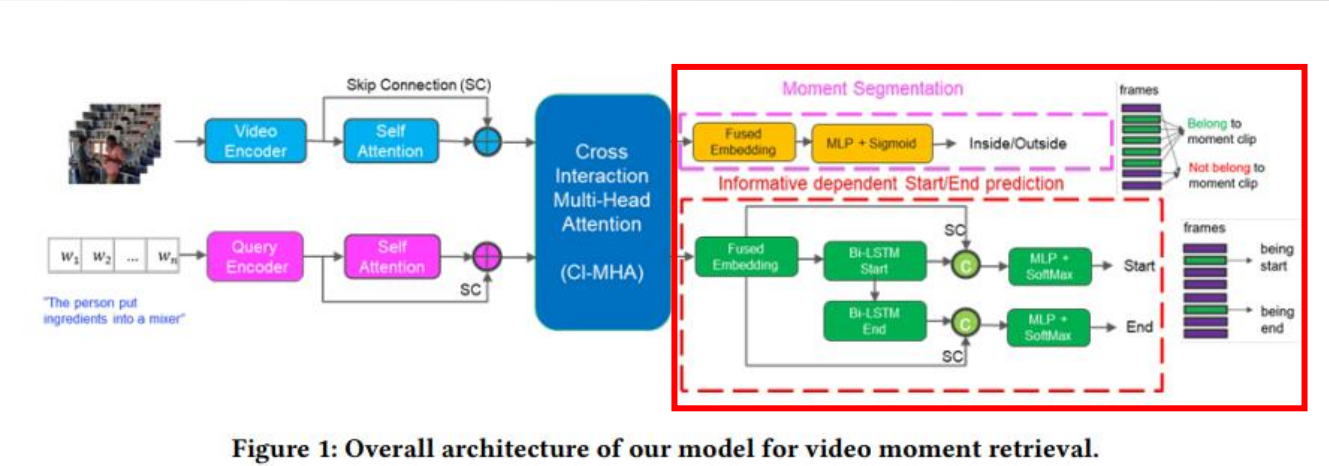


Figure 1: Overall architecture of our model for video moment retrieval.

- Annotators khó xác định chính xác **Being start** và **Being end**

→ Gây ra **start/end ground-truth labels** không những **thưa** mà còn **nhiễu**.

- Trong moment segmentation, the ground-truth bao hết khoảng thời gian từ start tới end (mọi frame trong ground-truth là **mẫu huấn luyện tích cực** (**positive training)**)

→ Task này tăng hiệu suất mô hình, chứng minh thông qua thực nghiệm.

4. Cross Interaction Network

Multi-task Training Objectives (2 Task): *Moment Segmentation (MS)*

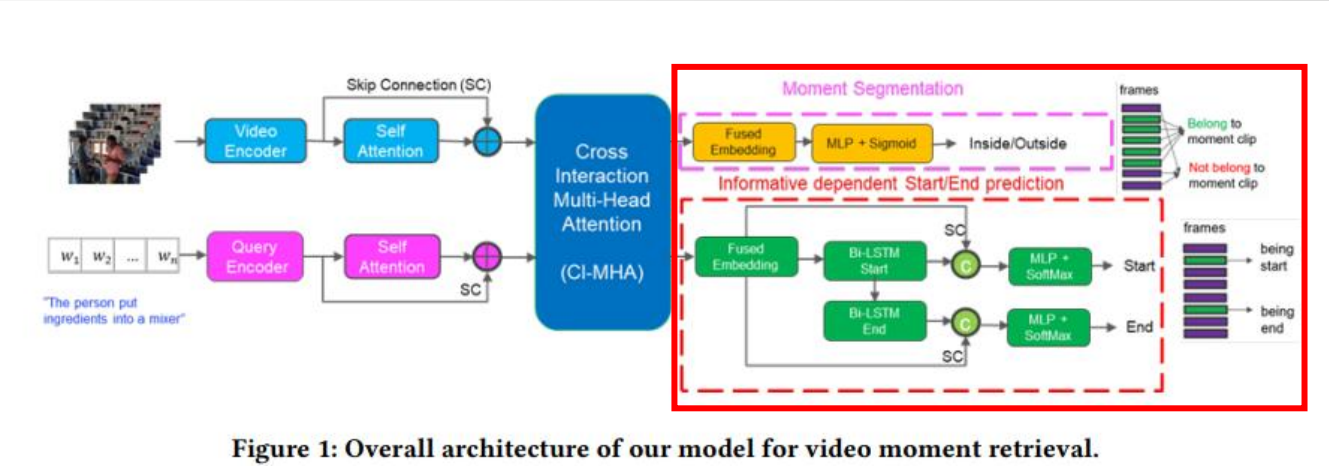


Figure 1: Overall architecture of our model for video moment retrieval.

- Tại mỗi thời điểm t, sử dụng **MLP + Sigmoid function**: dự đoán khả năng t có thuộc Thời điểm dự đoán hay không.

$$P_{in}(t) = \text{Sigmoid}(\text{MLP}_{\text{segment}}(\mathbf{s}_t^{v,q})), \quad (15)$$

$$P_{out}(t) = 1 - P_{in}(t). \quad (16)$$

- **Loss function**: binary cross entropy.
- **Ground-truth labels:** Inside/Outside.

5. Experiments

Datasets

- **The ActivityNet Captions dataset:** Có khoảng 20K video, với trung bình mỗi video chứa 3,65 câu đã được chú thích hoạt động theo thời gian và thời gian bắt đầu & kết thúc, tạo ra tổng cộng khoảng 100K câu.

- **Charades-STA dataset:** Chứa 9848 video với 157 hoạt động. Mỗi video chứa chú thích hoạt động theo thời gian và thời gian bắt đầu & kết thúc để phục vụ cho tác vụ xác định khoảnh khắc dựa trên truy vấn bằng ngôn ngữ.

5. Experiments

Implementation Details

- Huấn luyện mô hình end-to-end, với đầu vào là **raw video frames** và **natural language query**.
- Với query encoding:
  - Bi-LSTM được sử dụng để encode câu **query** GLOVE embedding 512 chiều
  - Độ dài tối đa của câu query: 25 từ
- Với video encoding
  - C3D features: 500 chiều
  - I3D features: 1024 chiều
  - Độ dài tối đa của frame: 128 frames
- Batch size: 100
- Adam optimizer
- Learning rate: 0.0001

5. Experiments

Comparison with State-of-the-art

Table 1: Evaluation on Charades-STA

Method	R@1, IoU=0.7	R@1, IoU=0.5	R@1, IoU=0.3
Random	3.03%	8.51%	-
LOGAN [13]	14.54%	34.68%	51.67%
MLVI [16]	15.80%	35.60%	-
ExCL [5]	22.40%	44.10%	-
MAN [18]	22.72%	46.23%	-
DRN [17]	31.75%	53.09%	-
MHA	30.49%	51.04%	60.85%
CI-MHA w/o MS	31.39%	52.41%	62.78%
CI-MHA	**35.27%**	**54.68%**	**69.87%**

5. Experiments

Comparison with State-of-the-art

Table 2: Evaluation on ActivityNet Captions

Method	R@1, IoU=0.7	R@1, IoU=0.5	R@1, IoU=0.3
MLVI [16]	13.60%	27.70%	45.30%
TripNet [6]	13.93%	32.19%	45.42%
ExCL [5]	23.9%	41.46%	**62.21%**
DRN [17]	23.24%	43.78%	-
MHA	21.76%	40.80%	59.73%
CI-MHA w/o MS	23.41%	41.73%	60.49%
CI-MHA	**25.13%**	**43.97%**	61.49%

References

- Tran, Du, et al. "Learning spatiotemporal features with 3d convolutional networks." *Proceedings of the IEEE international conference on computer vision*. 2015. Available: https://arxiv.org/pdf/1412.0767.pdf
- Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset." *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017. Available: https://arxiv.org/pdf/1705.07750.pdf
- Zhang, Hao, et al. "Span-based localizing network for natural language video localization." *arXiv preprint arXiv:2004.13931* (2020). Available: https://aclanthology.org/2020.acl-main.585.pdf
- Ghosh, Soham, et al. "Excl: Extractive clip localization using natural language descriptions." *arXiv preprint arXiv:1904.02755* (2019). Available: https://arxiv.org/pdf/1904.02755.pdf
- Sigurdsson, Gunnar A., et al. "Hollywood in homes: Crowdsourcing data collection for activity understanding." *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer International Publishing, 2016.
- Caba Heilbron, Fabian, et al. "Activitynet: A large-scale video benchmark for human activity understanding." *Proceedings of the ieee conference on computer vision and pattern recognition*. 2015.

Q&A

Cảm Ơn!