JOINT EVENT LOCALIZATION AND CAPTIONING USING PRETRAINED SEQUENCE-TO-SEQUENCE MODELS FOR DENSE VIDEO CAPTIONING

Huynh Ngoc Bao Long **Nguyen Huu Sang**

University of Information Technology - National University of Vietnam

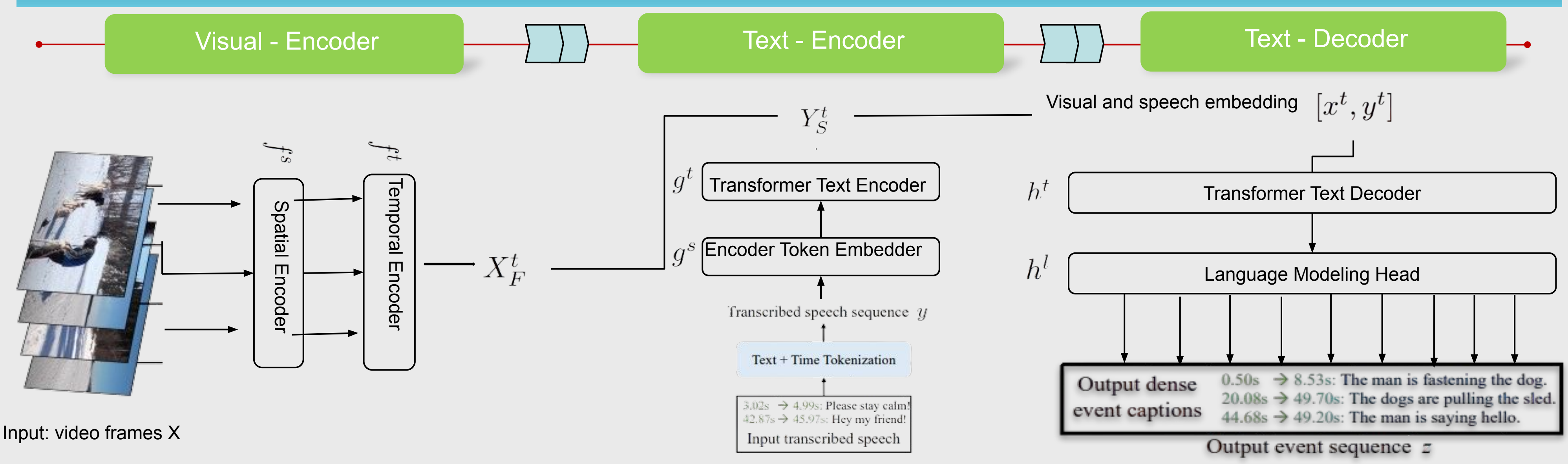
Motivations

The rapid growth of video content on digital platforms has made automatic video understanding a critical challenge. **Dense video captioning**, which identifies key events in untrimmed videos and generates concise descriptions, is crucial for multimedia analysis, accessibility, security, and data management. However, existing **two-stage methods** are inefficient and rely heavily on manual annotations. This research proposes a unified system integrating **event localization** and **captioning** using pretrained **sequence-to-sequence** deep learning models. By leveraging unannotated data and simplifying the processing pipeline, the approach improves scalability and performance, offering practical solutions for real-world applications and advancing video and natural language processing.

Targets

- **Develop an integrated sequence-to-sequence model**: Build a state-of-the-art model capable of simultaneously localizing and generating detailed, temporally coherent descriptions of events in videos.
- **Leverage video data with transcripts**: Utilize natural, unannotated video data to reduce costs while effectively capturing relationships between visuals, audio, and language.
- **Evaluate fine-tuning performance**: Measure model effectiveness through log-likelihood, assessing its predictive accuracy and the quality of generated descriptions compared to real-world data.

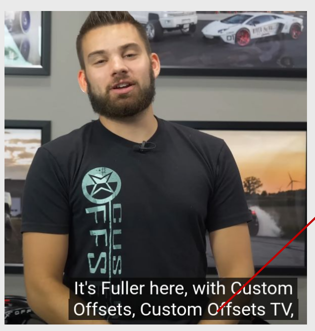
Overview



Description

1. Pre-training Dataset

- We utilized a large-scale dataset comprising narrated videos, offering an extensive collection of video-text pairs.
- This dataset encompasses diverse scenarios and natural narrations, providing valuable data for training models to produce detailed and coherent video descriptions.



caption string	timestamp_start float32	timestamp_stop float32
hey guys what's going on this fuller here with custom offsets custom offsets TV on the YouTube back at it again with another episode of from the inbox where we answer the questions that you guys sent us in the DMS or drop in the comments on YouTube and Facebook so	0	15

Figure 1 . From youtube subtitles to timestamp data

2. Pre Training Task

- **Generative objective:** Generate coherent speech outputs conditioned on visual inputs, enabling the model to learn a strong connection between visual and linguistic modalities.
- **Denoising objective:** Reconstruct masked tokens by leveraging both noisy speech and visual inputs, promoting the model's ability to reason across modalities and handle incomplete or noisy data effectively.

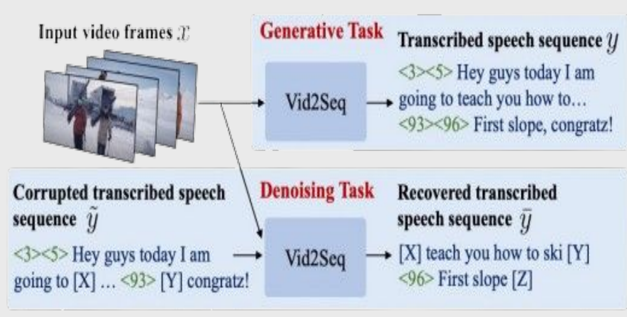


Figure 2 . Pre-training task process

3. Finetune

- Formally, given visual inputs x, encoder text sequence y and a decoder target text sequence z, both objectives are based on minimizing the following loss:

$$\mathcal{L}_\theta(x, y, z) = -\frac{1}{\sum_{k=1}^{L-1} w_k} \sum_{k=1}^{L-1} w_k \log p_\theta(z_{k+1} | x, y, z_{1:k})$$

Equation 1 . Likelihood loss function

4. Research Plan

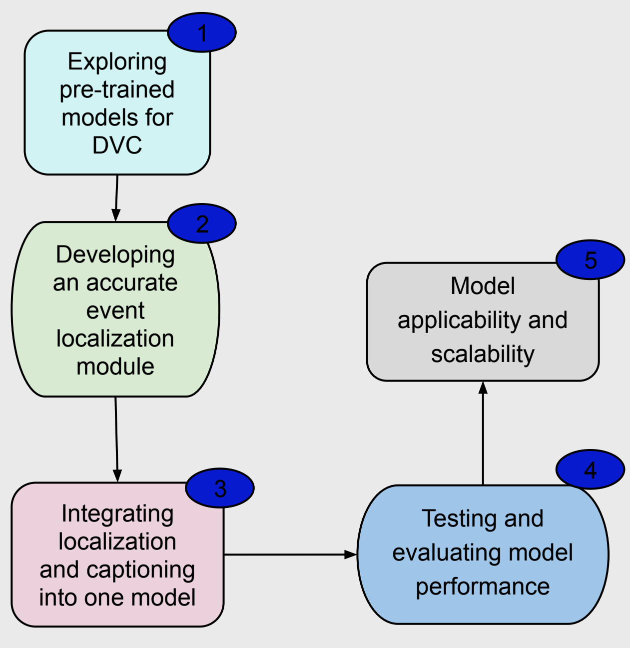


Figure 3 . Research plan diagram

NII

Huynh Ngoc Bao Long – University of Information Technology

TEL : 0946423648 Email : 22520811@gm.uit.edu.vn