FLASHFORMER: DESIGN AND EVALUATION OF A FAST AND EFFICIENT SINGLE-CHANNEL SPEECH SEPARATION MODEL

Dat Nguyen Duc[1,2]

[1] University of Information Technology, Ho Chi Minh City, Vietnam

[2] Vietnam National University, Ho Chi Minh City, Vietnam

What?

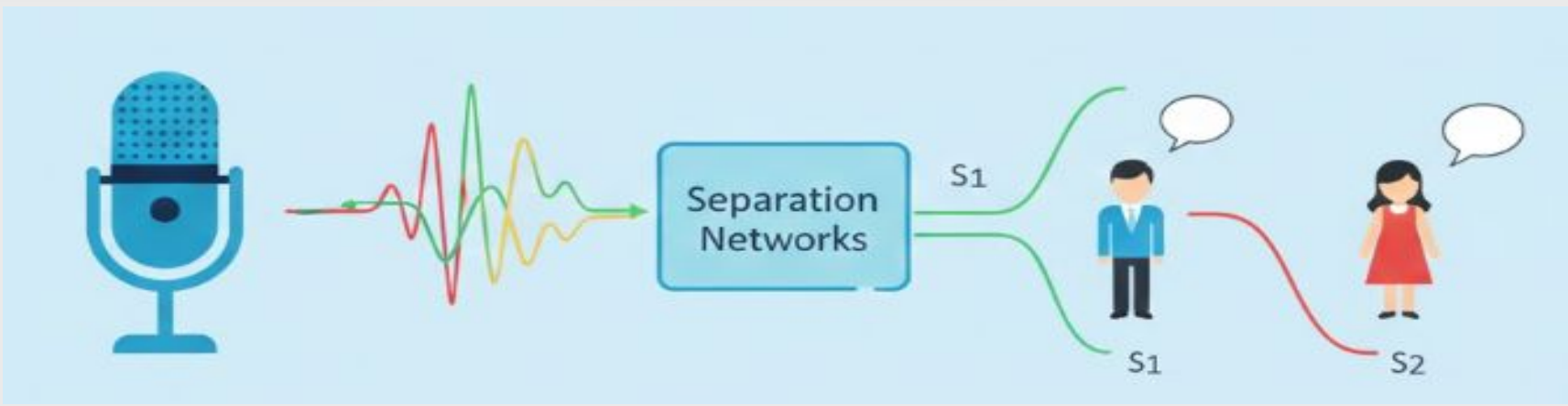
We introduce **FlashFormer**, a hybrid architecture designed to solve the single-channel speech separation problem:

- **Proposed Method:** Combines **Transformer** for global context, **FSMN** for efficient memory, and **Flash Attention** for hardware optimization.
- **Complexity Reduction:** Transitions computational complexity from quadratic to linear, significantly reducing memory footprint
- **Core Task:** Separating individual voices from a single mixed audio source

Why?

- **Computational Cost:** Current state-of-the-art models like Conv-TasNet or MossFormer2 are computationally **expensive** for long sequences.
- **Real-time Constraints:** Traditional Transformers suffer from quadratic memory growth, making them **difficult** to deploy on resource-constrained IoT devices or real-time systems
- **Optimization Gap:** There is a need for a model that balances high-quality separation (quality) with low processing latency (latency).

Overview



Description

1. Hybrid Transformer-FSMN Architecture

- The model utilizes **Transformer** layers to capture global dependencies and rich contextual representations across the audio signal
- To handle long-term dependencies with lower overhead, it integrates Feedforward Sequential Memory Networks (FSMN)
- FSMN layers expand the receptive field effectively without the high computational cost typically associated with recurrent structures.

2. Hardware-Optimized Flash Attention

- A core innovation is the integration of **Flash Attention** to optimize memory access at the hardware level
- This mechanism reduces the traditional Transformer's quadratic time and memory complexity down to **linear complexity**
- By making the attention mechanism **IO-aware**, the model achieves significantly faster inference speeds, particularly for long audio sequences.

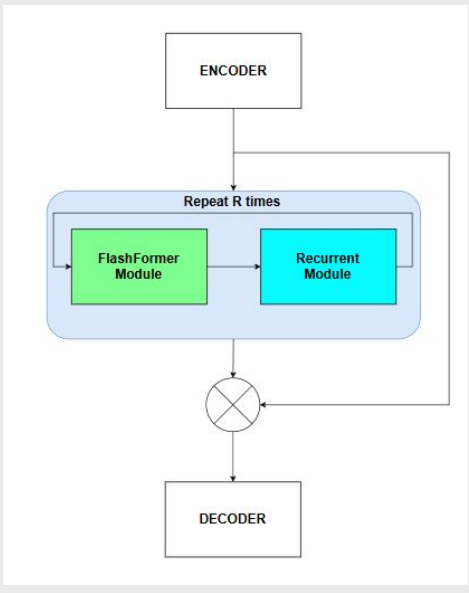


Fig. 1. The diagram of the proposed FlashFormer model.

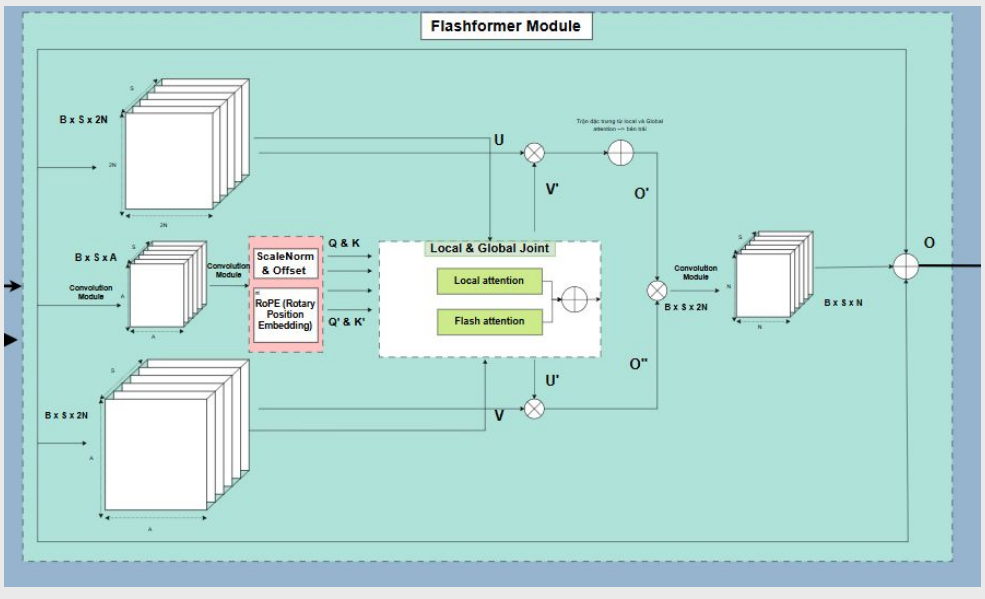


Fig. 2. The diagram of the FlashFormer Module

Dat Nguyen Duc – University of Information Technology, Ho Chi Minh City, Vietnam
TEL : 0779455498 Email : nddat1811@gmail.com