Author : Antoine BRIMONT

Title: «*A Multi-modal BLIP-2 Approach for Video Captioning*»

Short Abstract:

To address the challenge of aligning audio, visual, and textual data in video captioning, we introduce a novel multi-modal adapter inspired by BLIP-2's Q-Former that prioritizes early-stage cross-modal feature extraction. Unlike traditional methods, our approach bypasses heavy video–text pre-training by leveraging a strategic knowledge transfer from BLIP-2's image–text capabilities. By fostering fine-grained audio–visual interactions at the early layers, our model achieves state-of-the-art performance on MSR-VTT and Latest-VATEX, while remaining highly competitive on the audio-centric AudioCaps benchmark. Ultimately, this architecture proves that early fusion is key to scalable, robust, and human-aligned video understanding.

Télécom SudParis • Site d'Évry : 9 rue Charles Fourier - 91011 Évry-Courcouronnes Cedex - France • Site de Palaiseau : 19 place Marguerite Perey - 91120 Palaiseau France • Tél. +33 (0)1 60 76 40 40 • Siret : 180 092 025 00055 • NDA 11910803791 • APE : 8542Z
www.telecom-sudparis.eu