VideoBERT: A Joint Model for Video and Language Representation Learning

Literature Review

Authors:
Rasesh Srivastava (210123072)
Vidya Sagar G (210123073)

DA 421M: Multi-modal Data Processing and Learning Course Project

Instructor:
Dr. Prashant Wagambar Patil

Abstract

This report presents our study on video captioning using masking techniques inspired by transformer-based architectures. Specifically, we explore the application of masking in both text segments and video segments, drawing insights from models like MultiMAE, VideoBERT, and similar approaches. The goal of this project is to implement a novel method for video captioning that enhances the understanding of both spatial and temporal features in multimodal data.

1 Problem Statement

Video captioning is a crucial task at the intersection of computer vision and natural language processing, involving the automatic generation of descriptive text for a sequence of video frames. The primary challenge in this task lies in effectively capturing both the spatial information (features from individual video frames) and the temporal relationships (the progression of visual content over time). Existing models often struggle to balance these aspects, leading to captions that may overlook fine-grained details or fail to capture the temporal dynamics of events within the video.

2 Introduction

Video captioning, the task of generating textual descriptions from video content, has gained significant traction due to its applications in content indexing, accessibility, and video summarization. Recent models such as **VideoBERT** have demonstrated that masked token modeling—originally applied in natural language processing tasks like BERT—can effectively capture semantic information from video sequences. To address the inherent challenges in video captioning, particularly in capturing both spatial and temporal features, we propose to investigate the use of masked modeling techniques. In this project, we aim to extend this concept by applying **masking techniques** on both the visual and textual components of videos, inspired by the success of models like **MAE** in the vision domain.

3 Datasets

The following datasets have been commonly used in prior research for training and evaluating video captioning models:

- **MSR-VTT**: Consists of 10,000 video clips with over 200,000 captions. The dataset is split into 6,513 videos for training, 497 videos for validation, and 2,990 videos for testing.
- **YouCook2**: Comprises 2,000 cooking videos with instructional captions. It is divided into 1,333 videos for training, 457 videos for validation, and 210 videos for testing.
- **ActivityNet Captions**: A larger dataset with 20,000 videos featuring human activities. It includes 10,024 videos for training, 4,926 videos for validation, and 5,044 videos for testing, with dense captions provided for each event within the videos.

4 Related Works

This section will review key related works, which we will expand as the project progresses.

4.1 VideoBERT: A Joint Model for Video and Language Representation Learning

This paper by Chen Sun et al. (2019) introduces VideoBERT, a self-supervised joint model for video and language representation learning. Inspired by BERT, the model uses a bidirectional Transformer to capture high-level semantic features in both video and linguistic modalities. VideoBERT leverages large-scale unlabelled videos, applying vector quantization to video features and using ASR (Automatic Speech Recognition) for spoken words.

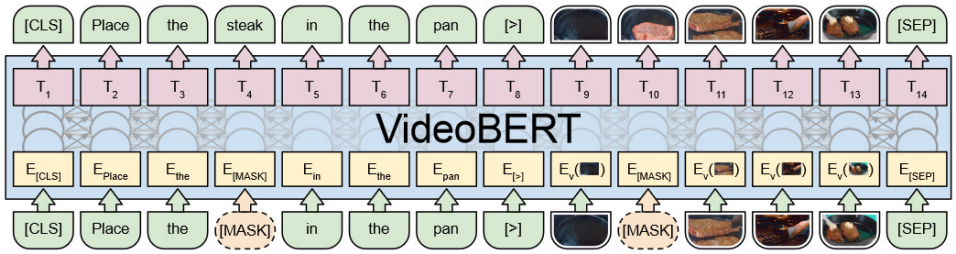

Figure 3: Illustration of VideoBERT in the context of a video and text masked token prediction, or *cloze*, task. This task also allows for training with text-only and video-only data, and VideoBERT can furthermore be trained using a linguistic-visual alignment classification objective (not shown here, see text for details).

Figure 1: VideoBERT: Text-to-video generation and future forecasting, as illustrated in the original paper.

Key Contributions

- **Joint Video-Language Representation**: VideoBERT models both video and text jointly, using BERT to learn bidirectional distributions over sequences of visual and linguistic tokens.
- **Self-Supervised Learning**: Trains without explicit labels, leveraging high-level features learned from vector quantized video tokens and ASR-derived linguistic tokens.
- **Long-Term Temporal Dynamics**: The model focuses on high-level semantic features and long-range temporal dependencies, surpassing low-level texture or motion patterns used in prior models.
- **Zero-shot Video Captioning and Classification**: Achieves state-of-the-art performance on video captioning tasks and allows zero-shot action classification.
- **Cross-Modal Pretraining**: Combines video and text inputs during pretraining, learning representations that improve performance on downstream tasks.

4.2 VisualBERT: A Simple and Performant Baseline for Vision and Language

VisualBERT, introduced by Li et al. (2019), is one of the first transformer-based models designed to handle both vision and language tasks. It extends the BERT architecture by incorporating visual information alongside textual input, allowing for multimodal representation learning. The model takes an image and a corresponding text input, such as a caption, and processes them together to create joint representations, enabling it to perform tasks like image-caption matching and visual question answering (VQA).

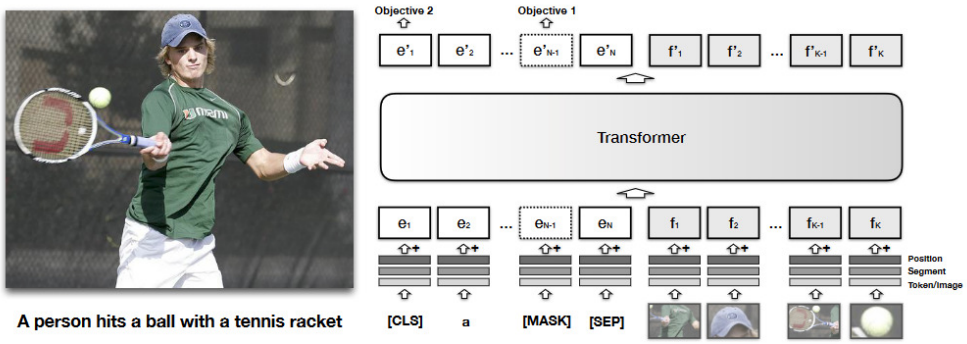

Figure 2: The architecture of VisualBERT. Image regions and language are combined with a Transformer to allow the self-attention to discover implicit alignments between language and vision. It is pre-trained with a masked language modeling (Objective 1), and sentence-image prediction task (Objective 2), on caption data and then fine-tuned for different tasks. See §3.3 for more details.

Figure 2: The pipeline of VisualBERT from the original paper.

Key Contributions

- **Multimodal Input**: VisualBERT processes both images and text inputs by embedding visual regions of interest (ROIs) from the image and token embeddings from the text, merging them within the transformer's attention layers.
- **BERT-Based Architecture**: The model leverages BERT's pre-trained weights, adapting them to vision-language tasks, making it easy to fine-tune for specific tasks such as image-captioning, VQA, and more.
- **Vision-Language Fusion**: VisualBERT effectively fuses visual and textual information, allowing for fine-grained cross-modal attention between image regions and their corresponding textual descriptions.
- **Performance on Downstream Tasks**: VisualBERT achieves strong performance on a variety of benchmarks, including image-caption matching and VQA, setting a high standard for transformer-based multimodal models.

4.3 End-to-End Dense Video Captioning with Masked Transformer

This paper by Luowei Zhou et al. (2018) introduces an end-to-end approach for dense video captioning using a Masked Transformer model. Unlike previous methods, which treat event detection and captioning separately, this model handles both tasks jointly. The architecture consists of a video encoder, a proposal decoder for event detection, and a captioning decoder that generates captions for specific events within the video.

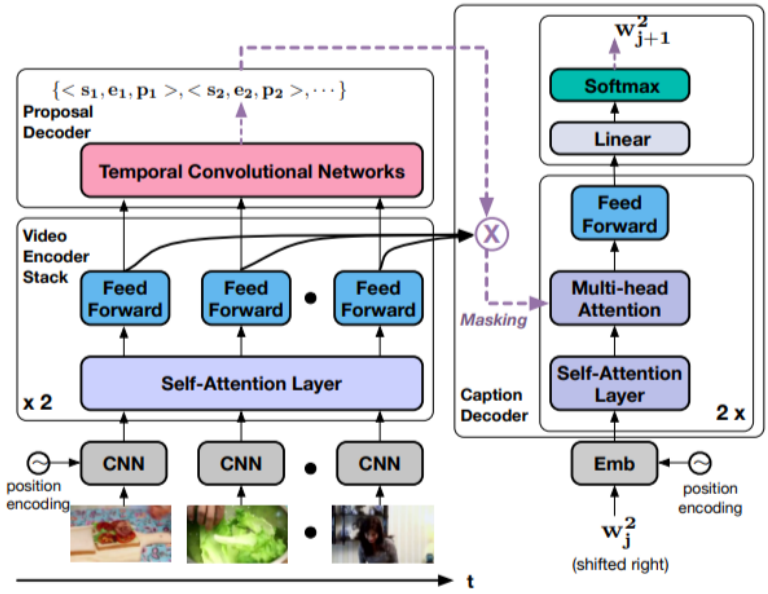

Figure 3: The pipeline of the End-to-End Dense Video Captioning with Masked Transformer from the original paper.

Key Contributions

- **Masked Transformer**: The model uses a Transformer with masking to focus on specific video segments and generate captions for individual events.
- **Event Proposal Module**: Detects temporal segments in videos and generates captions for each detected event.
- **End-to-End Learning**: The pipeline is trained end-to-end, optimizing both event detection and caption generation together.
- **Self-Attention Mechanism**: Captures both short-term and long-term dependencies within video sequences.
- **Differentiable Masking**: Ensures the captioning focuses on the relevant event, improving coherence and robustness.

4.4 UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation

UniVL was introduced by Huaishao Luo et al. (2020) to extend the transformer-based pre-training paradigm by focusing on both *multimodal understanding* and *generation tasks*. The architecture consists of two single-modal encoders (for text and video), a cross-modal encoder, and a decoder. Five pre-training objectives were introduced to enhance the model's ability to process both video and textual data jointly, including a **conditioned masked language model** and **conditioned masked frame model**.

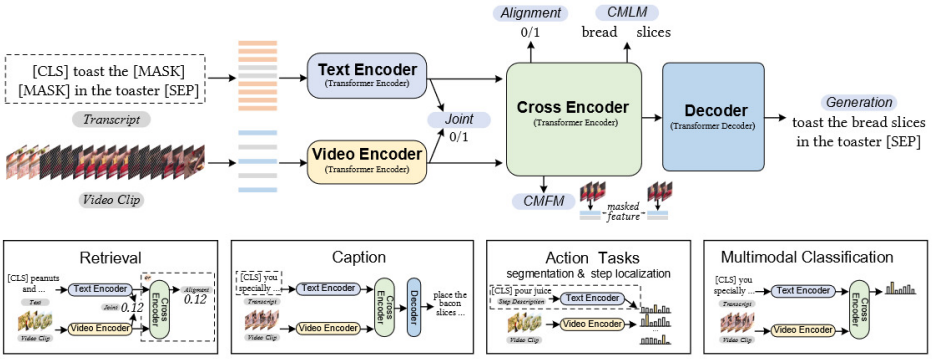

Figure 3: The main structure of our UniVL, which comprises four components, including two single-modal encoders, a cross encoder, and a decoder. The model is flexible for many text and video downstream tasks. Four possible tasks are listed.

Figure 4: The pipeline of UniVL from the original paper.

Key Contributions The key contributions of the UniVL paper are as follows:

- **Multimodal Inputs**: The model processes both video frames and their corresponding transcripts, pre-processing them into tokenized text and feature-extracted video sequences.
- **Single-Modal Encoders**: UniVL includes two separate encoders—one for text (based on BERT) and one for video (using a transformer to handle video frame sequences). These encoders learn distinct representations before fusion.
- **Cross Encoder**: A transformer-based cross encoder fuses the text and video features, generating a unified multimodal representation for further processing.
- **Decoder**: A transformer decoder is employed to handle generation tasks such as video captioning, where textual descriptions are generated sequentially based on the video and transcript inputs.
- **Pre-Training Tasks**: UniVL incorporates five pre-training objectives: conditioned masked language modeling (CMLM), conditioned masked frame modeling (CMFM), video-text alignment, and language reconstruction. These tasks help the model learn strong representations for multimodal data.

4.5 Masked Autoencoders (MAE): Scalable Vision Learners

The **Masked Autoencoders (MAE)** paper, introduced by Kaiming He et al. (2021), proposes a scalable self-supervised learning approach for computer vision. The key idea behind MAE is to mask random patches of an image and train a model to reconstruct the missing parts. This approach significantly reduces redundancy in images and creates a challenging task that encourages the model to learn more meaningful representations.

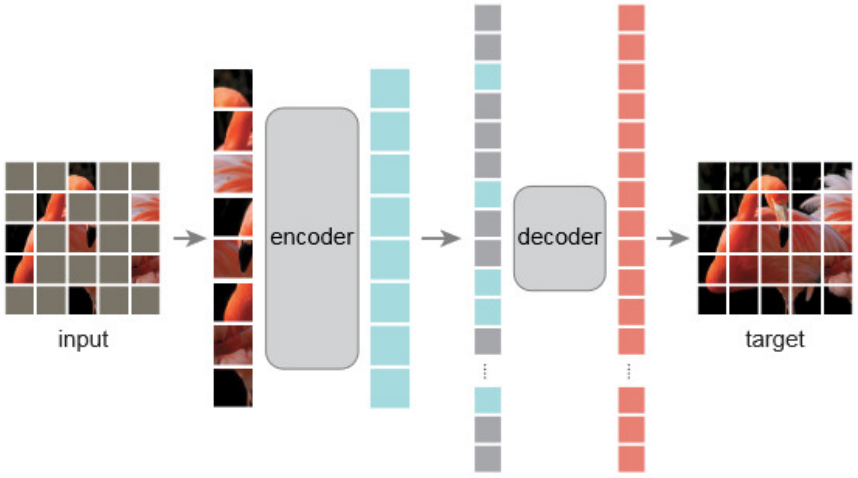

Figure 1. **Our MAE architecture**. During pre-training, a large random subset of image patches (*e.g.*, 75%) is masked out. The encoder is applied to the small subset of *visible patches*. Mask tokens are introduced *after* the encoder, and the full set of encoded patches and mask tokens is processed by a small decoder that reconstructs the original image in pixels. After pre-training, the decoder is discarded and the encoder is applied to uncorrupted images (full sets of patches) for recognition tasks.

Figure 5: The MAE pipeline from the original paper.

Key Contributions The major contributions of the MAE paper are as follows:

- **Asymmetric Encoder-Decoder Architecture**: The MAE framework uses an encoder that operates only on the visible (unmasked) patches of an image. A lightweight decoder then reconstructs the original image using both the encoded patches and mask tokens. This design reduces computation significantly by keeping the decoder smaller and focusing most of the complexity in the encoder.
- **High Masking Ratio**: MAE applies a high masking ratio (up to 75% or higher) making the reconstruction task more challenging, forcing the model to capture higher-level semantic information rather than relying on local features.

- **Pre-training and Fine-tuning Efficiency**: Reduction of pre-training time by up to 3x while improving performance. When fine-tuned on ImageNet, MAE achieves competitive accuracy (87.8%) using a ViT-Huge model, outperforming many supervised methods trained on the same data.
- **Generalization to Downstream Tasks**: The representations learned by MAE during pre-training transfer well to other tasks, such as object detection and semantic segmentation, showing better performance than models that use supervised pre-training.

5 References

- **Main Research Paper:** VideoBERT: A Joint Model for Video and Language Representation Learning
- **Related works' Research Papers:**
  1. VisualBERT: A Simple and Performant Baseline for Vision and Language
  2. End-to-End Dense Video Captioning with Masked Transformer
  3. UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation
  4. Masked Autoencoders Are Scalable Vision Learners