Video Captioning with Transformer-Based Models: An Exploration from VideoBERT to UniVL

Final Report

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 project investigates transformer-based architectures for video captioning, a task demanding precise capture of both spatial details and temporal dynamics in multimodal data. Leveraging recent advancements in self-supervised multimodal transformers, our approach centers on implementing and training VideoBERT for video captioning, followed by modifying VideoBERT to UniVL. The goal is to evaluate how each model independently processes video-language representations, examining metrics such as caption coherence, temporal awareness, and richness of generated content. Through this study, we aim to understand the impact of evolving multimodal transformers on video captioning quality, providing insights into their potential and limitations within real-world applications like content indexing and accessibility.

1 Problem Statement

Video captioning integrates computer vision and natural language processing to generate temporally coherent descriptions of video sequences. This task requires capturing spatial details within frames as well as temporal progression across frames—a dual challenge that current models often struggle to balance. While transformer architectures, particularly with self-supervised learning, show promise, they frequently fall short in aligning spatial understanding with long-term temporal dynamics. This project addresses this gap by implementing VideoBERT, an early video-language transformer, and comparing it with UniVL, a model that builds on VideoBERT using advanced multimodal objectives. This comparison aims to assess improvements in captioning performance and contribute insights into transformer effectiveness for multimodal tasks.

2 Introduction

Video captioning, the task of generating accurate textual descriptions from video content, is a valuable tool in many real-world applications, such as content indexing, accessibility enhancement, and video summarization. The need for such systems has grown with the expansion of video data, but traditional models have struggled with the inherent challenges of video captioning—specifically, the effective capture of spatial detail in individual frames and temporal relationships across sequences. Modern transformer architectures, inspired by the success of masked modeling techniques in language tasks (e.g., BERT), have emerged as strong contenders in addressing these challenges for multimodal data.

3 Literature Review

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

This paper by Chen Sun et al. (2019) introduces VideoBERT, a self-supervised model designed for video and language representation learning. Inspired by BERT's masked token

approach, VideoBERT uses a bidirectional Transformer to capture high-level semantic features across video and language. By leveraging unlabelled video data, the model applies vector quantization to video features and uses Automatic Speech Recognition (ASR) to incorporate spoken language, enabling multimodal representation learning.

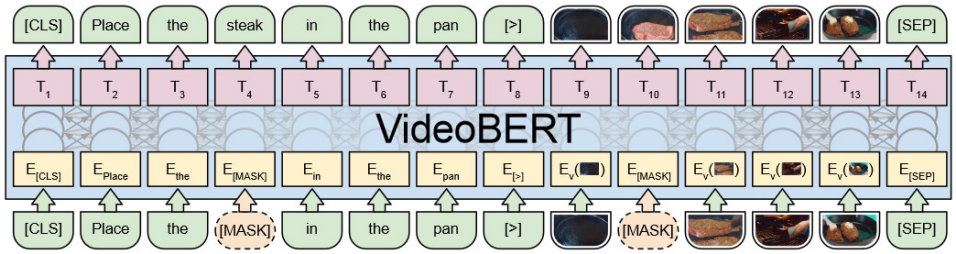

Figure 3: Illustration of VideoBERT in the context of a video and text masked token prediction, or *cloze*, task. This task also allows for training with text-only and video-only data, and VideoBERT can furthermore be trained using a linguistic-visual alignment classification objective (not shown here, 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**: Models both video and text jointly, with BERT learning bidirectional distributions over visual and linguistic tokens.
- **Self-Supervised Learning**: Operates without explicit labels, leveraging vector quantized video tokens and ASR-derived text tokens.
- **Long-Term Temporal Dynamics**: Focuses on high-level semantic and temporal relationships over low-level visual patterns.
- **Zero-shot Video Captioning and Classification**: Demonstrates state-of-the-art performance on video captioning and supports zero-shot action classification.
- **Cross-Modal Pretraining**: Integrates video and text inputs in pretraining, enhancing downstream task performance.

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

VisualBERT, introduced by Li et al. (2019), extends the BERT architecture for vision-language tasks by processing both image and text inputs simultaneously. Although designed primarily for image-language tasks like Visual Question Answering (VQA) and image-caption matching, VisualBERT contributed to multimodal learning by showing how BERT could fuse visual and textual data within a transformer.

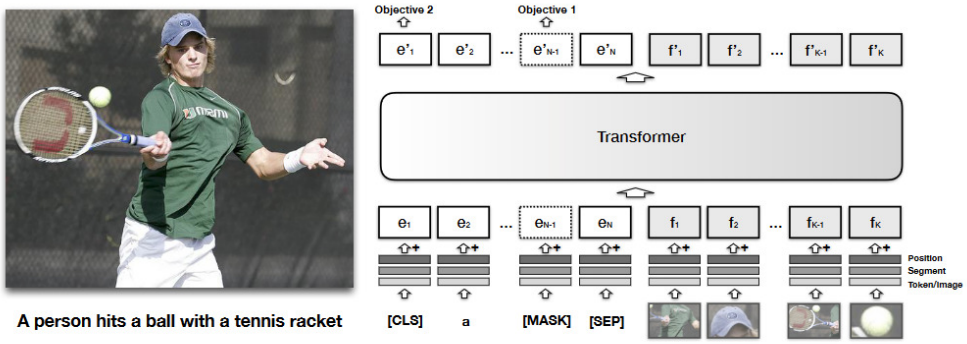

Figure 2: The architecture of VisualBERT. Image regions and language are combined with a Transformer to allow the self-attention to discover implicit alignments between language and vision. It is pre-trained with a masked language modeling (Objective 1), and sentence-image prediction task (Objective 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**: Incorporates both images and text by embedding regions of interest (ROIs) from images and text tokens.
- **BERT-Based Architecture**: Adapts pre-trained BERT weights for vision-language tasks, allowing easy fine-tuning for VQA and image-captioning.
- **Vision-Language Fusion**: Achieves fine-grained cross-modal attention between visual and textual elements.
- **Strong Performance**: Sets a high standard for transformer-based multimodal models on benchmarks like VQA and image-caption matching.

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

Luowei Zhou et al. (2018) introduced a model for dense video captioning using a masked transformer, enabling the model to perform both event detection and caption generation within a single framework. This approach marked a shift towards handling temporal segmentation and textual description jointly, leveraging self-attention to capture dependencies within video sequences.

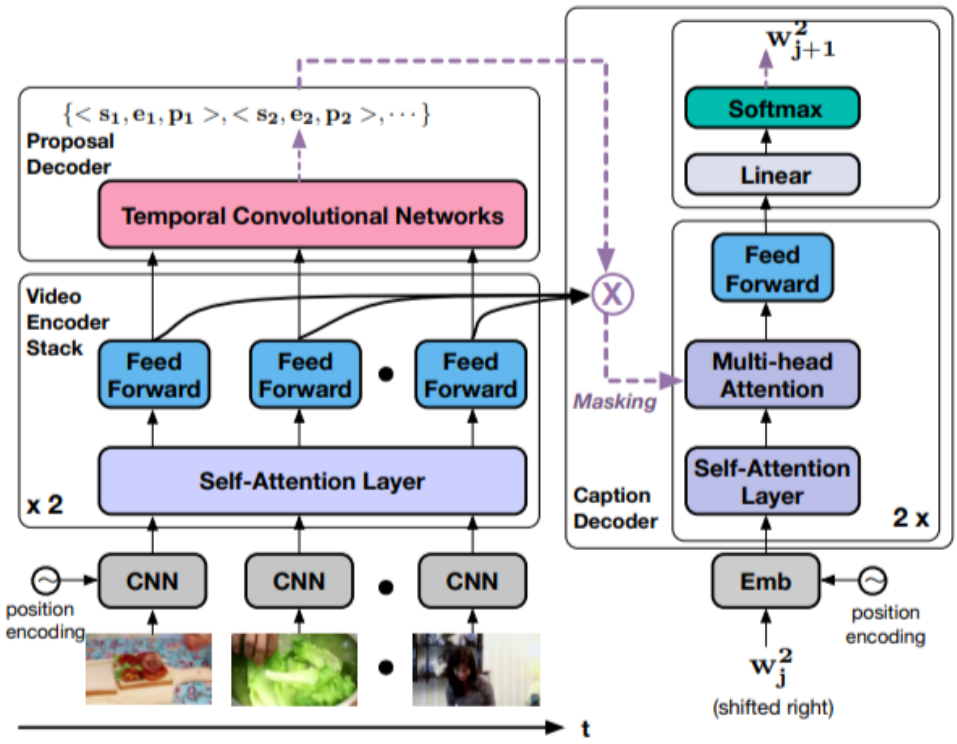

Figure 3: Pipeline of End-to-End Dense Video Captioning with Masked Transformer.

Key Contributions

- **Masked Transformer**: Uses masking to focus on specific video segments, producing captions for discrete events.
- **Event Proposal Module**: Detects temporal segments in videos and generates captions for each detected event.
- **End-to-End Learning**: Combines event detection and caption generation within a single, optimized pipeline.
- **Self-Attention Mechanism**: Captures both short-term and long-term dependencies within video sequences.
- **Differentiable Masking**: Ensures that captions are event-specific, enhancing caption coherence and relevance.

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

Huaishao Luo et al. (2020) introduced UniVL to enhance transformer-based multimodal pretraining, focusing on tasks requiring both understanding and generation in video-language contexts. UniVL employs distinct encoders for text and video, a cross-modal encoder, and a

decoder for generation tasks, pretraining on multiple objectives to strengthen its multimodal capabilities.

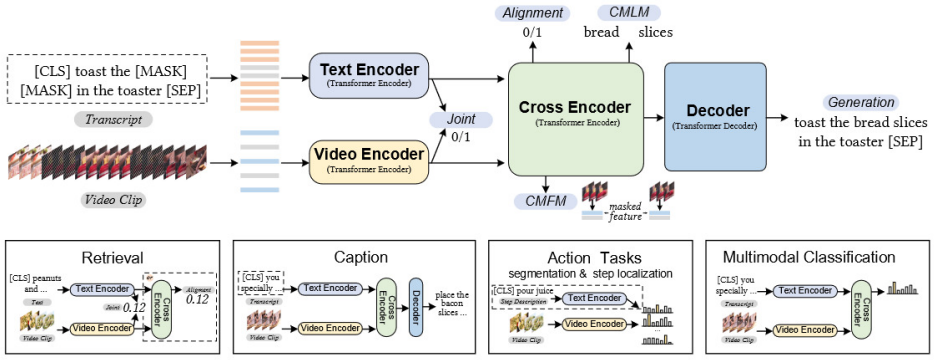

Figure 3: The main structure of our UniVL, which comprises four components, including two single-modal encoders, a cross encoder, and a decoder. The model is flexible for many text and video downstream tasks. Four possible tasks are listed.

Figure 4: The UniVL pipeline as presented in the original paper.

Key Contributions

- **Multimodal Inputs**: Processes both video frames and associated transcripts, tokenizing text and feature-extracting video frames.
- **Separate Encoders**: Uses distinct encoders for text (based on BERT) and video, enhancing specialized representations before fusion.
- **Cross Encoder**: Fuses text and video features, producing unified multimodal representations.
- **Decoder for Generation**: Supports generation tasks like video captioning, using sequential inputs to produce captions based on multimodal content.
- **Multi-Objective Pretraining**: Trains on multiple objectives, including masked language and frame modeling, aligning video-text pairs for enhanced multimodal representation.

3.5 Masked Autoencoders (MAE): Scalable Vision Learners

Masked Autoencoders (MAE) by Kaiming He et al. (2021) introduced a high masking ratio to focus the model on reconstructing missing portions of images. While primarily a vision model, its masking techniques indirectly inspired similar approaches in video-language transformers, emphasizing high-level feature learning over local patterns.

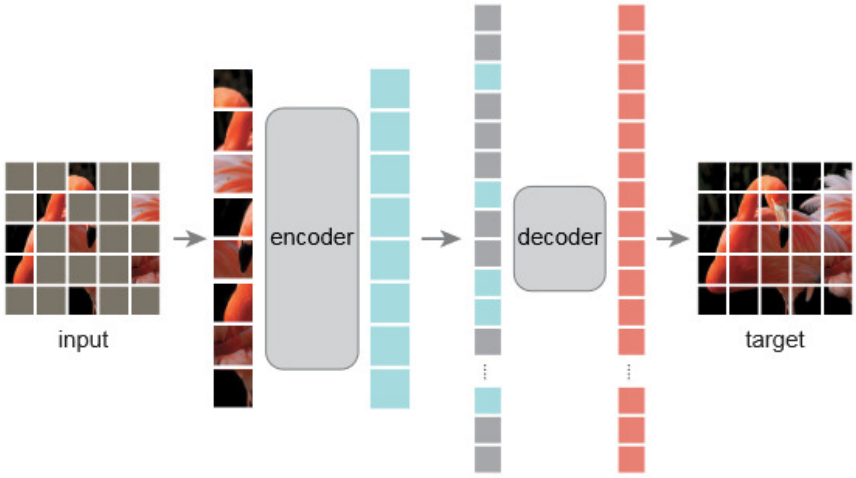

Figure 1. **Our MAE architecture**. During pre-training, a large random subset of image patches (*e.g.*, 75%) is masked out. The encoder is applied to the small subset of *visible patches*. Mask tokens are introduced *after* the encoder, and the full set of encoded patches and mask tokens is processed 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

- **Asymmetric Encoder-Decoder Architecture**: Uses an encoder for visible image patches and a lightweight decoder for reconstruction.
- **High Masking Ratio**: Applies up to 75% masking, challenging the model to capture high-level semantics over local features.
- **Efficient Pretraining and Fine-Tuning**: Reduces pretraining time while achieving competitive performance.
- **Improved Generalization**: Transferrable representations suited for tasks like object detection and semantic segmentation.

4 Additional Papers

4.1 CLIP: Contrastive Language–Image Pretraining

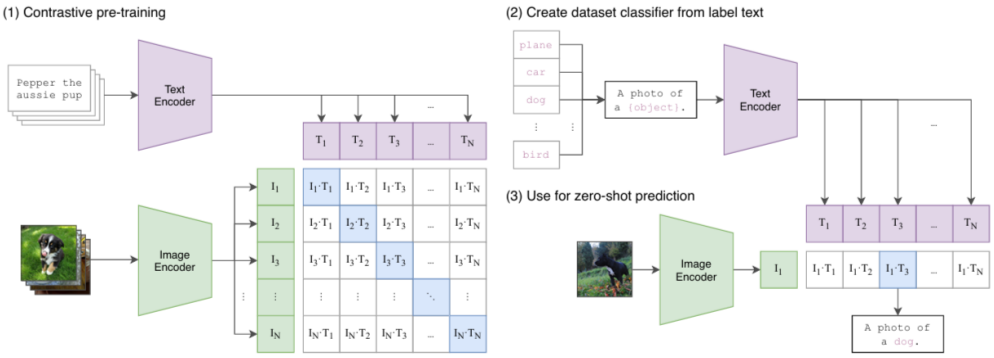

Figure 6: The CLIP pipeline from the original paper.

CLIP (Contrastive Language–Image Pretraining) uses a two-stream pipeline with independent image and text encoders trained to map images and language to a shared embedding space. The model employs contrastive learning to train the encoders, maximizing similarity between matched image-text pairs (e.g., a specific image and its caption) while minimizing similarity for unmatched pairs. By aligning image and text embeddings in this way, CLIP creates a powerful multimodal representation where images and descriptions of similar concepts are located close together in the embedding space. This structure enables the model to perform a range of zero-shot vision tasks, as it can leverage its generalizable, high-quality representations learned during pretraining.

4.2 MViT: Multiscale Vision Transformers

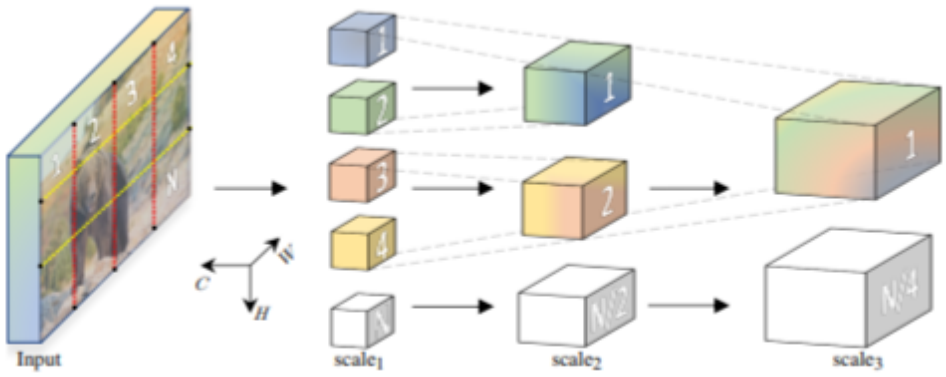

Figure 7: The MViT implementation from the original paper.

MViT (Multiscale Vision Transformers) introduces a pipeline that captures video context at multiple scales using adaptive attention. Unlike standard transformers that apply a single

resolution across the entire input, MViT dynamically adjusts the resolution at different layers, starting with lower resolutions for broader context and increasing detail as processing progresses. This approach enables the model to balance fine-grained spatial details with larger contextual understanding, making it especially suitable for video tasks where objects and scenes vary in scale. The multiscale pipeline effectively captures temporal dependencies by processing video sequences with this multi-resolution attention, improving performance in video understanding tasks.

4.3 ALBEF: Align Before Fuse

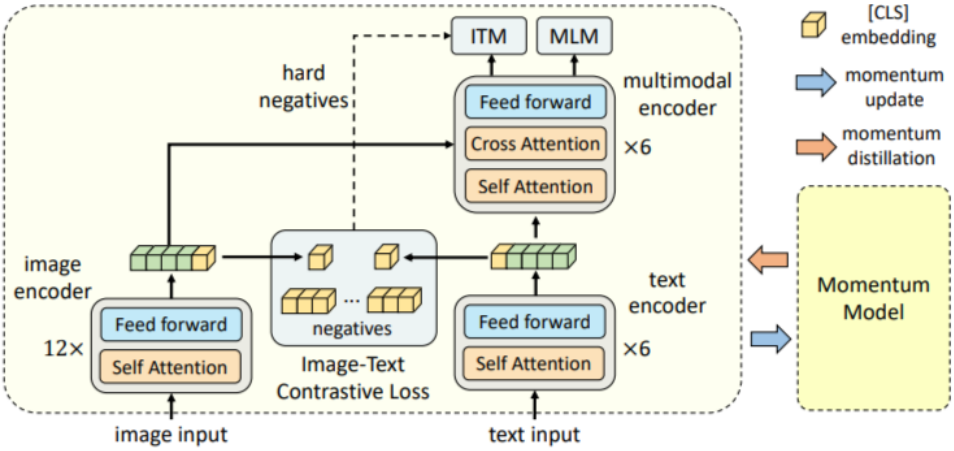

Figure 8: The ALBEF pipeline from the original paper.

In ALBEF (Align Before Fuse), the pipeline consists of an initial alignment phase followed by a fusion stage for multimodal inputs. First, separate encoders process image and text inputs independently, aligning them in a shared representation space through contrastive learning. This alignment stage ensures that both image and text features are positioned close in the embedding space if they are semantically related. Once aligned, the image and text features are then passed through a cross-modal attention layer that fuses the modalities for downstream tasks like image-caption matching or Visual Question Answering (VQA). This "align-before-fuse" design enhances model performance by enforcing feature consistency before combining modalities, allowing for more accurate multimodal grounding.

5 Proposed Approach

In our project, we implemented the VideoBERT model as the foundational transformer architecture for video captioning tasks. This implementation was followed by a modification using UniVL, an evolved model that builds on VideoBERT's framework to enhance multimodal understanding and caption generation. Below, we detail the pipeline steps used in our implementation of VideoBERT, outlining the preprocessing, feature extraction, and training processes that support video-language representation learning.

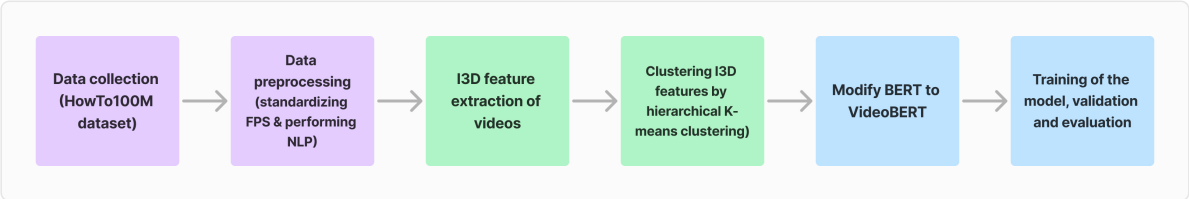

Figure 9: Proposed pipeline of training VideoBERT

5.1 VideoBERT Pipeline

The VideoBERT pipeline comprises several key stages, each essential for transforming raw video data and textual annotations into a format suitable for video-language pretraining. Figure 9 gives a brief of the pipeline we followed.

- **Step 1: Collection of Training Data**
  Video and text annotations are sourced from the HowTo100M dataset, a large-scale dataset for video-language tasks. We use a subset of 10,000 videos, with high-quality, aligned text annotations obtained from the official dataset source.

- **Step 2: Transforming the Data**
  Video frame rates are standardized to 10 frames per second (fps) for consistency and computational efficiency. Text annotations are processed to restore punctuation, ensuring language coherence.

- **Step 3: Extraction of I3D Features**
  I3D (Inflated 3D) network features are extracted from the processed frames to encode motion and object-level details, leveraging the I3D model's strength in capturing spatiotemporal information.

- **Step 4: Clustering the I3D Features**
  I3D features are grouped using hierarchical k-means clustering, with a cluster size $k = 12$ and hierarchy level $h = 4$. This clustering reduces feature dimensionality while retaining essential information.

- **Step 5: Converting BERT to VideoBERT**
  Modifications are made to BERT, including a new visual vocabulary and a custom class `VideoBertForPreTraining`, enabling joint text and video input for video-language tasks.

- **Step 6: Training the Model**
  VideoBERT is trained on the processed data, with an alignment task added to help the model learn meaningful associations between visual and textual components, enhancing cross-modal representation.

- **Step 7: Evaluation**
  We evaluate the trained model on the YouCookII validation dataset, assessing VideoBERT's zero-shot classification and captioning performance against ground truth annotations.

5.1.1 Quantitative Results:

Model	Top-1% Verb	Top-5 % Verb	Top-1% Noun	Top-5% Noun
VideoBERT (text only)	0.3%	9.7%	0.5%	11.2%
VideoBERT (text + video)	0.7%	13.8%	2.1%	16.2%

Table 1: Performance comparison of different VideoBERT models

5.1.2 Qualitative Results:

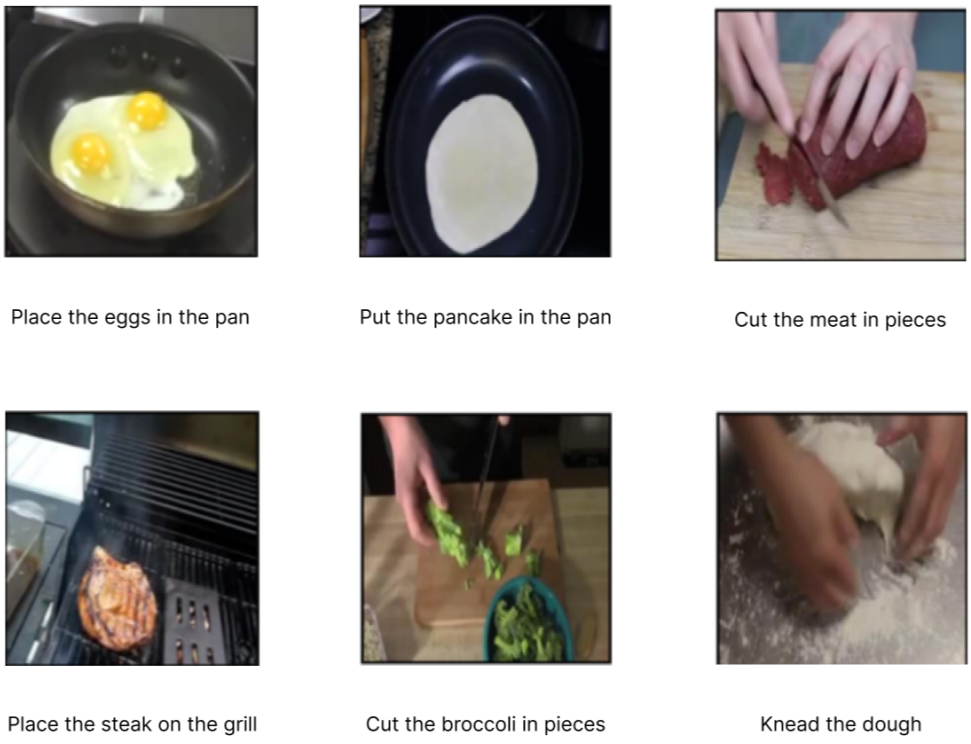

Figure 10: Examples of generated captions

Figure 11: Using VideoBERT to predict nouns and verbs given a video clip

5.2 Modification with UniVL

Following the initial VideoBERT implementation, we introduced UniVL as a modification to extend VideoBERT's capabilities. UniVL builds upon the VideoBERT framework with distinct encoders for video and language, a cross-modal encoder to improve video-language fusion, and an enhanced pretraining strategy. By implementing UniVL, we aim to assess the impact of these architectural changes on video captioning performance, particularly in terms of improved temporal awareness and finer-grained content representation in generated captions.

5.3 UniVL Pipeline

UniVL builds on VideoBERT with an advanced architecture that enables deeper multimodal understanding and generation. Unlike VideoBERT, which uses a single encoder for text and video, UniVL employs separate encoders for each modality, a cross-modal encoder for integration, and a dedicated decoder for sequence generation. Key stages of the UniVL pipeline are:

- **Step 1: Dataset Preparation**
  UniVL is fine-tuned on the YouCookII dataset for video captioning, supporting both video-only and text-video inputs for flexibility across multimodal data types.

- **Step 2: Feature Extraction**
  S3D (Shuffled and Separable 3D) network extracts 1024-dimensional feature vectors per second of video, providing robust spatiotemporal information compared to VideoBERT's I3D network.

- **Step 3: Independent Text and Video Encoding**
  UniVL uses a separate text encoder (BERT) and a video encoder for S3D features, preserving unique structural properties of each modality. VideoBERT, by contrast, combines video and text tokens in a single encoder.

- **Step 4: Cross-Modal Fusion**
  The cross-modal encoder integrates text and video features, explicitly modeling their interactions. This approach enables UniVL to create a unified, contextually rich representation, surpassing VideoBERT's simple concatenation method.

- **Step 5: Decoder for Caption Generation**
  UniVL's decoder generates captions by predicting tokens sequentially from fused multimodal features. For video-only captioning, it relies solely on visual context. VideoBERT lacks this dedicated generation mechanism, making UniVL better suited for descriptive tasks.

- **Step 6: Fine-Tuning on YouCookII**
  Fine-tuning uses pretrained weights (`univl.pretrained.bin`) and optimizes UniVL on caption quality metrics like BLEU and METEOR, extending beyond VideoBERT's focus on representation learning.

- **Step 7: Evaluation**
  Evaluated on YouCookII's validation set with BLEU, METEOR, ROUGE_L, and CIDEr metrics, UniVL demonstrates enhanced retrieval and generation capabilities, showing advantages over VideoBERT's representation-centered design.

5.4 Comparative Result Analysis

Model	BLEU - 3	BLEU - 4
VideoBERT (our implementation)	6.98%	5.23%
UniVL (our implementation)	21.76%	16.1%

Table 2: Performance comparison of VideoBERT and UniVL models

6 Code Availability

The Python code that supports this report is publicly available at the following GitHub repository: github.com/Rasesh-Srivastava/Multi-modal-Data-Processing-and-Learning-Course-Project

7 Acknowledgement

We extend our deepest gratitude to our project supervisor, Dr. Prashant Wagambar Patil, whose unwavering support, motivation, and insightful guidance were instrumental in the successful completion of this project. This accomplishment would not have been possible without his expertise and encouragement.

8 References

- **Main Research Paper:** VideoBERT: A Joint Model for Video and Language Representation Learning
- **Related works' Research Papers:**
    1. UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation
    2. VisualBERT: A Simple and Performant Baseline for Vision and Language
    3. End-to-End Dense Video Captioning with Masked Transformer
    4. Masked Autoencoders Are Scalable Vision Learners
    5. Learning Transferable Visual Models From Natural Language Supervision
    6. Multiscale Vision Transformers
    7. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation