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

Huaishao Luo[1]*, Lei Ji[2,3,4], Botian Shi[5], Haoyang Huang[2], Nan Duan[2], Tianrui Li[1], Jason Li[6], Taroon Bharti[6], Ming Zhou[2]
[1]Southwest Jiaotong University, Chengdu, China
[2]Microsoft Research Asia, Beijing, China
[3]Institute of Computing Technology, Chinese Academy of Science, Beijing, China
[4]University of Chinese Academy of Sciences, Beijing, China
[5]Beijing Institute of Technology, Beijing, China
[6]Microsoft STCA, Beijing, China
huaishaoluo@gmail.com, leiji@microsoft.com

Abstract

With the recent success of the pre-training technique for NLP and image-linguistic tasks, some video-linguistic pre-training works are gradually developed to improve video-text related downstream tasks. However, most of the existing multimodal models are pre-trained for understanding tasks, leading to a pretrain-finetune discrepancy for generation tasks. This paper proposes UniVL: a **Uni**fied **V**ideo and **L**anguage pre-training model for both multimodal understanding and generation. It comprises four components, including two single-modal encoders, a cross encoder, and a decoder with the Transformer backbone. Five objectives, including video-text joint, conditioned masked language model (CMLM), conditioned masked frame model (CMFM), video-text alignment, and language reconstruction, are designed to train each of the components. We further develop two pre-training strategies, stage by stage pre-training (StagedP) and enhanced video representation (EnhancedV), to make the training process of the UniVL more effective. The pre-train is carried out on a sizeable instructional video dataset HowTo100M. Experimental results demonstrate that the UniVL can learn strong video-text representation and achieves state-of-the-art results on five downstream tasks.

1 Introduction

With the recent advances of self-supervised learning, pre-training techniques play a vital role in learning visual and language representation. The paradigm is to pre-train the model on a large scale *unlabeled* data and fine-tune the downstream tasks using task-specific *labeled* data. Inspired by the BERT (Devlin et al., 2019) model's success for NLP tasks, numerous multimodal image-language pre-training models (Lu et al., 2019; Li et al., 2019a,b) have been proposed. Their results have demonstrated the effectiveness of pre-training on various visual and language tasks such as visual question answering. Different from previous text pre-training or image-language pre-training, we focus on video-linguistic pre-training in this paper.

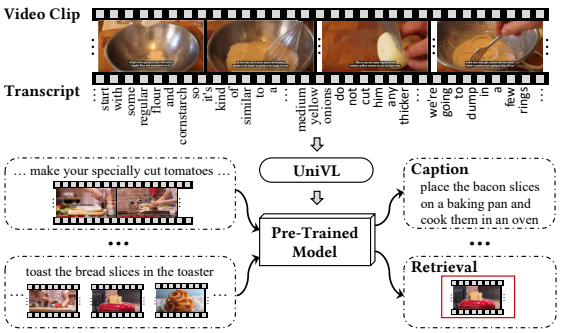


Figure 1: A showcase of video and language pre-train based model for multimodal understanding (e.g., retrieval) and generation (e.g., captioning).

Videos contain rich visual, acoustic, and language information for people to acquire knowledge or learn how to perform a task. This motivates researchers to investigate whether AI agents can learn task completion from videos like humans with both low-level visual and high-level semantic language signals. Therefore, multimodal video-language tasks are of great importance to investigate for both research and applications. In this work, we first propose to pre-train a unified video-language model using video and acoustic speech

*This work was done during the first author's internship in MSR Asia

recognition (ASR) transcript in instructional videos to learn a joint representation of both video and language. Then, we fine-tune this model on five typical multimodal tasks, including understanding and generation targets. Figure 1 presents a showcase of our pre-training and fine-tuning flow. Take multimodal video captioning as an example. The model inputs video and ASR transcript and predicts a captioning sentence.

VideoBERT (Sun et al., 2019b) and CBT (Sun et al., 2019a) are the first pioneers to investigate video-language pre-training with regard to video representation on instructional videos. They have demonstrated the effectiveness of the BERT based model for capturing video temporal and language sequential features. Besides the above two works, there is some concurrent progress to our model. ActBERT (Zhu and Yang, 2020) leverages global action information to catalyze mutual interactions between linguistic texts and local regional objects. Moreover, a transformer block is introduced to encode global actions, local regional objects, and linguistic descriptions. HERO (Li et al., 2020) hierarchically encodes multimodal inputs. Furthermore, two new pre-training tasks, video-subtitle matching and frame order modeling, are designed to improve the representation learning. VideoAsMT (Korbar et al., 2020) takes a generative modeling approach that poses the objective as a translation problem between modalities.

However, most of previous models only pre-train the model on understanding tasks. In this paper, we pre-train on both understanding and generation tasks through an encoder-decoder paradigm. Although the concurrent work VideoAsMT has a similar encoder-decoder as ours, it is not flexible for downstream tasks with only one single unified framework. In this paper, we develop a flexible approach to learn video and language joint representation and adapt downstream multimodal tasks.

We propose UniVL: a **Uni**fied **V**ideo and **L**anguage pre-training model for multimodal understanding and generation. Our UniVL model adopts Transformer (Vaswani et al., 2017) as the backbone and has four components, including two single-modal encoders, a cross encoder, and a decoder. In detail, we first encode the text and visual separately by two single-modal encoders. A video-text joint objective performs on these two encoders, which aims to learn better representation for each modality before fusing them. Such a two-stream design is natural to retrieval tasks due to its scalability to very large datasets. The proposed representation can be indexed and has linear complexity in the number of videos. Then we adopt the Transformer based encoder-decoder model to perform the understanding and generation pre-training by four tasks: conditioned masked language model (CMLM for language corruption), conditioned masked frame model (CMFM for video corruption), video-text alignment, and language reconstruction.

Furthermore, we design two pre-training strategies, including stage by stage pre-training strategy (StagedP) and Enhanced video representation (EnhancedV), to promote the UniVL pre-training. The StagedP has two parts in our setting. We only pre-train the text encoder and video encoder by the video-text joint objective for the first stage. Then all modules will be pre-trained under the whole objectives in the second stage. Besides, we adopt an entire masking strategy EnhancedV on text to enhance video representation.

Our contributions are summarized as follows:

1) We propose a multimodal video-language pre-training model trained on a large-scale instructional video dataset. It is a flexible model for both video-language understanding and generation tasks.

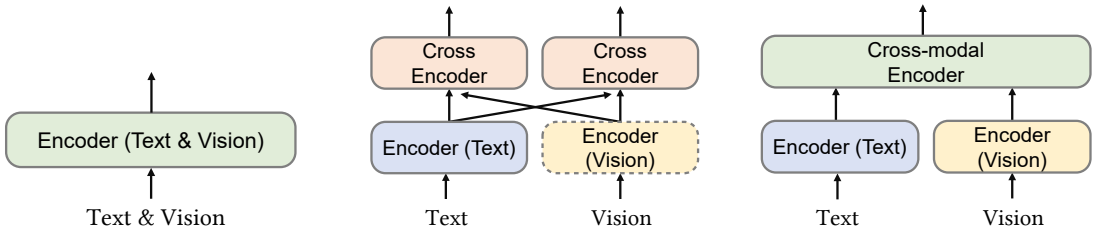
2) The pre-training consists of five objectives, including video-text joint, conditioned masked language model, conditioned masked frame model, video-text alignment, and language reconstruction. Two pre-training strategies are proposed to make these objectives work harmoniously.

3) We fine-tune our pre-trained model on five typical multimodal video-language tasks: text-based video retrieval, multimodal video captioning, action segmentation, action step localization, and multimodal sentiment analysis. Extensive experiments demonstrate our model's effectiveness on downstream tasks and achieve state-of-the-art results.

2 Related Works

2.1 Single Modal Pre-Training

Self-supervised representation learning has been shown to be effective for sequential data, including language and video. Language pre-training models, including BERT (Devlin et al., 2019), GPT (Radford et al., 2018), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), MASS (Song et al., 2019), UniLM (Dong et al., 2019), and BART (Lewis et al., 2019), have achieved great success



(a) Share Type (b) Cross Type (c) Joint Type

Figure 2: Various paradigms for multimodal pre-training.

on NLP tasks. BERT (Devlin et al., 2019) is a denoising auto-encoder network using Transformer with MLM (masked language model) and NSP (next sentence prediction) as pre-training tasks. It has a strong performance for understanding tasks. MASS (Song et al., 2019) focuses on pre-training for generation tasks. UniLM (Dong et al., 2019) and BART (Lewis et al., 2019) continuously study a unified pre-training model for both understanding and generation tasks.

Video representation learning mostly focuses on the video sequence reconstruction or future frames prediction as pre-training (pretext) tasks. Early works like (Mathieu et al., 2015; Srivastava et al., 2015; Han et al., 2019) aim to synthetic video frames through the image patches. Similarly, Wang and Gupta (2015) adopt a siamese-triplet network to rank continuous patches more similar than patches of different videos. Other works predict the feature vectors in latent space using auto-regressive models with the noise-contrastive estimation (NCE) (Lotter et al., 2016; Oord et al., 2018). Sun et al. (2019a) adopt NCE to predict corrupted (masked) latent space using the auto-encoder model.

2.2 Multimodal Pre-Training

Recently, numerous visual-linguistic pre-training models are proposed for multimodal tasks. For image and text pre-training, ViLBERT (Lu et al., 2019), LXMERT (Tan and Bansal, 2019) adopt two separate Transformers for image and text encoding independently. Other models like Visualbert (Li et al., 2019b), Unicoder-VL (Li et al., 2019a), VL-BERT (Su et al., 2020), UNITER (Zhou et al., 2019) use one shared BERT model. These models employ MLM and image-text matching as pre-training tasks which are effective for downstream multimodal tasks. VLP (Zhou et al., 2019) proposes a unified image-language model for understanding and generation tasks. Different from these works, we focus on video and text pre-training for universal representation.

For video and text pre-training, VideoBERT (Sun et al., 2019b) and CBT (Sun et al., 2019a) are the first works to explore the capability of pre-training models. Although VideoBERT and CBT pre-train the model on multimodal data, the downstream tasks mainly take video representation for further prediction. ActBERT (Zhu and Yang, 2020) leverages global action information to catalyze mutual interactions between linguistic texts and local regional objects, and introduces a transformer block to encode global actions, local regional objects, and linguistic descriptions. HERO (Li et al., 2020) encodes multimodal inputs in a hierarchical fashion. Besides, two new pre-training tasks, video-subtitle matching and frame order modeling, are designed to improve the representation learning. However, ActBERT and HERO are only pre-train the models on understanding tasks. VideoAsMT (Korbar et al., 2020) takes a generative modeling approach that poses the objective as a translation problem between modalities. The difference between our work with VideoAsMT is that our model contains two more separate encoders instead of one unified encoder-decoder, while VideoAsMT is inflexible for downstream tasks due to one single unified framework.

We summarize three pre-training paradigms to cover the previous vision-text pre-training model considering different encoder architecture in literature, as presented in Figure 2. Unicoder-VL (Li et al., 2019a), VL-BERT (Su et al., 2020), UNITER (Zhou et al., 2019), VLP (Zhou et al., 2019), VideoBERT (Sun et al., 2019b), ActBERT (Zhu and Yang, 2020), and VideoAsMT (Korbar et al., 2020) belong to share-type in Figure 2(a), where the text and vision sequences are combined

as the input of one shared Transformer encoder. ViLBERT (Lu et al., 2019) and LXMERT (Tan and Bansal, 2019) are cross-type shown in Figure 2(b). CBT (Sun et al., 2019a) and HERO (Li et al., 2020) are joint-type shown in Figure 2(c). The cross-type and joint-type architectures have two-stream input, and the difference is the interaction across both modalities. Compared with the single-stream input in the share-type, the two-stream input can accommodate each modality's different processing needs and interact at varying representation depths (Lu et al., 2019). Besides, the joint-type structure has one cross-modal encoder for full interaction between the two streams comparing with the cross-type. We adopt the joint-type as our encoder in this paper.

3 Method

The problem is defined as: given the input video and the corresponding ASR transcript pairs, pre-train a model to learn joint video and text representation with the self-supervision approach, and fine-tune downstream tasks. In this section, we describe the architecture and pre-training tasks in detail.

3.1 Model Architecture

Figure 3 presents the UniVL as an encoder-decoder architecture. First, the model extracts representations of the input text tokens and the video frame sequences using various feature extractors. A text encoder then adopts the BERT model to embed the text, and a video encoder utilizes the Transformer encoder to embed the video frames. Next, we employ a Transformer based cross encoder for interacting between the text and the video. Finally, a Transformer decoder is used to reconstruct the input text.

3.1.1 Pre-processing.

We first pre-process video and language before feeding to the UniVL. For the input text, we tokenize all words by WordPieces (Wu et al., 2016) following the pre-processing method in BERT to obtain the token sequence $\mathbf{t} = \{t_i | i \in [1, n]\}$, where t_i is the i-th token and n is the length of the token sequence. For each video clip, we sample a frame sequence $\mathbf{v} = \{v_j | j \in [1, m]\}$ and adopt them to extract features, where v_j is the j-th group of video frames and m is the group length of the frame sequence.

3.1.2 Single Modal Encoders.

We encode the text and video separately. Such a two-stream design has two advantages: module reusing and retrieval orienting. The module reusing means the text module can benefit from the existing text-based pretrained model, e.g., BERT. The retrieval orienting means the two-stream design is natural to retrieval tasks due to its scalability to extensive datasets. The extracted representation can be indexed, and the calculation of similarity has linear complexity in the number of videos. In this paper, we adopt the BERT-Base uncased model to generate the text representation $\mathbf{T} \in \mathbb{R}^{n \times d}$ after feeding the token sequence $\mathbf{t}$,

$$\mathbf{T} = \text{BERT}(\mathbf{t}), \tag{1}$$

where d is the hidden size of text representation.

For the video frame sequence $\mathbf{v}$, we adopt the off-the-shelf image feature extractors, e.g., S3D (Xie et al., 2018), to generate video feature $\mathbf{F}_v \in \mathbb{R}^{m \times d_v^f}$, where d_v^f is the hidden size. A Transformer encoder is utilized to embed the contextual information of video as follows,

$$\mathbf{V} = \text{Transformer}(\mathbf{F}_v). \tag{2}$$

The dimension of $\mathbf{V}$ is $\mathbb{R}^{m \times d}$.

3.1.3 Cross Encoder.

The text encoder and video encoder mainly focus on individual modality. To make the text and video fully interact, we design across encoder, which takes both the text and video modality features as input. Specifically, we first combine the text encoding $\mathbf{T}$ and the video encoding $\mathbf{V}$ to get the encoding $\mathbf{M} \in \mathbb{R}^{(n+m) \times d}$. Then, a Transformer encoder takes the encoding $\mathbf{M}$ as input to generate the attended encoding $\mathbf{M} \in \mathbb{R}^{(n+m) \times d}$,

$$\mathbf{M} = \text{Transformer}([\mathbf{T}; \mathbf{V}]), \tag{3}$$

where $[;]$ denotes the combination operation. It is noted that the combination is operated along with the dimension of sequence, not the dimension of hidden size. One reason is that the text length n and video clip length m are always different. Another reason is that the semantic between text and video are not absolutely aligned. People are likely to describe an event after or before performing it in the video (Miech et al., 2020).

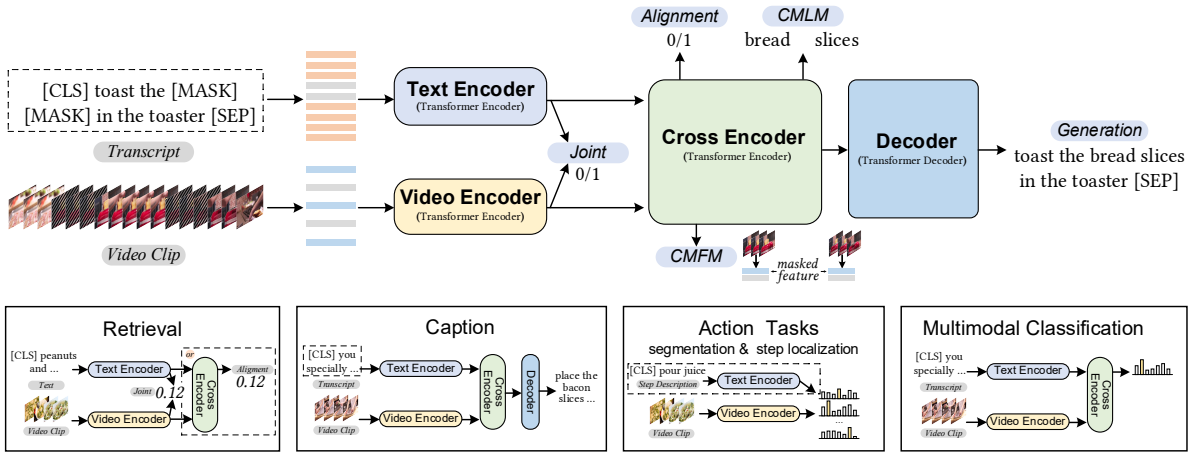


Figure 3: The main structure of our UniVL, which comprises four components, including two single-modal encoders, a cross encoder, and a decoder. The model is flexible for many text and video downstream tasks. Four possible tasks are listed.

3.1.4 Decoder.

We empower our pre-trained model to have the capability of learning from and then benefiting for generation tasks by attaching a decoder, which is usually a unidirectional recurrent/attention model to generate tokens one by one. Such a decoder module is proved useful in text-based pre-training tasks, e.g., T5 (Raffel et al., 2019) and BART (Lewis et al., 2020). It is noted that the decoder has a different target at different phases. The decoder learns to reconstruct the input text (e.g., transcripts) during pre-training because of no available text label. When fine-tuning, the decoder is used to generate results, e.g., video caption, where inputs transcripts and video and outputs caption. The input is the attended encoding $\mathbf{M}$ of text and video. We unexceptionally exploit Transformer to get the decoded feature $\mathbf{D} \in \mathbb{R}^{l \times d}$ from $\mathbf{M}$,

$$\mathbf{D} = \text{Transformer}(\mathbf{M}), \tag{4}$$

where l is the decoder length.

3.2 Pre-training Objectives

We have five pre-training objectives: 1) video-text joint, 2) conditioned masked language model (for text corruption), 3) conditioned masked frame model (for video corruption), 4) video-text alignment, and 5) language reconstruction.

3.2.1 Video-Text Joint.

As our text encoder, the BERT-Base uncased model is a robust extractor of text representation. So, we utilize a video-text joint objective to enhance the capability of the video encoder. It seems a retrieval oriented operation, which is to align the space of representation between text and video. Considering the misalignment between the text and video clip in narrated videos, we adopt MIL-NCE (Miech et al., 2020) on $\mathbf{T}$ and $\mathbf{V}$ as our joint objective,

$$\mathcal{L}_{Joint}(\theta) = -E_{(\mathbf{t},\mathbf{v})\sim\mathbf{B}} \log \text{MIL-NCE}\left(\mathbf{t}, \mathbf{v}\right), \tag{5}$$

$$\text{MIL-NCE}\left(\mathbf{t}, \mathbf{v}\right) = \frac{\sum_{(\hat{\mathbf{v}},\hat{\mathbf{t}})\in\mathcal{P}_{\mathbf{v},\mathbf{t}}} \exp(\hat{\mathbf{v}}\hat{\mathbf{t}}^\top)}{\mathcal{Z}}, \tag{6}$$

$$\mathcal{Z} = \sum_{(\hat{\mathbf{v}},\hat{\mathbf{t}})\in\mathcal{P}_{\mathbf{v},\mathbf{t}}} \exp(\hat{\mathbf{v}}\hat{\mathbf{t}}^\top) + \sum_{(\tilde{\mathbf{v}},\tilde{\mathbf{t}})\in\mathcal{N}_{\mathbf{v},\mathbf{t}}} \exp(\tilde{\mathbf{v}}\tilde{\mathbf{t}}^\top), \tag{7}$$

where $\mathcal{P}_{\mathbf{v},\mathbf{t}}$ is a set of positive video-transcript pairs. E.g., $\{(\mathbf{v},\mathbf{t}), (\mathbf{v},\mathbf{t}_{-1}), (\mathbf{v},\mathbf{t}_{+1})\}$, where $\mathbf{t}_{-1}$ and $\mathbf{t}_{+1}$ are two closest transcripts in time to $\mathbf{t}$. The negative pairs $\mathcal{N}_{\mathbf{v},\mathbf{t}}$ take negative transcripts (or video clips) from other instances within the batch $\mathbf{B}$ after fixing $\mathbf{v}$ (or $\mathbf{t}$). $\hat{\mathbf{v}}$, $\tilde{\mathbf{v}}$ and $\hat{\mathbf{t}}$, $\tilde{\mathbf{t}}$ are generated through mean-pooling on $\mathbf{V}$ and $\mathbf{T}$, respectively. θ is the trainable parameters.

3.2.2 CMLM: Conditioned Masked Language Model.

Following BERT, we also randomly mask 15% tokens with the special token [MASK] in the sentence and re-produce the masked tokens under the condition of video input and known tokens. This loss function is defined on the feature matrix of the text part in $\mathbf{M}$ as:

$$\mathcal{L}_{CMLM}(\theta) = -E_{t_m\sim\mathbf{t}} \log P_\theta\left(t_m \mid t_{\neg m}, \mathbf{v}\right), \tag{8}$$

where $t_{\neg m}$ means the contextual tokens surrounding the masked token t_m, θ is the trainable parameters.

3.2.3 CMFM: Conditioned Masked Frame Model.

Similarly, we also propose a masked frame model to predict the correct frames given contextual frames and the input text for semantic constraints. However, it is hard to reconstruct the original RGB frame. We adopt the contrastive learning method to maximize the MI (Mutual information) between the masked output features and the original features. This loss function is NCE (Sun et al., 2019a). We randomly mask 15% vectors (also 15% frames) with zeros. The objective is to identify the correct frame compared to negative distractors. The loss is defined as:

$$\mathcal{L}_{CMFM}(\theta) = -E_{v_m \sim \mathbf{v}} \log \mathrm{NCE}\left(v_m \mid v_{\neg m}, \mathbf{t}\right), \tag{9}$$

$$\mathrm{NCE}\left(v_m \mid v_{\neg m}, \mathbf{t}\right) = \frac{\exp(\mathbf{f}_{v_m} \mathbf{m}_{v_m}^{\top})}{\mathcal{Z}}, \tag{10}$$

$$\mathcal{Z} = \exp(\mathbf{f}_{v_m} \mathbf{m}_{v_m}^{\top}) + \sum\nolimits_{v_j \in \mathcal{N}(v_m)} \exp(\mathbf{f}_{v_m} \mathbf{m}_{v_j}^{\top}), \tag{11}$$

where $v_{\neg m}$ means the surrounding frames except v_m, $\mathbf{f}_{v_m} \in \mathbb{R}^{1\times d}$ is a linear output of $\mathbf{f}^{v}_{v_m} \in \mathbf{F}_v$, $\mathbf{F}_v$ is the real-valued vectors of video features, $\mathbf{m}_{v_m} \in \mathbf{M}^{(v)}$, and $\mathbf{M}^{(v)}$ is the feature matrix of the video part in $\mathbf{M}$. We take other frames in the same batch as negative cases defined as $\mathcal{N}(v_m)$.

3.2.4 Video-Text Alignment.

We use the fused representation that corresponds to the special token [CLS] to predict scores for the video-text alignment, which is similar to the BERT sentence pair classification task. We adopt the NCE loss to learn to discriminate against the positive from negative video-text pairs. To enhance this capability, we not only randomly sample negative cases but also re-sample video clips from the same video (Han et al., 2019). The reason is that the frames inside the same video are more similar than frames of different videos. This loss function is defined as follows,

$$\mathcal{L}_{Align}(\theta) = -E_{(\mathbf{t},\mathbf{v})\sim \mathbf{B}} \log \frac{\exp\left(s(\mathbf{t},\mathbf{v})\right)}{\mathcal{Z}}, \tag{12}$$

$$\mathcal{Z} = \exp\left(s(\mathbf{t},\mathbf{v})\right) + \sum\nolimits_{\mathbf{u}\in\mathcal{N}(\mathbf{v})} \exp\left(s(\mathbf{t},\mathbf{u})\right), \tag{13}$$

where $s(\cdot)$ means two linear layers with a $Tanh$ activation function between them, which is performed on the first hidden state of $\mathbf{M}$. We take other video clips in the same batch $\mathbf{B}$ as negative cases $\mathcal{N}(\mathbf{v})$.

3.2.5 Language Reconstruction.

To reconstruct the input sentence to endow the pre-trained model with the generation capability, we employed an auto-regressive decoder with reconstruction objective, and the loss function is,

$$\mathcal{L}_{Decoder}(\theta) = -E_{\hat{t}_i \sim \hat{\mathbf{t}}} \log P_\theta\left(\hat{t}_i \mid \hat{t}_{<i}, \mathbf{t}, \mathbf{v}\right). \tag{14}$$

It is noted that $\mathbf{t}$ is the masked version of ground-truth text $\hat{\mathbf{t}}$ when pre-training. As shown in BART (Lewis et al., 2019), pre-training decoder benefits generation tasks.

We jointly optimize our model by a weighted loss:

$$\begin{aligned}\mathcal{L}_{UniVL} =& \mathcal{L}_{Joint} + \mathcal{L}_{CMLM} + \mathcal{L}_{CMFM} \\ &+ \mathcal{L}_{Align} + \mathcal{L}_{Decoder}.\end{aligned} \tag{15}$$

3.3 Pre-training Strategies

We develop two pre-training strategies to train the UniVL model effectively.

3.3.1 StagedP: Stage by Stage Pre-training.

The UniVL can benefit from the pre-trained BERT-Base uncased model in the text encoder module. The natural idea is to train a peer to peer video encoder as the BERT-Base. We adopt a two-stage training fashion. For the first stage, we only preserve the text BERT and video Transformer to learn the weights using the Video-Text Joint loss Eq. (5). Next, we decrease the learning rate and continue to further pre-train the UniVL by all five objectives. One advantage is to fasten the pre-training speed, and the other advantage is to make the pre-training progress more smoothing on weights.

3.3.2 EnhancedV: Enhanced Video Representation.

To further enhance the video representation, we adopt a masked modality strategy to make the video to generate transcripts without text input. Specifically, we mask the whole text tokens with a 15% possibility. In other words, there are 15% text-video pairs with entire text tokens masked in each mini-batch, and the model utilizes the video information to complete generation. Such a strategy is a more challenging task for the model to learn a better video representation.

4 Experiments

We first pre-train our model on the large scale dataset. We download videos with ASR transcripts from Howto100M dataset (Miech et al., 2019)[1]. After filtering the unavailable ones, we get 1.2M videos for pre-training our model. On average, the duration of each video is 6.5 minutes with 110 clip-text pairs.

Then, we fine-tune our pre-trained model on five diverse downstream tasks using five datasets, including text-based video retrieval, multimodal video captioning, action segmentation, action step localization, and multimodal sentiment analysis.

4.1 Datasets

4.1.1 Youcook2

Youcook2 (Zhou et al., 2018a) contains 2,000 cooking videos on 89 recipes with 14K video clips. The overall duration is 176 hours (5.26 minutes on average). Each video clip is annotated with one captioning sentence. We evaluate both text-based video retrieval and multimodal video captioning task on this dataset.

For the text-based video retrieval task, we follow the same experimental setting in (Miech et al., 2019), and use the captions as the input text queries to find the corresponding video clips. For the video captioning task, we use the same setting as in (Shi et al., 2019). We filter the data and make sure there is no overlap between pre-training and evaluation data. In all, we have 1,261 training videos and 439 test videos, that is, 9,776 training clip-text pairs and 3,369 test clip-text pairs.

4.1.2 MSR-VTT

MSR-VTT (Xu et al., 2016) is the open-domain dataset for video retrieval tasks. It has open domain video clips, and each clip has 20 captioning sentences labeled by human. In all, there are 200K clip-text pairs from 10K videos in 20 categories including sports, music, etc. Following JSFusion (Yu et al., 2018), we randomly sampled 1,000 clip-text pairs as test data to evaluate the performance of our model on text-based video retrieval task.

4.1.3 COIN

COIN (Tang et al., 2019) is to evaluate action segmentation task, which contains 180 different tasks and 11,827 videos. Each video is labeled with 3.91 step segments. In total, the dataset contains videos of 476 hours, with 46,354 annotated segments.

4.1.4 CrossTask

CrossTask (Zhukov et al., 2019) is to evaluate the action step localization task. It contains 83 different tasks and 4.7k videos. For each task, an ordered list of steps with manual descriptions are provided.

4.1.5 CMU-MOSI

Multimodal Opinion Sentiment and Emotion Intensity (Zadeh et al., 2016) is sentence-level sentiment analysis and emotion recognition in online videos. CMU-MOSI contains 2,199 opinion video clips, each annotated with real-valued sentiment intensity annotations in the range [-3, +3]. We evaluate the performance of our model on multimodal sentiment analysis.

4.2 Experimental Details

For text encoding, we apply WordPiece embeddings (Wu et al., 2016) with a 30,000 token vocabulary to input to BERT model. We exploit the BERT-base model (Devlin et al., 2019) with 12 layers of Transformer blocks. Each block has 12 attention heads and the hidden size is 768.

For video encoding, we first extract the 3D feature from video clips using the S3D model pretrained by Miech et al. (2020). The basic visual feature can significantly affect the results from our preliminary experiments. The fps of the 3D feature extractor is 16 and the dimension is 1,024. We then employ Transformer Encoder with 6 layers to capture the sequential information on the 3D feature. Each block has 12 attention heads and the hidden size is 768.

The model consumes the clip-text pairs. The maximal input tokens of text is 32 and the maximal number of video features is 48. For short sentence and clip, we concatenate contextual tokens and frames. For cross encoder and decoder, we use a 2 layers Transformer Encoder as the encoder and a 3 layer Transformer Decoder as the decoder with 12 heads and 768 hidden size. For generation task during the inference stage, we use the beam search with the size of 5. As previously mentioned, the generated sequence is the ground-truth input transcripts in the pre-training phase. Its target is to sequentially learn full information from the masked transcripts and video features.

We pre-train our model on 8 NVIDIA Tesla V100 GPUs. There are two sets of hyper-

[1] https://www.di.ens.fr/willow/research/howto100m/

Methods	R@1	R@5	R@10	Median R
Random	0.03	0.15	0.3	1675
HGLMM (Klein et al., 2015)	4.6	14.3	21.6	75
HowTo100M (Miech et al., 2019)	8.2	24.5	35.3	24
MIL-NCE (Miech et al., 2020)	15.1	38.0	51.2	10
ActBERT (Zhu and Yang, 2020)	9.6	26.7	38.0	19
VideoAsMT (Korbar et al., 2020)	11.6	-	43.9	-
UniVL (FT-Joint)	22.2	52.2	66.2	5
UniVL (FT-Align)	**28.9**	**57.6**	**70.0**	**4**

Table 1: Results of text-based video retrieval on Youcook2 dataset.

Methods	R@1	R@5	R@10	Median R
Random	0.1	0.5	1.0	500
C+LSTM+SA (Torabi et al., 2016)	4.2	12.9	19.9	55
VSE (Kiros et al., 2014)	3.8	12.7	17.1	66
SNUVL (Yu et al., 2016)	3.5	15.9	23.8	44
Kaufman et al. (2017)	4.7	16.6	24.1	41
CT-SAN (Yu et al., 2017)	4.4	16.6	22.3	35
JSFusion (Yu et al., 2018)	10.2	31.2	43.2	13
HowTo100M (Miech et al., 2019)	14.9	40.2	52.8	9
MIL-NCE (Miech et al., 2020)	9.9	24.0	32.4	29.5
ActBERT (Zhu and Yang, 2020)	8.6	23.4	33.1	36
VideoAsMT (Korbar et al., 2020)	14.7	-	52.8	-
UniVL (FT-Joint)	20.6	49.1	62.9	6
UniVL (FT-Align)	**21.2**	**49.6**	**63.1**	**6**

Table 2: Results of text-based video retrieval on MSR-VTT dataset.

parameters considering the stage by stage pre-training strategy. In the first stage, the batch size is set to 600 and the model is trained 50 epochs for 1.5 days. In the second stage, the batch size is set to 48 and the model is trained 50 epochs for 12 days. We use the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 1e-3 in the first stage and 1e-4 in the second stage, and employ a linear decay learning rate schedule with a warm-up strategy.

4.3 Main Results

4.3.1 Text-based Video Retrieval.

Text-based video retrieval is defined to retrieve a relevant video/clip given an input text query. As shown in Figure 3 (retrieval block), the model encodes the input text query and candidate video clips through the text encoder and video encoder respectively. Then calculate the matching scores using two different approaches: one is UniVL (FT-Joint), which calculates the score through dot product as in Eq. (6), and use $\mathcal{L}_{Joint}$ as the loss during the fine-tuning stage; the other is UniVL (FT-Align), which feeds the encodings to both single encoders and the cross encoder to get unified representation and predict the match score through $s(\cdot)$ in Eq. (12) on the first token '[CLS]'. During the fine-tuning stage, the loss is $\mathcal{L}_{Align}$. We use the Adam optimizer with an initial learning rate of 3e-5 and a batch size of 32 video-caption pairs for Youcook2, an initial learning rate of 5e-5 and a batch size of 128 video-caption pairs for MSR-VTT as hyper-parameters to fine-tune for 5 epochs.

We fine-tune our pre-trained model for text-based video retrieval task on both Youcook2 and MSR-VTT datasets. The evaluation metrics are Recall@n (R@n) and Median R. Tables 1 and 2 list the retrieval results of all baselines and our model on Youcook2 and MSR-VTT separately. We can see that our model achieves the best performance

Methods	Input	B-3	B-4	M	R-L	CIDEr
Bi-LSTM (Zhou et al., 2018a)	V	-	0.87	8.15	-	-
EMT (Zhou et al., 2018b)	V	-	4.38	11.55	27.44	0.38
VideoBERT (Sun et al., 2019b)	V	6.80	4.04	11.01	27.50	0.49
CBT (Sun et al., 2019a)	V	-	5.12	12.97	30.44	0.64
ActBERT (Zhu and Yang, 2020)	V	8.66	5.41	13.30	30.56	0.65
VideoAsMT (Korbar et al., 2020)	V	-	5.3	13.4	-	-
AT (Hessel et al., 2019)	T	-	8.55	16.93	35.54	1.06
DPC (Shi et al., 2019)	V + T	7.60	2.76	18.08	-	-
AT+Video (Hessel et al., 2019)	V + T	-	9.01	17.77	36.65	1.12
UniVL	V	16.46	11.17	17.57	40.09	1.27
UniVL	T	20.32	14.70	19.39	41.10	1.51
UniVL	V + T	**23.87**	**17.35**	**22.35**	**46.52**	**1.81**

Table 3: The multimodal video captioning results on Youcook2 dataset. 'V' means video and 'T' means Transcript.

over all baselines to a large extent. We present several baseline methods with or without pre-training. Our model outperforms the Howto100M and VideoAsMT models pre-trained on the same dataset on all metrics. Besides, the experimental results present the a large performance gain with pre-training.

We also notice that UniVL (FT-Align) performs better than UniVL (FT-Joint), which demonstrates that fusion representation generated by the cross encoder is better. Nevertheless, the UniVL (FT-Joint) inference speed is 50 times for Youcook2 and 10 times for MSR-VTT faster than that of the UniVL (FT-Align). Therefore, it is a trade-off between performance and efficiency in practical applications. In the following ablation experiment, we exploit UniVL (FT-Joint) in the retrieval task.

4.3.2 Multimodal Video Captioning.

Multimodal video captioning aims to generate a sequence of descriptive sentences. As shown in Figure 3 (caption block), the model encodes the input video frames as well as transcripts inside the clips through the video encoder and text encoder respectively, then feeds the encodings to the cross encoder to get unified representation, and finally generates token sequence by the decoder. We use $\mathcal{L}_{Decoder}$ as the loss during the fine-tuning stage. The hyper-parameters are an initial learning rate of 3e-5, a batch size of 32 samples, and fine-tune for 5 epochs.

Table 3 lists the caption results of all baselines and our models on Youcook2. This generation task adopts the corpus-level generation evaluation metric using the pen-source tool[2], including BLEU (BLEU-3, B-3; BLEU-4, B-4) (Papineni et al., 2002), METEOR (M) (Banerjee and Lavie, 2005), ROUGE-L (R-L) (Lin and Och, 2004), and CIDEr (Vedantam et al., 2015). We compare our pre-trained model with several baseline methods. We classify the methods with the setting that the input is video-only or video+transcript. Zhou et al. (2018a) propose an end-to-end model for both procedural segmentation and captioning. Sun et al. (2019b,a); Zhu and Yang (2020); Korbar et al. (2020) adopt the pre-training strategy and evaluate the captioning with the only video as input. Shi et al. (2019) and Hessel et al. (2019) discuss the multimodal input with both video and transcript. Our pre-trained model achieves state-of-the-art results and outperforms the existing pre-trained models, even only considering video as input.

4.3.3 Action Segmentation.

We fine-tune our pre-train model on action segmentation task using COIN dataset, which is to predict one pre-defined label for each frame of the given video. As shown in Figure 3 (action tasks block), the model encodes the input video frames through the video encoder, followed by a linear classifier upon the output encodings for frame labeling. We do not use the text encoder due to no text description in the dataset. The evaluation metric is frame-wise accuracy (FA). The hyper-parameters are an initial learning rate of 3e-5, a batch size of 32 samples, and fine-tune for 5 epochs. The results are shown in Table 4. The UniVL significantly outperforms the baselines with more than 14% im-

[2]https://github.com/Maluuba/nlg-eval

Methods	Frame Accuracy (%)
NN-Viterbi (Richard et al., 2018)	21.17
VGG (Simonyan and Zisserman, 2014)	25.79
TCFPN-ISBA (Ding and Xu, 2018)	34.30
CBT (Sun et al., 2019a)	53.90
MIL-NCE (Miech et al., 2020)	61.00
ActBERT (Zhu and Yang, 2020)	56.95
UniVL	**70.02**

Table 4: Action segmentation results on COIN.

Methods	Average Recall (%)
Alayrac et al. (2016)	13.3
Zhukov et al. (2019)	22.4
Supervised (Zhukov et al., 2019)	31.6
HowTo100M (Miech et al., 2019)	33.6
MIL-NCE (Miech et al., 2020)	40.5
ActBERT (Zhu and Yang, 2020)	41.4
UniVL	**42.0**

Table 5: Action step localization results on CrossTask.

provements. It shows that the pre-trained UniVL actually learns a good visual representation, even absent of linguistic descriptions.

4.3.4 Action Step Localization.

We evaluate the action step localization on CrossTask dataset. As shown in Figure 3 (action tasks block), the model encodes the step description (action) and video clip through the text encoder and the video encoder respectively. And then calculate the relevance scores through dot product similar to the retrieval task. To fairly compare to (Miech et al., 2019, 2020; Zhu and Yang, 2020), we do not fine-tune on the CrossTask dataset. We perform the evaluation protocol by reporting the average recall (CTR) metric for the localization task[3]. The results are shown in Table 5. Our results are even better than the supervised baseline, which demonstrates our UniVL model can learn better joint text-video representation.

4.3.5 Multimodal Sentiment Analysis.

We evaluate the multimodal sentiment analysis on CMU-MOSI dataset, the goal of which is to identify the sentiment of speaker based on the speakers display of verbal and nonverbal behaviors. We employ video and corresponding transcripts to accomplish this task. As shown in Figure 3 (multimodal classification block), the model encodes the input video frames as well as transcripts inside the clips through the video encoder and text encoder, respectively. Then feeds the encodings to the cross encoder to get unified representation, and finally predicts the sentiment score by a linear on the first token '[CLS]'. The hyper-parameters are an initial learning rate of 1e-5, a batch size 32, and fine-tune for 3 epochs.

The results are shown in Table 6. Following (Zadeh et al., 2019), the evaluation metrics are binary accuracy (BA), F1 score, Mean-Absolute Error (MAE), and Pearson Correlation Coefficient (Corr). Compared with the baseline using video, transcript, and audio inputs, our model trained with video and language still achieves the best results without audio information.

4.4 Ablation Studies

We analyze the effectiveness of our model design on pre-training objectives and strategies through ablation studies over text-based video retrieval and multimodal video captioning tasks. We also discuss the effectiveness of various visual features.

[3]The result is generated following the evaluation process of official project: https://github.com/DmZhukov/CrossTask

Methods	BA	F1	MAE	Corr
MV-LSTM (Rajagopalan et al., 2016)	73.9/-	74.0/-	1.019	0.601
TFN (Zadeh et al., 2017)	73.9/	73.4/-	1.040	0.633
MARN (Zadeh et al., 2018b)	77.1/	77.0/-	0.968	0.625
MFN (Zadeh et al., 2018a)	77.4/	77.3/-	0.965	0.632
RMFN (Liang et al., 2018)	78.4/	78.0/-	0.922	0.681
RAVEN (Wang et al., 2019)	78.0/	-/-	0.915	0.691
MulT (Tsai et al., 2019)	/83.0	-/82.8	0.870	0.698
FMT (Zadeh et al., 2019)	81.5/83.5	81.4/83.5	0.837	0.744
UniVL	**83.2/84.6**	**83.3/84.6**	**0.781**	**0.767**

Table 6: Multimodal sentiment analysis results on CMU-MOSI dataset. BA means binary accuracy, MAE is Mean-absolute Error, and Corr is Pearson Correlation Coefficient. For BA and F1, we report two numbers following Zadeh et al. (2019): the number on the left side of / is calculated based on the approach from Zadeh et al. (2018b), and the right side is by Tsai et al. (2019).

Methods	Dataset	R@1	R@5	R@10	Median R
UniVL	Youcook2	22.2	52.2	66.2	5
-w/o Joint	Youcook2	19.5	48.0	62.7	6
-w/o Alignment	Youcook2	16.3	42.3	56.2	8
-w/o EnhancedV	Youcook2	16.1	41.3	55.8	8
-w/o Decoder	Youcook2	14.6	40.3	55.5	8
-w/o StagedP	Youcook2	11.9	35.0	48.9	11
-w/o Pre-training	Youcook2	7.7	23.9	34.7	21
UniVL	MSR-VTT	20.6	49.1	62.9	6
-w/o Joint	MSR-VTT	19.6	45.9	62.6	6
-w/o Alignment	MSR-VTT	19.3	44.6	60.1	7
-w/o EnhancedV	MSR-VTT	18.0	45.3	59.3	7
-w/o Decoder	MSR-VTT	18.9	44.9	57.8	7
-w/o StagedP	MSR-VTT	18.0	44.3	57.7	8
-w/o Pre-training	MSR-VTT	16.7	44.0	55.9	8

Table 7: Ablation study on retrieval task. '-w/o' means reducing the condition above the previous line.

Methods	B-3	B-4	M	R-L	CIDEr
UniVL	23.87	17.35	22.35	46.52	1.81
-w/o Joint	23.96	17.54	22.48	46.77	1.84
-w/o Alignment	23.51	17.24	22.02	45.90	1.77
-w/o EnhancedV	23.15	17.04	21.83	45.89	1.76
-w/o Decoder	19.01	13.22	19.43	43.62	1.53
-w/o StagedP	18.13	12.49	18.78	42.64	1.46
-w/o Pre-training	14.23	9.46	16.27	37.44	1.15

Table 8: Ablation study on caption task of Youcook2 dataset. '-w/o' means reducing the condition above the previous line.

Method	Visual Feature	R@1	R@5	R@10	Median R
UniVL on Youcook2	RS152 + RX101	11.5	29.1	40.1	17
	S3D	22.2	52.2	66.2	5
UniVL on MSR-VTT	RS152 + RX101	18.7	44.4	58.9	7
	S3D	20.6	49.1	62.9	6

Table 9: Ablation study of visual features for retrieval task. RS152 denotes ResNet-152, RX101 means ResNeXt-101.

Method	Visual Feature	B-3	B-4	M	R-L	CIDEr
UniVL	RS152 + RX101	20.42	14.31	19.92	42.35	1.47
	S3D	23.87	17.35	22.35	46.52	1.81

Table 10: Ablation study of visual features for multimodal video captioning results on Youcook2 dataset. RS152 denotes ResNet-152, RX101 means ResNeXt-101.

4.4.1 Modules and Strategies.

Table 7 shows the effectiveness of each objective or strategy on the retrieval task. The results are reported on both Youcook2 and MSR-VTT datasets. Simultaneously, Table 8 demonstrates the effectiveness of each objective or strategy on the caption task. For the retrieval task, we exploit UniVL (FT-Joint) fine-tuning strategy to study the objectives: Joint loss, Alignment loss, and Decoder loss, and the strategies: StagedP and EnhancedV show consistent improvement. From the result, we can see that the cross encoder and decoder modules can promote the joint representation of video and text. For the caption task, we find that the decoder module shows great advantage and achieves more than 3 points gain on the BLUE-4 metric. Another finding is that the Joint loss decreases the generation task a little, although it performs well in the retrieval task. Excessive emphasis on coarse-grained matching can affect the fine-grained description at the generation task.

4.4.2 Visual Features.

We compare the S3D video feature pre-trained on Howto100M and ResNet-152 plus ResNeXt-101 pre-trained on labeled ImageNet and Kinetics respectively. The ResNet-152 (RS152) and ResNeXt-101 (RX101) are used to extract 2D and 3D features from video clips respectively similar to Miech et al. (2019)'s work.

As shown in Table 9 and Table 10, the visual feature is important in our pre-training model and the downstream tasks. It is worth studying an end to end training from raw videos instead of extracted fixed video features in the future. However, the time-cost and the memory-cost are enormous. The key bottleneck is visual representation, and we propose two possible approaches: designing a lightweight training scheme, e.g., training on keyframes of video, using a small feature dimension size.

5 Conclusion and Discussion

This paper proposes UniVL with self-supervised learning for video and language representation on large scale videos. The UniVL is designed with four modules and five objectives for both video-language understanding and generation tasks. It is a flexible model for most of the multimodal downstream tasks considering both efficiency and effectiveness. We conduct extensive experiments on evaluating our model for five downstream tasks, e.g., text-based video retrieval and multimodal video captioning. The experimental results demonstrate that our pre-trained model can improve the performance to a large extent over the baseline models and achieve state-of-the-art results on five typical multimodal tasks. Besides, we will investigate our model's performance on more massive datasets and more downstream tasks for future work.

References

Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. 2016. Unsupervised learning from narrated instruction videos. In *CVPR*, pages 4575–4583.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Li Ding and Chenliang Xu. 2018. Weakly-supervised action segmentation with iterative soft boundary assignment. In *CVPR*, pages 6508–6516.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language

understanding and generation. *arXiv preprint arXiv:1905.03197*.

Tengda Han, Weidi Xie, and Andrew Zisserman. 2019. Video representation learning by dense predictive coding. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0.

Jack Hessel, Bo Pang, Zhenhai Zhu, and Radu Soricut. 2019. A case study on combining asr and visual features for generating instructional video captions. In *CoNLL*.

Dotan Kaufman, Gil Levi, Tal Hassner, and Lior Wolf. 2017. Temporal tessellation: A unified approach for video analysis. In *ICCV*, pages 94–104.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR*.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2015. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*, pages 4437–4446.

Bruno Korbar, Fabio Petroni, Rohit Girdhar, and Lorenzo Torresani. 2020. Video understanding as machine translation. *arXiv preprint arXiv:2006.07203*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880.

Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2019a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *arXiv preprint arXiv:1908.06066*.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019b. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. In *EMNLP*, pages 150–161.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *ACL*, page 605.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

William Lotter, Gabriel Kreiman, and David Cox. 2016. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23.

Michael Mathieu, Camille Couprie, and Yann LeCun. 2015. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *CVPR*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. *ICCV*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrusaitis, and Roland Goecke. 2016. Extending long short-term memory for multi-view structured learning. In *ECCV*, pages 338–353.

Alexander Richard, Hilde Kuehne, Ahsan Iqbal, and Juergen Gall. 2018. Neuralnetwork-viterbi: A framework for weakly supervised video learning. In *CVPR*, pages 7386–7395.

Botian Shi, Lei Ji, Yaobo Liang, Nan Duan, Peng Chen, Zhendong Niu, and Ming Zhou. 2019. Dense procedure captioning in narrated instructional videos. In *ACL*, pages 6382–6391.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. 2015. Unsupervised learning of video representations using lstms. In *ICML*, pages 843–852.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *ICLR*.

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019a. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019b. Videobert: A joint model for video and language representation learning. *ICCV*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. COIN: A large-scale dataset for comprehensive instructional video analysis. In *CVPR*, pages 1207–1216.

Atousa Torabi, Niket Tandon, and Leonid Sigal. 2016. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, pages 6558–6569.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575.

Xiaolong Wang and Abhinav Gupta. 2015. Unsupervised learning of visual representations using videos. In *ICCV*, pages 2794–2802.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *AAAI*, pages 7216–7223.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, pages 318–335.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, pages 487–503.

Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. 2016. Video captioning and retrieval models with semantic attention. In *ECCVLSMDC2016 Workshop*.

Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. 2017. End-to-end concept word detection for video captioning, retrieval, and question answering. In *CVPR*, pages 3261–3269.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, pages 1103–1114.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. *AAAI*.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018b. Multi-attention recurrent network for human communication comprehension. In *AAAI*.

Amir Zadeh, Chengfeng Mao, Kelly Shi, Yiwei Zhang, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. 2019. Factorized multimodal transformer for multimodal sequential learning. *arXiv preprint arXiv:1911.09826*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. 2019. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*.

Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018a. Towards automatic learning of procedures from web instructional videos. In *AAAI*.

Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018b. End-to-end dense video captioning with masked transformer. In *CVPR*, pages 8739–8748.

Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *CVPR*.

Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. 2019. Cross-task weakly supervised learning from instructional videos. In *CVPR*.