Multilingual Multimodal Pretraining for Zero-Shot Cross-lingual Transfer of Vision-Language Models

Anonymous submission

Abstract

This paper studies zero-shot cross-lingual transfer of vision-language models. Specifically, we focus on multilingual text-to-video search and propose a Transformer-based model that learns contextualized multilingual multimodal embeddings. Under a zero-shot setting, we empirically demonstrate that performance degrades significantly when we query the multilingual text-video model with non-English sentences. To address this problem, we introduce a multilingual multimodal pre-training strategy, and collect a new multilingual instructional video dataset (Multi-HowTo100M) for pre-training. Experiments on VTT show that our method significantly improves video search in non-English languages without additional annotations. Furthermore, when multilingual annotations are available, our method outperforms recent baselines by a large margin in multilingual text-to-video search on VTT and VATEX; as well as in multilingual text-to-image search on Multi30K. Our model and Multi-HowTo100M will be made available.[1]

1 Introduction

One of the key challenges at the intersection of computer vision (CV) and natural language processing (NLP) is building versatile vision-language models that not only work in English, but in all of the world's approximately 7,000 languages. Since collecting and annotating task-specific parallel multimodal data in all languages is impractical, a framework that makes vision-language models generalize across languages is highly desirable.

One technique that has shown promise to greatly improve the applicability of NLP models to new languages is *zero-shot cross-lingual transfer*, where models trained on a source language are applied as-is to a different language without any additional annotated training data (Täckström et al., 2012; Klementiev et al., 2012; Cotterell and Heigold, 2017; Chen et al., 2018; Neubig and Hu, 2018). In particular, recent techniques for cross-lingual transfer have demonstrated that by performing unsupervised learning of language or translation models on many languages, followed by downstream task fine-tuning using only English annotation, models can nonetheless generalize to a non-English language (Wu and Dredze, 2019a; Lample and Conneau, 2019; Huang et al., 2019a; Artetxe et al., 2020; Hu et al., 2020). This success is attributed to the fact that many languages share a considerable amount of underlying vocabulary or structure. At the vocabulary level, languages often have words that stem from the same origin, for instance, "desk" in English and "Tisch" in German both come from the Latin "discus". At the structural level, all languages have a recursive structure, and many share traits of morphology or word order.

For cross-lingual transfer of vision-language models, the visual information is clearly an essential element. To this end, we make an important yet under-explored step to incorporate visual-textual relationships for improving multilingual models (Delvin et al., 2018; Artetxe et al., 2020). While spoken languages could be different, all humans share similar vision systems, and many visual concepts can be understood universally (Sigurdsson et al., 2020; Zhang et al., 2020). For example, while  is termed "cat" for an English speaker and "chat" for a French speaker; they understand  similarly. We leverage this observation to learn to associate sentences in different languages with visual concepts for promoting cross-lingual transfer of vision-language models.

In this work, we focus on multilingual text-to-video search tasks and propose a Transformer-based video-text model to learn contextualized mul-

[1] http://github.com/anonymity/xxxxxx

tilingual multimodal representations. Our vanilla model yields state-of-the-art performance in multilingual text→video search when train with multilingual annotations. However, under the zero-shot setting, rather surprisingly, there is a significant performance gap between English and non-English queries (see §5.5 for details). To resolve this problem, motivated by recent advances in large-scale language model (Artetxe et al., 2020) and multimodal pre-training (Lu et al., 2019; Miech et al., 2019; Patrick et al., 2020), we propose a multilingual multimodal pre-training (MMP) strategy to exploit the weak supervision from large-scale multilingual text-video data. We construct the Multilingual-HowTo100M dataset, that extends the English HowTo100M (Miech et al., 2019) dataset to contain subtitles in 9 languages for 1.2 million instructional videos.

Our method has two important benefits. First, compared to pre-training on English-video data only, pre-training on multilingual text-video data exploits the additional supervision from a variety of languages, and therefore, enhances the search performance on an individual language. Second, by exploiting the visual data as an implicit "pivot" at scale, our methods learns better alignments in the multilingual multimodal embedding space (*e.g.*, "cat"-  -"chat"), which leads to improvement in zero-shot cross-lingual transfer (*e.g.*, from "cat"-  to "chat"- ) of vision-language models.

In our experiments on VTT (Xu et al., 2016) and VATEX (Wang et al., 2019), our method yields state-of-the-art English→video search performance. For zero-shot cross-lingual transfer, the proposed multilingual multimodal pre-training improves English-video pre-training by $2 \sim 2.5$ in average R@1 across 9 languages. Additionally, when training with in-domain multilingual annotations as other baselines, our method outperforms them by a large margin in multilingual text→video search on VATEX and text→image search on Multi30K (Elliott et al., 2016).

To summarize, we make the following contributions: (1) We propose a transformer-based video-text model that learns contextualized multilingual multimodal representations (§3.1). (2) We empirically demonstrate that vision-language models, unlike NLP models, have limited zero-shot cross-lingual transferrability. (§5.5). (3) We introduce the multilingual multimodal pre-training strategy and construct a new Multi-HowTo100M dataset (§4) for pre-training to improve zero-shot cross-lingual capability of vision-language models. (4) We demonstrate the effectiveness of our approach, by achieving state-of-the-art multilingual text→video search performance in both the zero-shot (§5.5) and fully supervised setup (§5.6).

2 Related Work

Cross-lingual representations. Early work on learning non-contextualized cross-lingual representations used either parallel corpora (Gouws and Søgaard, 2015; Luong et al., 2015) or a bilingual dictionary to learn a transformation (Faruqui and Dyer, 2014; Mikolov et al., 2013). Later approaches reduced the amount of supervision using self-training (Artetxe et al., 2017). With the advances in monolingual transfer learning (McCann et al., 2017; Howard and Ruder, 2018; Peters et al., 2018; Devlin et al., 2019), multilingual extensions of pre-trained encoders have been proven effective in learning deep contextualized cross-lingual representations (Eriguchi et al., 2017; Lample and Conneau, 2019; Wu and Dredze, 2019b; Siddhant et al., 2020; Pires et al., 2019; Pfeiffer et al., 2020). We extend prior work to incorporate visual context.

Video-text representations. The HowTo100M dataset (Miech et al., 2019) has spurred significant interest in leveraging pre-training for text→video search (Korbar et al., 2020), captioning (Iashin and Rahtu, 2020), and unsupervised translation (Sigurdsson et al., 2020). This work studies a challenging and unexplored task: Zero-shot cross-lingual transfer of vision-language models. Moreover, unlike prior image/video-text work that utilizes RNN with word embeddings (Dong et al., 2019; Chen et al., 2020a; Burns et al., 2020; Kim et al., 2020) and/or a inter-modal contrastive objective (Sigurdsson et al., 2020; Liu et al., 2019; Huang et al., 2019b), our work employs Transformers to learn contextualized multilingual multimodal representations and uniquely models cross-lingual instances.

Cross-lingual Transfer. Cross-lingual transfer has proven effective in many NLP tasks including dependency parsing (Schuster et al., 2019), named entity recognition (Rahimi et al., 2019), sentiment analysis (Barnes et al., 2019), document classification (Schwenk and Li, 2018), and question answering (Lewis et al., 2020; Artetxe et al., 2020). Recently, XTREME (Hu et al., 2020) was proposed to evaluate the cross-lingual transfer capabilities of multilingual representations across a diverse set of

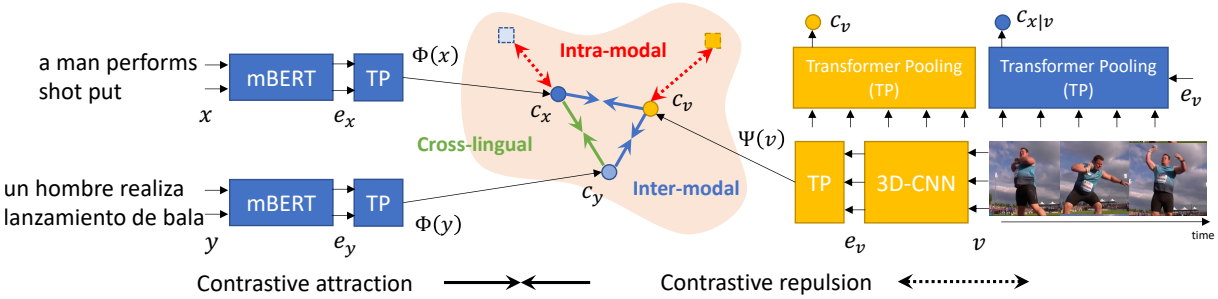


Figure 1: An overview of our video-text model for learning contextualize multilingual multimodal representations. We utilize *intra-modal*, *inter-modal*, and conditional *cross-lingual* contrastive objectives to align (x, v, y) where x and y are the captions or transcriptions in different languages of a video v. TP: Transformer pooling head.

NLP tasks and languages. However, a comprehensive evaluation of multilingual multimodal models on zero-shot cross-lingual transfer capabilities is still missing. To our best knowledge, we are the first work that investigates and improves zero-shot cross-lingual transfer of vision-language models.

3 Method

We consider the problem of learning multilingual multimodal representations from a corpus $\mathcal{C}$ of video-text pairs $\{(x_i, v_i)\}_{i=1}^{C}$, where v_i is a video clip and x_i is its corresponding text (caption or transcription) that is written in one of K languages. Our goal is to learn a shared multilingual text encoder $c_x = \Phi(x)$ and a video encoder $c_v = \Psi(v)$, both of which project the input to a shared D-dimensional embedding space $c_v, c_t \in \mathbb{R}^D$, where semantically similar instances (*i.e.*, paired (x_i, v_i)) are closer to each other than the dissimilar ones (*i.e.*, $(x_i, v_j), i \neq j$). In the following, we denote a batch of multilingual text-video samples as $\mathcal{B} = \{(x_i, v_i)\}_{i=1}^{B}\}$ where $\mathcal{B} \subset \mathcal{C}$.

3.1 Multilingual Multimodal Transformers

Figure 1 gives an overview of the proposed method. Our text encoder consists of a multilingual Transformer (*e.g.* multilingual BERT (Delvin et al., 2018)) and a text Transformer pooling head (explained below). Similarly, our video encoder consists of a 3D-CNN (*e.g.* R(2+1)D network (Tran et al., 2018)) and a video Transformer pooling head. We use these multilingual multimodal Transformers to encode text and video for alignment.

Unlike prior multilingual text-image models (Gella et al., 2017; Kim et al., 2020; Huang et al., 2019b) that utilize word embeddings and RNNs, our multilingual text encoder is built on a multilingual Transformer that generates contextualized multilingual representations $e_x \in \mathbb{R}^{N \times D}$ to encode a sentence x containing N words. We employ an additional 2-layer Transformer which we will call a "Transformer pooling head (TP)" as it serves as a pooling function to selectively encode variable-length sentences and aligns them with the corresponding visual content. We use the first output token of the second Transformer layer as the final sentence representation. Precisely, we set $c_x = \text{Trans}_x^{(2)}(\text{query=key=value}=e_x)[0]$ where $\text{Trans}_x^{(2)}$ is a 2-layer stack of Transformers (Vaswani et al., 2017) with e_x as the (query,key,value) in the multi-head attention. Note that we use the same text encoder to encode sentences in all languages.

For encoding videos, our model uses pre-trained 3D-CNNs that encode spatial-temporal context in a video. For a M-second video v, we apply R(2+1)D (Tran et al., 2018) and S3D (Miech et al., 2020) networks to its frames, concatenate network outputs, and apply a linear layer to encode the visual input, $e_v \in \mathbb{R}^{M \times D}$, to our model. Similarly to the text part, we employ a two-layer Transformer as the pooling head to encode videos with different lengths into fixed-length representations. Formally, we set $c_v = \text{Trans}_v^{(2)}(\text{query=key=value}=e_v)[0]$. Since videos are typically long and have a high frame rate (*e.g.*, 30 fps), it is infeasible to update 3D-CNNs simultaneously and therefore, we use pre-extracted video features. Our model is parameterized by $\theta = \theta_{\text{mBERT}} \cup \theta_{\text{Trans}_x} \cup \theta_{\text{Trans}_v}$.

3.2 Multilingual Text-Video Alignment

For learning multimodal representations, the common practice is to minimize a contrastive objective to map the associated (video, text) embeddings to be near to each other in a shared embedding space. The inter-modal max-margin triplet loss has been widely studied in video-text (Yu et al., 2018; Liu et al., 2019) and image-text (Kim et al., 2020; Burns et al., 2020; Huang et al., 2019b) re-

search. In this work, we generalize and model all *inter-modal*, *intra-modal*, and *cross-lingual* instances with a noise contrastive estimation objective (NCE) (Gutmann and Hyvärinen, 2010; Oord et al., 2018; Chen et al., 2020b).

Inter-modal NCE. Let $\mathcal{X}$ and $\mathcal{V}$ denote the subsets of the sampled sentences in multiple languages and videos in $\mathcal{B}$, respectively. And let $s(a,b) = \frac{a^T b}{\|a\|\|b\|}$ be the similarity measure. We use an (inter-modal) NCE objective defined as:

$$\mathcal{L}(\mathcal{X},\mathcal{V}) = -\frac{1}{B}\sum_{i=1}^{B}\log \ell^{\text{NCE}}(\Phi(x_i),\Psi(v_i)), \quad (1)$$

where

$$\ell^{\text{NCE}}(c_x,c_v) = \frac{e^{s(c_x,c_v)}}{e^{s(c_x,c_v)} + \sum_{(x',v')\sim\mathcal{N}} e^{s(c_{x'},c_{v'})}}. \quad (2)$$

In inter-modal NCE, the noise $\mathcal{N}$ is a set of "negative" video-text pairs sampled to enforce the similarity of paired ones are high and and those do not are low. Following Miech et al. (2020), we set the negatives of (x_i, v_i) as other x_j and $v_j, j \neq i$ in $\mathcal{B}$.

Intuitively, inter-modal NCE draws paired (semantically similar) instances closer and pushes apart non-paired (dissimilar) instances. Note that we do not distinguish language types in $\mathcal{X}$ and the sentences in all possible languages will be drawn towards their corresponding videos in the shared multilingual text-video embedding space.

Intra-modal NCE. Beyond cross-modality matching, we leverage the intra-modal contrastive objective to learn and preserve the underlying structure within the video and text modality. For example, *Corgi* should be closer to *Husky* than *Balinese*. Prior image-text work (Gella et al., 2017; Huang et al., 2019b) utilizes a triplet loss to maintain such neighborhood relationships. Inspired by recent success in self-supervised image and video representation learning (Yalniz et al., 2019; Ghadiyaram et al., 2019), our model leverages intra-modal NCE that constrains the learned representations to be invariant against noise and to maintain the within-modality structure simultaneously. We minimize the following intra-modal NCE loss:

$$\mathcal{L}^{\text{intra}} = \mathcal{L}(\mathcal{X},\mathcal{X}^m) + \mathcal{L}(\mathcal{V},\mathcal{V}^m), \quad (3)$$

where $\mathcal{X}^m$ and $\mathcal{V}^m$ are the noised version of the original sentences and videos. For noising, we randomly mask 5% of the multilingual text tokens and video clips. We optimize our model by $\min_\theta \mathcal{L}^{\text{inter}} + \mathcal{L}^{\text{intra}}$.

3.3 When Visually-Pivoted Multilingual Annotations Are Available

In many multilingual multimodal datasets, there are sentences in different languages that describe a shared visual context. For example, 10 English and 10 Chinese descriptions are available for each video in VATEX. With these visually-pivoted (weakly paralleled) sentences (x, y), we further revise the contrastive objectives to leverage this additional supervisory signal. Given a visually-pivoted corpus $\mathcal{C}^p$ that contains all possible combination of visually-pivoted pairs $\{(x_i, v_i, y_i)\}_{i=0}^{C_p}$, we sample batches $\mathcal{B}^p = \{(x_i, v_i, y_i)\}_{i=1}^{\mathcal{B}^p}, \mathcal{B}^p \subset \mathcal{C}^p$ and revise the contrastive objective as:

$$\mathcal{L}^{\text{inter}} = \mathcal{L}(\mathcal{X},\mathcal{V}) + \mathcal{L}(\mathcal{Y},\mathcal{V}) \quad (4)$$

$$\mathcal{L}^{\text{intra}} = \mathcal{L}(\mathcal{X},\mathcal{X}^m) + \mathcal{L}(\mathcal{Y},\mathcal{Y}^m) + \mathcal{L}(\mathcal{V},\mathcal{V}^m) \quad (5)$$

Visual-pivoted Cross-lingual NCE. Inspired by Translation Language Modeling (TLM) in XLM (Lample and Conneau, 2019), we propose a multimodal TLM-like contrastive objective which promotes alignments of descriptions in different languages that describe the same video. We use the intuition that conditioned on a video, the descriptions (need not to be translation pairs) in different languages would likely be semantically similar. To this end, we set the cross-lingual NCE as:

$$\mathcal{L}^{\text{cross}} = \mathcal{L}(\mathcal{X}|\mathcal{V},\mathcal{Y}|\mathcal{V}) \quad (6)$$

For visually-pivoted sentences, as shown in Fig. 1, we generate their representations conditioned on the video they describe. We extend the *key* and *value* of multihead attention with the additional visual content e_v and generate new $c_{x|v}$ and $c_{y|v}$ for matching. Specifically, our model employs $c_{x|v} = \text{Trans}_x^{(2)}(\text{query}{=}e_x, \text{key}{=}\text{value}{=}e_x||e_v)[0]$. With the access to (visually-pivoted) multilingual annotations, we optimize our model by $\min_\theta \mathcal{L}^{\text{inter}} + \mathcal{L}^{\text{intra}} + \mathcal{L}^{\text{cross}}$.

For inference, we use $c_x = \Phi(x)$ and $c_v = \Psi(v)$ to encode multilingual text and video for search.

4 The Multilingual HowTo100M Dataset

As large-scale pre-training has been shown important in recent NLP and vision-language models, we construct the **Multilingual HowTo100M** dataset (Multi-HowTo100M) to facilitate research in multilingual multimodal learning. The original HowTo100M (Miech et al., 2019) dataset is a

large-scale video collection of 1.2 million instructional videos (138 million clips) on YouTube, along with their automatic speech recognition (ASR) transcriptions as the subtitles. For each video in HowTo100M, we collect the multilingual subtitles provided by YouTube, which either consist of user-generated subtitles or those generated by Google ASR and Translate in the absence of user-generated ones. Essentially, we collect video subtitles in 9 languages: English (*en*), German (*de*), French (*fr*), Russian (*ru*), Spanish (*es*), Czech (*cz*), Swahili (*sw*), Chinese (*zh*), Vietnamese (*vi*).

At the time of dataset collection, there are 1.1 million videos available, each with subtitles in 7-9 languages. We utilize Multi-HowTo100M for multilingual multimodal pre-training to exploit the weak supervision from large-scale multilingual text-video data. Please refer to the supplementary material for more details.

5 Experiment

In this section, we first describe our experimental setup (§5.1-5.3). In §5.4, we conduct ablation studies to validate the effectiveness of proposed multilingual text-video model . With the best models at hand, we investigate their zero-shot cross-lingual transferability in §5.5, where we showcase that the proposed multilingual multimodal pre-training serves as the key facilitator. We then verify the superior text→video search performance of our method under the monolingual, multilingual, and cross-modality settings in §5.6.

5.1 Evaluation Datasets

MSR-VTT (VTT) (Xu et al., 2016) contains 10K videos, where each video is annotated with 20 captions. Additionally, we created pseudo-multilingual data by translating the English captions into 8 languages with off-the-shelf machine translation models.[2] We use the official training set (6.5K videos) and validation set (497 videos). We follow the protocol in Miech et al. (2019); Liu et al. (2019) which evaluates on text→video search with the 1K testing set defined by Yu et al. (2018).

VATEX (Wang et al., 2019) is a multilingual (Chinese and English) video-text dataset with 35K videos. Five (*en*,*zh*) translation pairs and five non-paired *en* and *zh* descriptions are available for each video. We use the official training split (26K videos) and follow the testing protocol in Chen et al. (2020a) to split the validation set equally into 1.5K validation and 1.5K testing videos.

Multi30K (Elliott et al., 2016) is a multilingual extension of Flickr30K (Young et al., 2014). For each image, there are two types of annotations available: (1) One parallel (English,German,French,Czech) translation pair and (2) five English and five German descriptions collected independently. The training, validation, and testing splits contain 29K, 1K, and 1K images respectively.

5.2 Implementation Details

For the video backbone, we use a 34-layer, R(2+1)-D (Tran et al., 2018) network pre-trained on IG65M (Ghadiyaram et al., 2019) and a S3D (Miech et al., 2020) network pre-trained on HowTo100M. We pre-extract video features and concatenate the two 3D-CNN outputs to form $e_x \in \mathbb{R}^{M\times 1024}$ as a video input.

Our model uses multilingual BERT (mBERT) (Devlin et al., 2018) or XLM-Roberta-large (XLM-R) (Artetxe et al., 2020), where the latter achieves near SoTA cross-lingual transfer performance for NLP tasks. Following Hu et al. (2020), instead of using the top layer, we output the 12-th layer in XLM-R and mBERT. For vision-language tasks, we freeze layers below 9 as this setup empirically performs the best.

We use a 2-layer Transformer with 4-head attention for each TP module. The embedding dimension D is set to 1024. We use the Adam (Kingma and Ba, 2015) optimizer and a 0.0002 learning rate to train our model for 16 (pre-training) and 10 (fine-tuning) epochs.

5.3 Experimental Setup

We use Multi-HowTo100M for multilingual multimodal pre-training (MMP). For each video, we randomly sample the start and end time to construct a clip. For a clip, we sample one language type each time from 9 languages and use the consecutive ASR transcriptions that are closest in time to compose (video, text) pairs for training. For simplicity and speed purposes, we follow the training protocol of XLM-R to pretrain on a multilingual corpus *wihtout* using translation pairs *i.e.*, we use multilingual text-video pairs but no translation pairs from Multi-HowTo100M dataset and utilize only inter- and intra-modal NCE (Eq. 1-3) for MMP.

We fine-tune our model on VTT, VATEX, and Multi30K to evaluate on text→video search tasks. In the zero-shot cross-lingual transfer experiments,

[2] https://marian-nmt.github.io/

Text-B	Video-B	R@1↑	R@5↑	R@10↑
XLM-R	S3D	19.5	49.0	62.8
XLM-R	R(2+1)D	19.0	49.5	63.2
XLM-R	R+S	**21.0**	**50.6**	**63.6**
mBERT	R+S	19.9	49.8	62.5

Table 1: **Text and Video (B)ackbone comparison.**

T layers	V layers	R@1↑	R@5↑	R@10↑
1	1	20.0	50.3	63.2
2	1	20.1	50.5	63.8
2	2	**21.0**	**50.6**	63.6
2*	2*	20.7	50.5	63.3
4	4	20.8	50.4	**63.8**

Table 2: **Architecture comparison.** Number of multilingual multimodal transformer layers. *Weight sharing between video and text transformers.

Objective	Inter	Intra	Cross	R@1↑	R@5↑	R@10↑
Triplet	✓			13.3	36.0	55.2
Triplet	✓	✓		20.9	49.3	63.0
NCE	✓			21.4	49.3	61.1
NCE	✓	✓		21.0	50.6	63.6
NCE*	✓	✓		21.3	50.7	63.5
NCE*	✓	✓	✓	**21.5**	**51.0**	**63.8**

Table 3: **Objective comparison.** *Training with additional machine translated *de*-video and *fr*-video pairs.

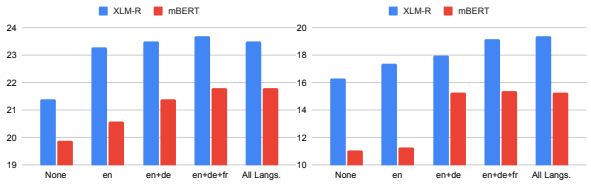
we use only English-video data and fine-tune with Eq. 1-3. We then test the model with non-English queries. When annotations in additional languages are available (by humans in VATEX and Multi30K; by MT models (*i.e.* *translate-train*) in VTT), we train our model with all available multilingual annotations (*i.e.* fully supervised) with Eq. 4-6 to demonstrate the upper bound of the zero-shot setup and to compare fairly with other baselines in multilingual text→video search. We report the standard recall at k (R@k) metrics (higher is better).

5.4 Comparison Experiments and Ablations

In this section, we ablate and compare different text/video backbones, model architectures, and learning objectives on VTT English→video search. Additional discussion can be found in the supplementary material.

Text and Video encoders. Table 1 compares different text and video encoders. While R(2+1)D outperforms S3D, the simple concatenation (*i.e.* early-fusion) of their output features provides a $1.5 \sim 2.0$ improvement in R@1. For the text encoder, XLM-R significantly outperforms mBERT.

Transformer Pooling. Table 2 compares various configurations of the proposed pooling module. We observe that a simple 2-layer Transformer achieves the best performance. Weight sharing of the video and text Transformer slightly degrades the performance. Therefore, we choose to separate them.



(a) English→Video (b) Zero-shot German→Video

Figure 2: R@1 trends in languages used for multilingual multimodal pre-training. Left: English→video search. Right: Zero-shot German→video search.

Learning Objective. From Table 3, the intra-modal contrastive objective is important for both NCE and Triplet loss. In general, the NCE loss outperforms the Triplet loss. The proposed inter-modal and intra-modal NCE objective achieves the best performance. When captions in different languages are available, cross-lingual NCE additionally provides a consistent improvement.

5.5 VTT Zero-Shot Cross-Lingual Transfer

Table 4 shows the multilingual text→video search results on VTT. With the best English-video models at hand (with mBERT or XLM-R as the text backbone), we first investigate how well these models transfer to other non-English languages under the zero-shot setting. We then demonstrate the benefit of the proposed multilingual multimodal pre-training.

The upper section shows the zero-shot results. Unlike cross-lingual transfer in NLP tasks, employing multilingual Transformers in vision-language tasks apparently does not generalize well across languages. For example, there is a significant drop in R@1 (19.9→11.1 (-44%) with mBERT, 21.0→16.3 (-24%) with XLM-R) when directly applying English-finetuned model to German→video search. For comparison, there is only a -10% degradation for XLM-R on $en \rightarrow de$ cross-lingual transfer in XNLI (Conneau et al., 2018). Multimodal (English-video) pre-training (MP) on HowTo100M only improves average R@1 (+0.1 or mBERT and +1.1 for XLM-R) compared to model-from-scratch. In contrast, our proposed multilingual multimodal pre-training (MMP) is shown to be the key facilitator for zero-shot cross-lingual transfer. MMP improves German→Video search (11.1→15.0, +35% for mBERT, and 16.3→19.4, +20% for XLM-R) and achieves $2.6 \sim 2.8$ improvement in average R@1. We attribute the effectiveness of MMP to

Model	*en*	*de*	*fr*	*cs*	*zh*	*ru*	*vi*	*sw*	*es*	Avg↑
mBERT	19.9	11.1	11.6	8.2	6.9	7.9	2.7	1.4	12.0	9.1
mBERT-MP	20.6	11.3	11.9	8.0	7.1	7.7	2.5	1.1	12.5	9.2
mBERT-MMP	21.8	15.0	15.8	11.2	8.4	11.0	3.7	3.4	15.1	11.7
XLM-R	21.0	16.3	17.4	16.0	14.9	15.4	7.7	5.7	17.3	14.7
XLM-R-MP	23.3	17.4	18.5	17.1	16.3	17.0	8.1	6.2	18.5	15.8
XLM-R-MMP	**23.8**	**19.4**	**20.7**	**19.3**	**18.2**	**19.1**	**8.2**	**8.4**	**20.4**	**17.5**
mBERT + translated VTT	19.6	18.2	18.0	16.9	16.2	16.5	8.4	13.0	18.5	16.1
mBERT-MMP + translated VTT	21.5	19.1	19.8	18.3	17.3	18.3	8.9	14.1	20.0	17.4
XLM-R + translated VTT	21.5	19.6	20.1	19.3	18.9	19.1	10.3	12.5	18.9	17.8
XLM-R-MMP + translated VTT	**23.1**	**21.1**	**21.8**	**20.7**	**20.0**	**20.5**	**10.9**	**14.4**	**21.9**	**19.4**

Table 4: **Recall@1 of multilingual text→video search on VTT.** Upper: Zero-shot cross-lingual transfer. Lower: Performance with synthesized pseudo-multilingual annotations for training. MMP: multilingual multimodal pre-training on Multi-HowTo100M. MP: Multimodal (English-Video) pre-training on HowTo100M.

Model	R@1↑	R@5↑	R@10↑
JSFusion (Yu et al., 2018)	10.2	31.2	43.2
JPoSE (Wray et al., 2019)	14.3	38.1	53.0
VidTrans (Korbar et al., 2020)	14.7	–	52.8
HT100M (Miech et al., 2019)	14.9	40.2	52.8
Noise (Amrani et al., 2020)	17.4	41.6	53.6
CE* [3] (Liu et al., 2019)	20.9	48.8	62.4
Ours(VTT:*en*-only)	21.0	50.6	63.6
Ours-MMP (VTT:*en*-only)	**23.8**	**52.6**	**65.0**

Table 5: English→video search performance on VTT. CE* uses 9K videos for training, while other baselines and our model use 6.5K videos for training.

Figure 3: English→video and Russian→video on VTT

learning improved alignments between multilingual textual and visual context in the shared embedding space, as relatively balanced improvements between English→video and non-English→video is observed with fine-tuning.

Fig 2 shows the trend of incrementally incorporating more languages in MMP. For XLM-R, improvement in R@1 asymptotically converges when pre-training with more multilingual text-video pairs. On the other hand, for zero-shot German→video search, pre-training with more languages keeps improving the search performance, even though the additional language (*e.g.* French) is different from the target language (*i.e.* German).

The lower section of Table 4 shows the results of models fine-tuned with (synthesized) pseudo-multilingual annotations. It can be regarded as

Model	English to Video			Chinese to Video		
	R@1↑	R@5↑	R10↑	R@1↑	R@5↑	R@10↑
VSE (Kiros et al., 2014)	28.0	64.3	76.9	-	-	-
VSE++ (Faghri et al., 2018)	33.7	70.1	81.0	-	-	-
Dual (Dong et al., 2019)	31.1	67.4	78.9	-	-	-
HGR (Chen et al., 2020a)	35.1	73.5	83.5	-	-	-
Ours (VATEX:*en*-only)	43.5	79.8	88.1	23.9	55.1	67.8
Ours-MMP (VATEX:*en*-only)	**44.4**	80.5	88.7	29.7	63.2	75.5
Ours-MMP (VATEX:*en*, *zh*)	44.3	**80.7**	**88.9**	**40.5**	**76.4**	**85.9**

Table 6: Multilingual text→video search on VATEX.

the *translate-train* scenario, which serves as a strong performance target for evaluating zero-shot cross-lingual transfer, as discussed in (Lample and Conneau, 2019; Hu et al., 2020). Both mBERT and XLM-R yield better performance across non-English languages with the in-domain translated pseudo-multilingual annotations. However, for English→video search, a 0.7 degradation is observed compared to the zero-shot setting. It is likely due to the noise in translated captions. Notably, there is still a performance gap between zero-shot and translate-train settings for models with mBERT. In contrast, the gap is much smaller for models with XLM-R. In the following sections, unless otherwise specified, we use our best model with XLM-R as the text backbone to compare with other baselines.

5.6 Comparison to Supervised State of the Art

English→Video Search on VTT. Table 5 shows the comparison of English→video models on VTT, where our model outperforms other baselines by a large margin. Essentially, our model achieves 8.9 R@1 improvement over the original HowTo100M model. Using a smaller set of visual features and training on a smaller (6,513 vs 9,000) training set, our model also outperforms CE (Liu et al., 2019) with or without pre-training. Fig. 3 shows an English→video and a zero-shot Russian→video search results with Ours-MMP. Our model retrieves

Model	M30K # lang.	English to Image R@1↑	English to Image R@5↑	English to Image R@10↑	German to Image R@1↑	German to Image R@5↑	German to Image R@10↑	Czech to Image R@1↑	Czech to Image R@5↑	Czech to Image R@10↑
OE†* (Vendrov et al., 2015)	2	21.0	48.5	60.4	25.8	56.5	67.8	-	-	-
VSE++* (Faghri et al., 2018)	2	31.3	62.2	70.9	39.6	69.1	79.8	-	-	-
Pivot† (Gella et al., 2017)	2	22.5	49.3	61.7	26.2	56.4	68.4	-	-	-
MULE (Kim et al., 2020)	4	42.2	72.2	81.8	35.1	64.6	75.3	37.5	64.6	74.8
SMALR (Burns et al., 2020)	10	41.8	72.4	82.1	36.9	65.4	75.4	36.7	68.0	78.2
MHA-D (Huang et al., 2019b)	2	50.1	78.1	85.7	40.3	70.1	79.0	-	-	-
Ours (M30K:en-only)	1	48.4	78.3	85.9	31.4	61.1	72.6	33.2	65.2	76.1
Ours-MMP (M30K:en-only)	1	50.0	79.2	86.8	33.8	63.3	74.7	37.9	68.8	78.2
Ours-MMP (M30K:en, de, cs, fr)	4	**51.6**	**80.1**	**87.3**	**45.1**	**75.6**	**85.0**	**46.6**	**75.9**	**83.4**

Table 7: Multilingual text→image search on Multi30K. MMP: Multilingual multimodal pre-training.

the correct videos and the other top-ranked videos share similar appearance to the correct one.

Multilingual Text→Video Search on VATEX. Table 6 summarizes English→video and Chinese→Video search performance on VATEX. Under the zero-shot setting where we train with only English-video pairs, our model already outperforms other baselines. However, a clear gap between English→video and Chinese→video is observed, indicating that cross-lingual transfer remains challenging even with XLM-R. With the proposed MMP, the gap is significantly closed by 5.8/8.1/7.7 in R@1/5/10. With the access to Chinese annotations, the performance of our model is further improved for both languages and our model achieves new state of the art performance.

Cross-Modality Transfer: From Video-Text to Image-Text on Multi30K. To extend our study on zero-shot cross-lingual transfer for image-text tasks, we investigate the feasibility of transferring our video-text model across modalities. We replace the 3D-CNN in the original video-text model with a 2D-CNN to encode the image. In practice, following MHA-D (Huang et al., 2019b), we utilize the Faster-RCNN (Ren et al., 2015) pre-trained in Visual Genome (Krishna et al., 2016) to extract regional visual features. Essentially, an image is encoded as $e_v = \mathbb{R}^{M \times H}$ where $M = 36$ is the maximum number of visual objects in an image. For models with MMP, we initialize their weights with the model pre-trained on Multi-HowTo100M. To tackle the feature mismatch between 2D-CNN and 3D-CNN, we leverage a linear layer with a doubled learning rate to map 2D-CNN features to the same dimension as 3D-CNN features.

Table 7 shows the results on Multi30K. For zero-shot cross-lingual transfer, when training from scratch (M30K:en-only), our model achieves comparable performance to MHA-D but lags in German→image search since it only uses English annotations. In Ours-MMP, pre-training improves all recall metrics even with modality gap. The average R@1 improvement is 3.2. A larger gain for (relatively) low-resource language such as Czech is observed. Without using any Czech annotations, our zero-shot model with MMP achieves comparable Czech→image search performance to SMALR (Burns et al., 2020), which uses 10 languages including Czech. However, when transferring across modalities and using only English annotations, there are performance gaps between English→Image and German/Czech→Image search, implying that transferring models across modalities is feasible but remains challenging. We consider zero-shot cross-modal cross-lingual transfer as our future work.

For a fair comparison with other baselines, when training with annotations in all 4 languages provided by Multi30K, our model greatly outperforms all baselines by large margins in multilingual text→image search.

6 Conclusion

We have presented a multilingual multimodal pre-training (MMP) strategy, the Multi-HowTo100M dataset, and a Transformer-based text-video model for learning contextualized multilingual multimodal representations. The results in this paper have convincingly demonstrated that MMP is an essential ingredient for zero-shot cross-lingual transfer of vision-language models. Meanwhile, there are many remaining challenges, such as resolving the performance gap between zero-shot and training with in-domain non-English annotations; as well as techniques to transfer varieties of vision-language models (*e.g.*, VQA (Goyal et al., 2017) or TVQA (Lei et al., 2020)). We believe the proposed methodology, and the corresponding resources we will release, will be an important first step towards spurring more research in this direction.

References

Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. 2016. Unsupervised learning from narrated instruction videos. In *Computer Vision and Pattern Recognition (CVPR)*.

Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. 2020. Noise estimation using density estimation for self-supervised multimodal learning. *arXiv preprint arXiv:2003.03186*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Jeremy Barnes, Lilja Øvrelid, and Erik Velldal. 2019. Sentiment analysis is not solved! assessing and probing sentiment classification. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 12–23, Florence, Italy. Association for Computational Linguistics.

Andrea Burns, Donghyun Kim, Derry Wijaya, Kate Saenko, and Bryan A. Plummer. 2020. Learning to scale multilingual representations for vision-language tasks. In *The European Conference on Computer Vision (ECCV)*.

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA. Association for Computational Linguistics.

Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020a. Fine-grained video-text retrieval with hierarchical graph reasoning. In *CVPR*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020b. A simple framework for contrastive learning of visual representations. In *ICML*.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Delvin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. 2019. Dual encoding for zero-example video retrieval. In *CVPR*.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics.

Akiko Eriguchi, Yoshimasa Tsuruoka, and Kyunghyun Cho. 2017. Learning to parse and translate improves neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 72–78, Vancouver, Canada. Association for Computational Linguistics.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image pivoting for learning multilingual multimodal representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2839–2845. Association for Computational Linguistics.

Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. 2019. Large-scale weakly-supervised pre-training for video action recognition. In *CVPR*.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390, Denver, Colorado. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019a. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. *arXiv preprint arXiv:1909.00964*.

Po-Yao Huang, Xiaojun Chang, and Alexander Hauptmann. 2019b. Multi-head attention with diversity for learning grounded multilingual multimodal representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1461–1467, Hong Kong, China. Association for Computational Linguistics.

Vladimir Iashin and Esa Rahtu. 2020. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 958–959.

Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, and Bryan A. Plummer. 2020. MULE: Multimodal Universal Language Embedding. In *AAAI Conference on Artificial Intelligence*.

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

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474.

Bruno Korbar, Fabio Petroni, Rohit Girdhar, and Lorenzo Torresani. 2020. Video understanding as machine translation.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, Online. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. 2019. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 13–23.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nicholas Johnston, Andrew Rabinovich, and Kevin Murphy. 2015. What's cookin'? interpreting cooking videos using text, speech and vision. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 143–152, Denver, Colorado. Association for Computational Linguistics.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6294–6305. Curran Associates, Inc.

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. In *ICCV*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

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

Mandela Patrick, Y. Asano, Ruth Fong, João F. Henriques, G. Zweig, and A. Vedaldi. 2020. Multi-modal self-supervision from generalized data transformations. *ArXiv*, abs/2003.04298.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. *arXiv preprint arXiv:2005.04816*.

Gunnar A. Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, João Carreira, Phil Blunsom, and Andrew Zisserman. 2020. Visual grounding in video for unsupervised word translation. In *CVPR*.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of*

the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 477–487, Montréal, Canada. Association for Computational Linguistics.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*.

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 *Advances in neural information processing systems*, pages 5998–6008.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4581–4591.

Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. 2019. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *ICCV*.

Shijie Wu and Mark Dredze. 2019a. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.

Shijie Wu and Mark Dredze. 2019b. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*.

I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. 2019. Billion-scale semi-supervised learning for image classification.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Shoou-I Yu, Lu Jiang, and Alexander Hauptmann. 2014. Instructional videos for unsupervised harvesting and learning of action examples. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, page 825–828, New York, NY, USA. Association for Computing Machinery.

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

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*.

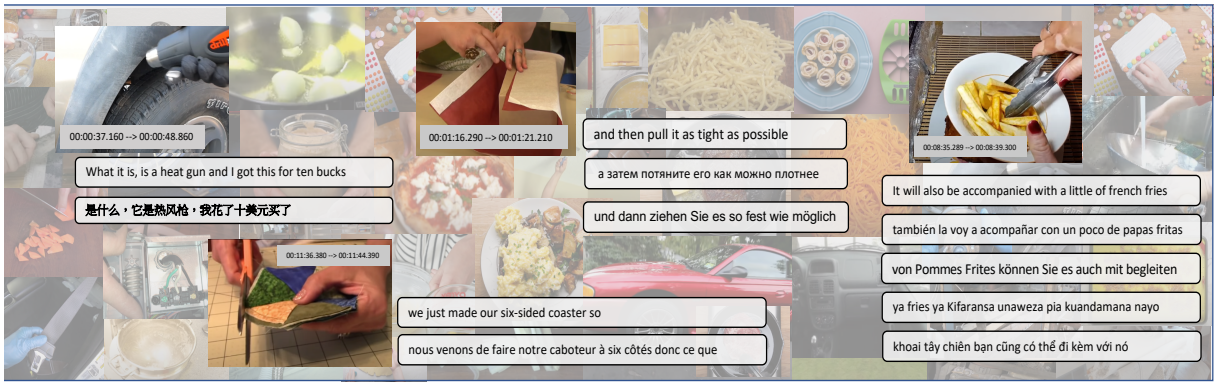


Figure 4: Video clips and the corresponding multilingual subtitles Multi-HowTo100M.

A Appendix Overview

This supplementary material is organized as the following: First we introduce the Multilingual HowTo100M dataset for multilingual multimodal pre-training in §B. Then we provide additional implementation details and experiment setup in §C. Additional ablation studies regarding choices of hyper parameters in our model architecture is discussed in §D. Then we provide and discuss additional experimental results in §E. Additional qualitative results on VTT can be found in §F.

B The Multilingual HowTo100M Dataset

In this section we provide the detailed statistics of the Multilingual HowTo100M (Multi-HowTo100M) dataset. We also provide a comparison to Sigurdsson et al. (2020) that also uses HowTo100M for unsupervised word translation.

The Multi-HowTo100M dataset is built upon the original English HowTo100M dataset (Miech et al., 2019) that contains 1.2 million instructional videos (138 million clips) on YouTube. We reuse the *raw* English subtitles in HowTo100M, where the subtitles in HowTo100M are either the automatic speech recognition (ASR) transcriptions or the user generated transcription. In most cases they are generated by Google ASR.

For Multi-HowTo100M, we use the same video collection as English HowTo100M. At the time of data collection, there were 1.09 million videos accessible. We collect the subtitles provided by YouTube, which either consist of user-generated subtitles or those generated by Google ASR and Translate in the absence of user-generated ones. Essentially, we collect video subtitles in 9 languages: English (*en*), German (*de*), French (*fr*), Russian (*ru*), Spanish (*es*), Czech (*cz*), Swahili (*sw*), Chinese (*zh*), Vietnamese (*vi*). Table 8 summarizes the dataset statistics for each language. In most cases there are more than 1 billion tokens a language.

Fig. 5 further shows the number of tokens per video. There are typically lengthy narrations that contains several hundreds of tokens available in each instructional video. Fig. 6 shows the distribution of number of tokens in a subtitle. For each subtitle segment, which ranges from 0 20 seconds, there are typically 15-25 words. The most of the cases, subtitles are well aligned in time for non-English languages. Fig. 4 visualizes a few examples in Multi-HowTo100M.

A similar HowTo100M variant has been recently reported (but not yet released) in MUVE (Sigurdsson et al., 2020) that is created for unsupervised word translation. Our Multil-HowTo100M differs from MUVE in the following perspectives: First, we collects 9 language for *all* videos in HowTo100M. MUVE only has 4 languages available (English, French, Japanese, and Korean) on HowTo100M. Also, MUVE divided HowTo100M into 4 non-overlapped sections for each language while there are parallel language pairs in Multi-HowTo100M. Essentially, there are more languages in Multi-HowTo100M (9 vs. 4) There are more than 1 billion tokens in most languages. To our best knowledge, our Multi-HowTo100M dataset is currently the largest multilingual text-video collection.

Beyond scale, instructional videos in Multi-HowTo100M are feasible resources for learning language-video models. Demonstrators in instructional videos typically perform intentionally and explain the visual object or action explicitly. According to the inspection by (Miech et al., 2019), for around 51% of clips, at least one object or action mention in the caption can be visually seen. Prior work has shown that instructional videos are useful for event recognition (Yu et al., 2014), action local-

Language	videos	#subtitle	#tokens
English	1238911	138429877	1.18B
German	1092947	69317890	1.26B
French	1093070	69399097	1.33B
Czech	1092717	68911940	1.22B
Russian	1092802	69117193	1.25B
Chinese	1092915	68939488	0.94B
Swahili	1092302	68898800	1.22B
Vietnamese	1092603	68887868	1.13B
Spanish	1092649	70143503	1.16B

Table 8: Multi-HowTo100M statistics

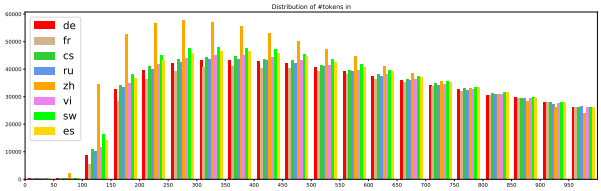


Figure 5: Distribution of #tokens/video in Multi-HowTo100M

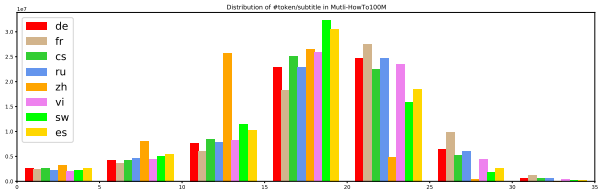


Figure 6: Distribution of #tokens/subtitle in Multi-HowTo100M

ization model (Alayrac et al., 2016), cross-modal alignments (Malmaud et al., 2015). We expect the previous success in the intersection of NLP and CV could be further translated into more languages to have a broaden impact to the world.

The are great potentials of using our Multi-HowTo100M dataset in related research field such as multilingual multimodal representation learning, multilingual multimodal translation, multilingual image/video captioning ... etc. We expect the release of Multi-HowTo100M will be a first step towards spurring more research in these directions.

C Implementation and Experiment Details

Pre-Processing For pre-possessing, we truncate the maximum length N of text to 192 for pre-training on Multi-HowTo100M or HowTo100M. The maximum length is set to 96 for fine-tuning VTT (Xu et al., 2016), VATEX (Wang et al., 2019) and Multi30K (Elliott et al., 2016). The maximum video length is set to 128 for pre-training on Multi-HowTo100M or HowTo100M and 36 for all fine-tuning tasks.

Model Architecture For the multilingual Transformers, either multilingual BERT (Delvin et al., 2018) or XLM-R-large (Artetxe et al., 2020), we use the pre-trained version provided by HuggingFace. [4] and use their corresponding tokenizers for tokenization. Detailed design choices regarding output layer and frozen layer is discussed in §D.

For the video backbone, we use a 34-layer, R(2+1)-D (Tran et al., 2018) network pre-trained on IG65M (Ghadiyaram et al., 2019) and a S3D (Miech et al., 2020) network pre-trained on HowTo100M (Miech et al., 2019). We apply a spatial-temporal average pooling over the last convolutional layer, resulting in a 512-dimensional vector for each 3D CNN network. We extract visual features at a rate of 1 feature per second. Since the 3D CNNs employs different size of input windows (*e.g.*, 8 frames for R(2+1)D and 16 for S3D), we re-sample videos to 30 fps and employs a window of size 8 or 30 that takes consecutive frames starting from the beginning of every second for encoding. We simply concatenate the two 3D-CNN outputs and use the 1024-dimension vector as the visual input stream to our model. Notably, instead of using 9 different types of visual features as in CE (Liu et al., 2019), we use only the above 2 features and achieve superior performance.

For the Transformer pooling head (TP), we use a 2-layer Transformer with 4-head attention for each TP module. The embedding dimension D is set to 1024. We do not use the positional embedding in both text and video TP as we do not find it beneficial. The softmax temperature in all NCE contrastive objectives is set to 0.1 as used in SimCLR (Chen et al., 2020b).

Training and Inference Details and Profiling. For the softmax temperature in NCE, we set to 0.1 as used in SimCLR (Chen et al., 2020b). We use the Adam (Kingma and Ba, 2015) optimizer with a initial learning rate $2 \cdot 10^{-4}$ and clip gradients greater than 0.2 during the training phase. Dropout rate is 0.3. Since the video length and token length is longer in the pre-training phase, we use a 64 batch size for pre-training. For fine-tuning, we use a batch size of 128.

Pre-training on 1.2 million HowTo100M videos takes around 10 GPU hours (NVIDA V100) for 16

[4] https://github.com/huggingface/transformers

epochs. We speed up the pre-training process by distributing the workload over 8 GPUs on a single server. We use 1 GPU for the fine-tuning or training from scratch experiments. For the MSR-VTT split, it takes 12 GPU hours to train our model on 180K video-text pairs for 20 epochs. For VATEX, it takes 32 GPU hours to train on 260K video-text pairs for 30 epochs. For inference, the encoding speed is around 250-300 videos/sec and 200-250 text queries/sec. The overall text→video search speed on 1,000 video-text pairs (1,000 text queries over 1,000 videos) is around 6 seconds including video/text encoding and ranking their similarity scores.

Experiment Details Our experiment consider three types of pre-training: (1) Multilingual multimodal pre-training (MMP), (2) Multimodal pre-training (MP), and (3) no pre-training (from scratch). For (1) and (2), we pre-train 16 epochs and use the model weight at 16-th epoch for fine-tuning experiments.

For multimodal pre-training, we pre-train on the original English HowTo100M dataset. We iterate over all videos in HowTo100M. For each video, we randomly sample the start and end time to construct a video clip. For each clip, we locate the nearest consecutive ASR transcriptions in time and use it as to construct the (video, text) pair for training.

For multilingual multimodal pre-training (MMP), we use Multi-HowTo100M for pre-trianing. For each video, we follow the same strategy as MP. For a clip, we sample one language type each time from 9 languages and use the consecutive ASR transcriptions that are closest in time to compose (video, text) pairs for training.

After the pre-training phase, we fine-tune our model on VTT or VATEX to evaluate on text→video search tasks. In the zero-shot cross-lingual transfer experiments, we use only English-video data. We then directly test the model with non-English queries to report the zero-shot performance. When annotations in additional languages are available (by humans in VATEX and Multi30K; by MT models (*i.e.* *translate-train*) in VTT), we train our model with all available multilingual annotations (*i.e.* fully supervised) to compare fairly with other baselines in multilingual text→video search.

Since pre-trained model has a faster convergence rate, we fine-tune for 10 epochs and use the model with best validation performance (summation of R@1, R@5, R@10) for testing. For models without pre-training (*i.e.*, from-scratch), we train for 20 epochs under the same training protocol.

D Additional Ablation Studies

As has been investigated in XTREME (Hu et al., 2020), choosing different output layers will affect the zero-shot transferability of multilingual Transformers in NLP tasks. For text→video search tasks, we conduct a series of experiments to identify the desirable hyper-parameter choices in the proposed multilingual multimodal Transformer that lead to best performance in English-to-video and (zero-shot) non-English-to-video search performance. Beyond our ablation studies in Sec.5, in this section we highlight our trials in the choice of the output layer and the layers to be frozen in our multilingual Transformer backbone (*i.e.* mBERT and XLM-R). There are 24 layers in XLM-R (large) and 12 layers in mBERT. We perform grid-search on VTT to identify the best choice of these two hyper-parameters.

D.1 Choosing the output layer and the layers to freeze in Multilingual Transformers

Table 9 and Table 10 compare different choice of hyper-parameters. The best output layer for mBERT and XLM-R is the 12-th layer. While output layer does not affect English→video search significantly, it greatly affects the zero-shot cross-lingual transfer performance. For both XLM-R and mBERT, the performance degrade significantly if fine-tuning all layers.

Meanwhile, we also observe that when freezing all layers (*i.e.* using the pre-extracted contextualized multilingual embeddings) does not lead to satisfactory results. For mBERT, R@1 drops from 19.9 to 18.9 in English→video search and 11.1 to 9.8 in German→video search. For XLM-R, R@1 drops from 21.0 to 18.9 in English→video search and 16.3 to 14.1 in German→video search. These results imply that text-only contextualized multilingual embeddings along are likely to be infeasible to be applied to vision-language tasks without proper fine-tuning.

An important tendency is that the best English→video search performance corresponds to the best German→video performance. This trend implies that for model selection, the configuration for the best English→video model usually translates to the best configuration for (zero-shot) cross-lingual model. This shared trend justifies the

Model	R@1	R@5	R@10
VSE (Kiros et al., 2014)	10.1	29.4	41.5
VSE++ (Faghri et al., 2018)	14.4	35.7	46.9
Dual (Dong et al., 2019)	13.7	36.1	48.2
HGR (Chen et al., 2020a)	16.4	38.3	49.8
Ours-Full	**24.0**	**50.5**	**62.1**

Table 12: Zero-shot generalization on YouTube2Text with VTT-finetuned model.

Output layer	Freeze lower	*en*	*de*
3	0	20.9	3.2
6	0	20.5	3.1
9	0	21.0	4.8
12	0	21.0	13.3
15	0	20.5	12.3
18	0	20.8	12.6
12	6	21.0	15.5
12	9	**21.0**	**16.3**
12	12	18.9	14.1

Table 9: Text→video R@1 of XLM-R output layers and layers to freeze on VTT

Output layer	Freeze lower	*en*	*de*
3	0	19.2	2.5
6	0	19.5	2.0
9	0	19.3	5.8
12	0	19.6	8.8
12	6	19.3	10.5
12	9	**19.9**	**11.1**
12	12	18.9	9.8

Table 10: Text→video R@1 of mBERT output layers and layers to freeze on VTT

text→video	English	Non-English
In-domain	✓	✓
Out-of-domain	✓	

Table 11: Coverage of our experiments

English→video ablation studies in the original paper. Note that we utilize the best English→video for all (zero-shot) cross-lingual experiment in our experiment section.

For multilingual text→video search, the best configuration we found in our experiments is to output the 12-th layer and freeze the layers below 9 for both mBERT and XLM-R.

E Additional Experimental Results

The coverage of our text→video search experiments is summarized in Table 11. Due to length constraint, we provide the additional experimental result in this supplementary material. Essentially, our experiments cover the following scenarios:

1. **In-domain, English**: Table 5 (VTT) and Table 6 (VATEX) in the original paper.
2. **In-domain, non-English**: Table 4 (VTT, 9 languages) and Table 6 (VATEX, Chinese).
3. **Out-of-domain, English**: We provide additional (zero-shot) generalization results across datasets in §E.1.
4. **Out-of-domain, non-English**: We consider as our future work.

E.1 Generalizability across English-Video Datasets

In this section. we provide additional experiment results regarding zero-shot generalization of the VTT-finetuned model on out-of-domain dataset. Specifically, we test on YouTube2Text (Chen and Dolan, 2011). The aim of this experiment is to test the cross-dataset generalizabilty of our model without using domain-specific training data.

Table 12 shows the comparison of English→video search results on the YouTube2Text testing set. Models in this table are only fine-tuned on VTT and use *no* YouTube2Text training data. As can be observed, our model with MMP generalizes well on YouTube2Text, outperforming HGR (Chen et al., 2020a) by 7.6 and DualEncoder (Dong et al., 2019) by 10.3 in R@1.

F Additional Qualitative Results

We provide addition qualitative multilingual text→video search results on VTT in Fig. 7. With a query in 9 possible languages, there is one and only one correct video to be retried out of the 1000 testing videos in VTT testing set. As can be observed, given a multilingual text query on top, in most cases, our model successfully retrieves the correct videos marked in green. Also, the top-ranked videos look semantically similar to the correct one.

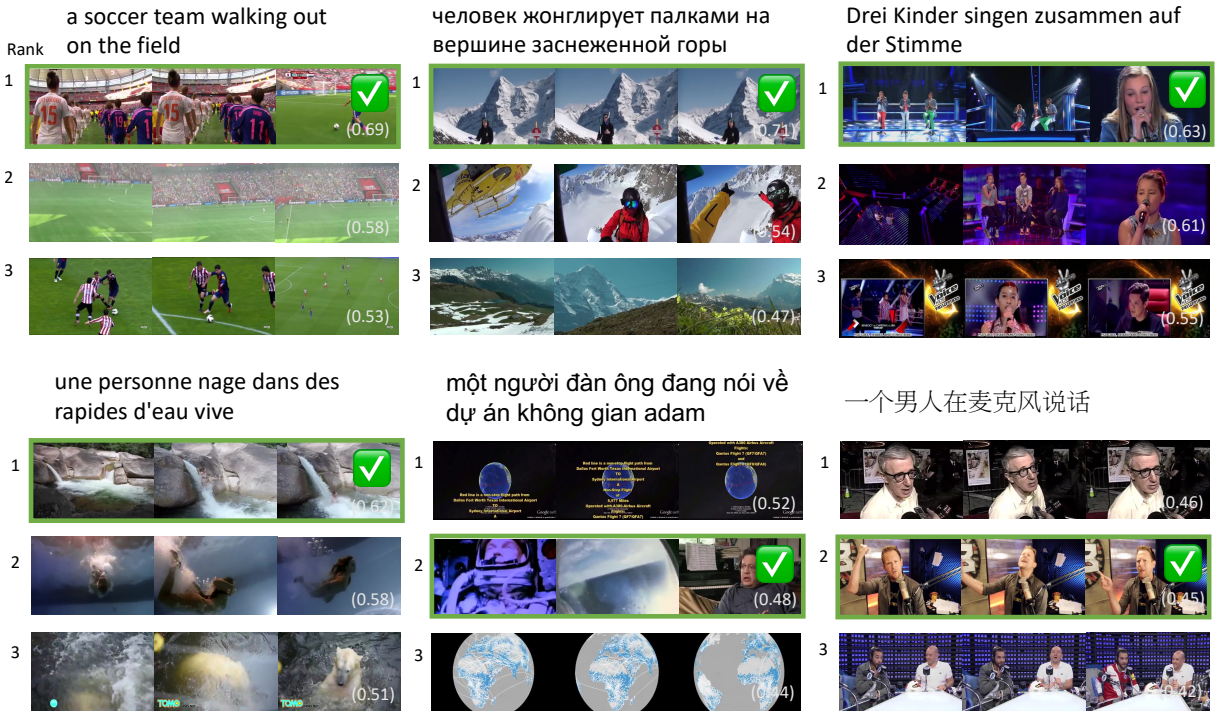


Figure 7: Qualitative examples of the top-3 multilingual text→video search results and cosine similarity scores on VTT. Only one correct video (colored in green) for each multilingual text query on the top.