VIDLANKD: Improving Language Understanding via Video-Distilled Knowledge Transfer

Anonymous Author(s)
Affiliation
Address
email

Abstract

Since visual perception can give rich information beyond text descriptions for world understanding, there has been increasing interest in leveraging visual grounding for language learning. Recently, vokenization [64] attracted attention by using the predictions of a text-to-image matching model as labels for language model supervision. Despite its success, the method suffers from approximation error of using finite image labels and lack of vocabulary diversity of a small image-text dataset. To overcome these limitations, we present VIDLANKD, a video-language knowledge distillation method for improving language understanding. We train a multi-modal teacher model on a video-text dataset, and then transfer its knowledge to a student language model on a text dataset. To avoid approximation error, we propose to use different knowledge distillation objectives. In addition, the use of a large-scale video-text dataset helps our models learn diverse and richer vocabularies. In our experiments, VIDLANKD achieves consistent improvements over text-only language models and vokenization models, on several downstream language understanding tasks including GLUE, SQuAD, and SWAG. We present comprehensive ablation studies to show the effectiveness of our proposed components. We also demonstrate the improved world knowledge and physical reasoning ability of our model by evaluating on GLUE diagnostics and PIQA datasets.

1 Introduction

Language learning can be aided by grounded visual cues, as they provide powerful signals for modeling a vastness of experiences in the world that cannot be documented by text alone [5; 27; 4]. While the recent trend of large-scale language model pre-training indirectly provides some world knowledge from text, most large text corpora (e.g., Wikipedia) do not provide enough multi-modal grounding information. Previous works have explored multiple ways of grounding language to visual information such as constructing a common vector space [36; 7] and supervising the model with token-wise generated vision labels [64]. However, the widely-used image-text datasets (e.g., MS COCO [45]) are much smaller than text-only corpora in terms of word counts and vocabulary diversity for language learning.

The recent method of 'vokenization' [64] is a promising initial step towards addressing this problem by supervising language models with weakly-aligned vision-language groundings. Firstly, an image-text matching model finds a corresponding image to each text token in a sentence, from pre-selected image candidates. Then a language model learns to predict the selected image (called 'voken') for each text token, out of the image candidates. This can be seen as a knowledge distillation (KD) process [31] from a vision-language grounding model to a language model. Although the voken classification task helps the language model to improve on natural language understanding (NLU) tasks, there exist several limitations: (1) images cannot represent words well that require

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

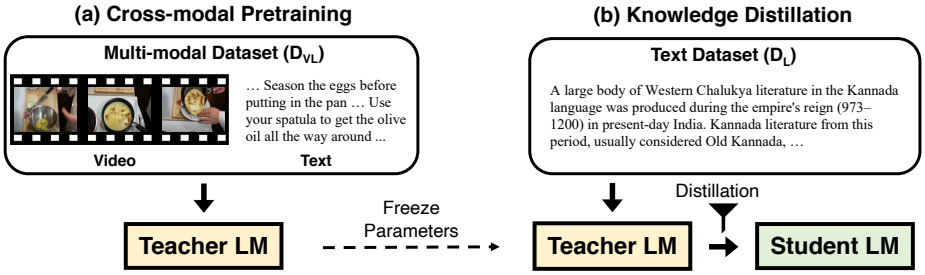

Figure 1: Overview of the proposed VIDLANKD method. We first pretrain a teacher language model on a multi-modal dataset (Sec. 3.2). Then we distill the knowledge of the teacher model (weights frozen) to a student language model on a text dataset (Sec. 3.3).

more activity-based and physical commonsense knowledge. (2) the voken supervision suffers from approximation/quantization error of the image-text matching model.

To address these problems, we propose VIDLANKD, a novel KD method based on video and language. Our teacher model consists of a video encoder and a language encoder. They are jointly trained with a video-language contrastive learning objective and a masked language modeling (MLM) objective. Then, we transfer the knowledge of a teacher language model to a student language model by minimizing the difference between contextualized text representations of two models on a text dataset (see Fig. 1). For this, we propose to use different KD objectives including neuron selectivity transfer (NST) [32] and contrastive representation distillation (CRD) [66] that avoid the approximation error from voken assignments [64]. For cross-modal pretraining of our teacher model, we use HowTo100M, a large-scale video dataset which has more diverse vocabulary and richer world commonsense and physical reasoning knowledge compared to MS COCO dataset.

In our experiments, student language models learned with the proposed video-language KD objectives outperform the baseline text-pretrained language models and the models distilled with vokenization, on several diverse natural language understanding benchmarks including GLUE [68], SQuAd [57], and SWAG [74]. Additionally, we show comprehensive ablation studies on video encoders, student KD objectives, teacher pretraining objectives, and video vs. image-based pretraining. We also conduct analyses to measure the linguistic world commonsense knowledge and physical reasoning ability learned from video, by evaluating our models on GLUE diagnostic dataset [68] and PIQA [6], and show improved performance on these abilities.

Overall, our contributions are: (1) a novel cross-modal knowledge distillation method for improving natural language understanding, (2) studying different knowledge distillation methods and using rich video-text data that can overcome the limitations of vokenization, and (3) empirical improvements on several language understanding benchmarks and comprehensive analyses on the effectiveness of the proposed components and physical reasoning and commonsense knowledge learned from videos.

2 Related Work

2.1 Knowledge Distillation

Knowledge distillation (KD) [31] is the process of transferring knowledge from a teacher model to a student model. Knowledge distillation has been successfully used in a wide range of applications, such as machine translation [38], visual recognition [29], speech recognition [10], and recommendation systems [65]. Due to its usefulness, many works advanced the field of knowledge distillation by proposing new architectures [72; 75; 1; 52] and objectives [32; 14].

While many KD works study the problem of knowledge transfer within the same modality, cross-modal knowledge distillation [25; 20; 66] tackles the knowledge transfer across different modalities. Gupta et al. [25] transfers the knowledge of a model trained with RGB images to another model for depth and optical flow images. Do et al. [20] proposes a KD method for visual question answering [2], where the trilinear (image-question-answer) relational representation of a teacher model is transferred to a bilinear (image-question) student model. Tian et al. [66] combines contrastive learning and knowledge distillation to improve the knowledge transfer between different modalities.

Our VIDLANKD transfers the knowledge of a multi-modal teacher model learned from a video dataset to a student language model that tackles natural language understanding tasks.

2.2 Language Pre-Training

Large-scale pre-training of contextualized language models has seen huge success in natural language processing in recent years. ELMo [53] proposes to pre-train and fine-tune a large recurrent language model, which improves performance on a diverse set of downstream natural language processing tasks. BERT [19] improves the scalability of the pre-train/fine-tune framework by using a transformer [67] language model with a masked language modeling objective. Since then, pre-training of transformer language models has been extensively explored [46; 73; 42; 21; 60; 55; 16] for various natural language understanding [57; 69; 74; 68] and generation tasks [23; 58; 57; 59].

2.3 Multi-modal Pre-Training

Following the success of language pre-training with transformer models, pre-training of image-text [63; 48; 13; 77; 43] and video-text [62; 51; 79; 50; 44] multi-modal transformers have achieved improvements on numerous multi-modal downstream tasks [2; 71; 78]. The multi-modal transformers take both visual and textual inputs and are pre-trained on image-text or video-text pairs with multi-modal masked language modeling objectives. Despite the success on multi-modal downstream tasks, Tan and Bansal [64] finds that the multi-modal pretraining does not improve (and sometimes even harms) the language understanding performance. This is because the scale and diversity of text vocabulary of image-text and video-text datasets are usually smaller than those of text datasets. To utilize the rich vocabulary of text dataset, our VIDLANKD transfers the knowledge of pretrained multi-modal model to a student language model on a large text dataset.

2.4 Visually-Grounded Language Learning

A line of works explore using visual information to aid language understanding and generation tasks including co-reference resolution [40; 15], machine translation [76], bilingual lexicon learning [37], and visually-supervised language model: vokenization [64], which is closest to our work. Vokenization proposes to supervise a language model to predict a visualized token, called 'voken' for each input text token. Vokens are obtained by a contextualized token-to-image matching model, pretrained on a MS COCO image captioning dataset [12]. In this work, we experiment with alternative objectives which avoid the approximation error from finite voken assignments. In addition, we use HowTo100M [51] video dataset, which provides a more diverse vocabulary as well as richer world commonsense and physical reasoning knowledge.

3 Video-Language Knowledge Distillation

3.1 Method Overview

We aim to learn a better language representation with the help of visual supervision. For this, we leverage two kinds of datasets: the aligned muti-modal dataset, D_{VL}: $\{(\mathbf{x}, \mathbf{v})\}$ (e.g., HowTo100M [51]); and the text dataset, D_{L}: $\{\mathbf{x}\}$ (e.g., Wikipedia), where $\mathbf{x}$ is a sentence and $\mathbf{v}$ is a video paired with $\mathbf{x}$. Our knowledge transfer is done in two stages: (1) cross-modal pretraining of a teacher model, M^T, on multi-modal data D_{VL} (Eq. 1) (2) distilling the knowledge of teacher model to a student model, M^S, on text data D_{L} (Eq. 2). We illustrate our two-stage knowledge transfer method in Fig. 1.

$$\min_{\theta^T} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{v}\sim D_{\mathrm{VL}}} \mathcal{L}^T(\mathrm{M}^T, \mathbf{x}, \mathbf{v}) \tag{1}$$

$$\min_{\theta^S} \mathop{\mathbb{E}}_{\mathbf{x}\sim D_{\mathrm{L}}} \mathcal{L}^{KD}(\mathrm{M}^T, \mathrm{M}^S, \mathbf{x}) \tag{2}$$

Our teacher model M^T consists a language model LM^T and a visual encoder V^T. Both LM^T and V^T have transformer [67] architectures, where LM^T takes text tokens $\mathbf{x}$ and V^T takes video frame features $\mathbf{v}$ as inputs. We use off-the-shelf image encoders to encode video frames (see also Sec. 4.2). Our student model M^S is a transformer language model LM^S sharing the same architecture with teacher language model LM^T. As illustrated in Fig. 1(a), we first train teacher models LM^T and V^T

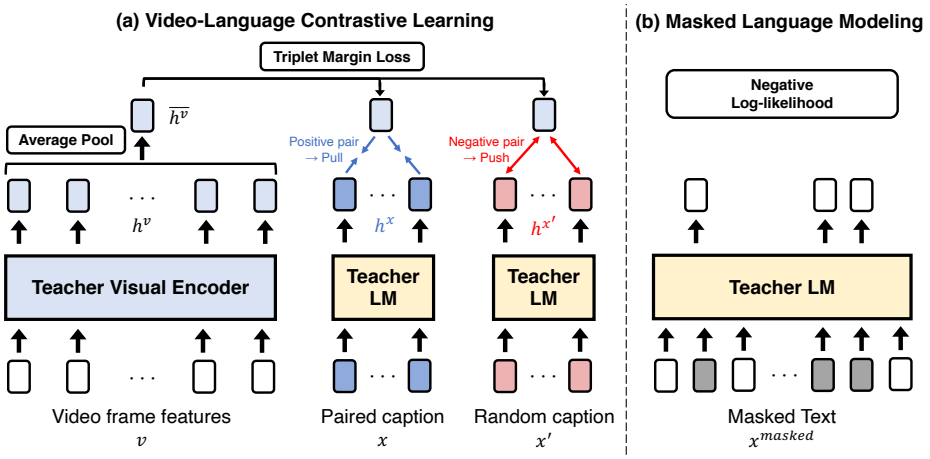

Figure 2: Cross-modal pretraining of our teacher model on a multi-modal dataset (Sec. 3.2). We train our teacher model with (a) video-language contrastive learning and (b) masked language modeling. For video-language contrastive learning, we only illustrate the negative sampling for text for brevity.

with contrastive learning and masked language modeling. Then, we transfer the knowledge of teacher models to student model LM^S as in Fig. 1(b). In the following subsections, we discuss the detailed training procedure of teacher (Sec. 3.2, Fig. 2) and student models (Sec. 3.3, Fig. 3).

3.2 Teacher Model

We train our teacher model on a multi-modal dataset with two objectives: video-language contrastive learning (Fig. 2(a)) and masked language modeling (Fig. 2(b)): $\mathcal{L}^T = \mathcal{L}_{CT} + \mathcal{L}_{MLM}$

Architecture. As shown in Figure 2, our teacher model M^T consists of a language encoder LM^T and a visual encoder V^T. Both LM^T and V^T have similar transformer architecture.[1] For each sentence $\mathbf{x}$, we tokenize it and append a special token `[CLS]` that represents the entire sentence following Devlin et al. [19]. LM^T takes $\mathbf{x}$ and outputs contextualized representation $h^x = \{h^x_{\texttt{[CLS]}}, h^x_1 \cdots h^x_{|x|}\}$. For each video $\mathbf{v}$, we extract frame-level features with an off-the-shelf image encoder (see more details in Sec. 4.2). Note that the parameters of the image encoder are not updated to save computation. We feed the frame features $\{v_1 \cdots v_{|v|}\}$ to our visual encoder V^T to get contextualized video frame features $h^v = \{h^v_1 \cdots h^v_{|v|}\}$. We get the final video representation $\bar{h}^v$ by temporally averaging frame-level features: $\bar{h}^v = \frac{1}{|v|}\sum_{i=1}^{|v|} h^v_i$. Following Tan and Bansal [64], both LM^T and V^T parameters are trained from scratch.

Video-Language Contrastive Learning. To learn multi-modal grounding, we use a contrastive learning objective that encourages the distance between representations of an aligned video-text pair to be closer than unaligned pairs, as shown in Fig. 2(a). For each $\mathbf{x}$, we randomly sample another text $\mathbf{x}'$ from its batch where $\mathbf{x}' \neq \mathbf{x}$. Similarly, for each $\mathbf{v}$, we randomly sample another video $\mathbf{v}'$ from its batch where $\mathbf{v}' \neq \mathbf{v}$. Then, we use hinge loss $\max\{0, \alpha - pos + neg\}$ with cosine similarity:

$$\mathcal{L}_{CT}(\mathbf{x}, \mathbf{x}', \mathbf{v}, \mathbf{v}') = \sum_{i}^{|\mathbf{x}|}[\max\{0, \alpha - \cos(\mathbf{h}^x_i, \bar{\mathbf{h}}^v) + \cos(\mathbf{h}^{x'}_i, \bar{\mathbf{h}}^v)\} \tag{3}$$

$$+ \max\{0, \alpha - \cos(\mathbf{h}^x_i, \bar{\mathbf{h}}^v) + \cos(\mathbf{h}^x_i, \bar{\mathbf{h}}^{v'})\}] \tag{4}$$

where α is the margin between the similarities within a positive pair and a negative pair. Different from previous models [7; 30] that exploits sentence-level contrastive loss, we follow [64] to construct a token-level contrastive loss that grounds the visual information to each contextualized token output. This fine-grained contrastive loss will help the token-level knowledge distillation in Sec. 3.4.

[1] In our experiments, we use the BERT architecture with two different configurations: 12 layers/768 hidden dimensions ($\mathrm{BERT}_{12L/768H} = \mathrm{BERT}_{BASE}$) and 6 layers/512 hidden dimensions ($\mathrm{BERT}_{6L/512H}$).

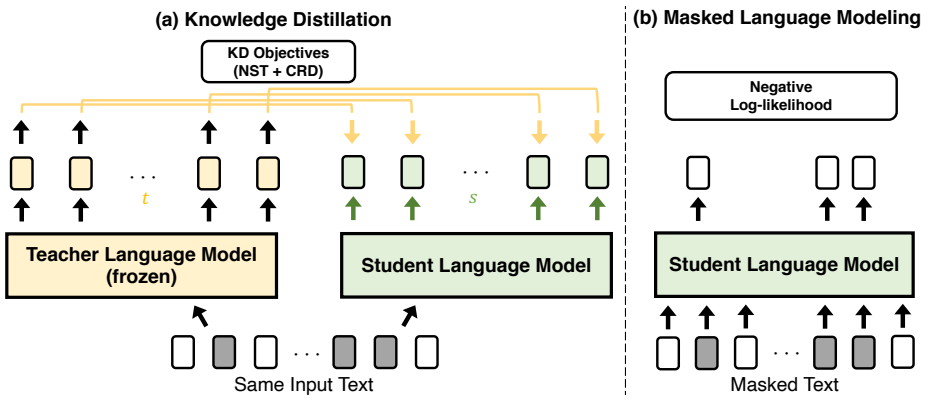

Figure 3: Illustration of our knowledge distillation from teacher language model LM^T to student language model LM^S on a text dataset (Sec. 3.3). We train our student model with (a) knowledge distillation objectives and (b) masked language modeling.

Masked Language Modeling. For textual reasoning and cross-modal grounding, we follow BERT [19] to use masked language modeling (MLM) objective (Fig. 2(b)). By replacing 15% of $\mathbf{x}$ tokens with a special token `[MASK]`, we obtain a masked text $\mathbf{x}^{\text{masked}}$ with the same length. The model takes $\mathbf{x}^{\text{masked}}$ as input and predicts the masked tokens. Then we minimize the negative log-likelihoods: $\mathcal{L}_{\text{MLM}}(\mathbf{x}, \mathbf{x}^{\text{masked}}) = -\sum_{i \in M} \log p(\mathbf{x}_i \mid \mathbf{x}^{\text{masked}})$, where M refers to the indices of masked tokens.

3.3 Student Model

After we train a teacher model on a multi-modal dataset, we transfer its knowledge to a student model on a text dataset. Following Kim and Rush [38], we train our student model with a sum of masked language modeling and two knowledge distillation objectives, NST and CRD (see Sec. 3.4):

$$\mathcal{L}^{\mathcal{S}} = \mathcal{L}_{\text{MLM}} + \mathcal{L}^{\text{KD}}_{\text{NST}} + \mathcal{L}^{\text{KD}}_{\text{CRD}} \tag{5}$$

Architecture. As shown in Fig. 3(a), our student model M^S is a language model LM^S with the same transformer architecture as the teacher language model LM^T. We train LM^S from scratch. Following previous works [11; 66], we introduce a multi-layer perceptron (MLP) distillation head on top of the last hidden states of LM^S. In our ablation study in appendix, we find that adding a distillation head slightly improves the distillation performance.

3.4 Knowledge Distillation Objectives

We next describe the knowledge distillation (KD) objectives used to transfer knowledge from the teacher model LM^T (Sec. 3.2) to this student model LM^S. Note that the weights of the teacher model LM^T are frozen during the knowledge distillation process since the teacher model should not be affected by the student model's performance. Following Kim and Rush [38], we use the knowledge distillation objective combined with the MLM objective (Fig. 3(b)). Concretely, we use the same input text mask for MLM and KD objectives. While we calculate the MLM loss only on masked positions, we calculate KD losses using all hidden states following Clark et al. [16] (Fig. 3).

We study the following KD objectives: Soft-label [31], L2 Regression [3], Neuron Selectivity Transfer (NST) [32], Contrastive Representation Distillation (CRD) [66], and Vokenization [64]. In our experiments comparing different KD objectives (Table 3), NST and CRD perform best, and the combination of them improved the performance even further. Therefore, we propose NST+CRD for our cross-modal knowledge distillation objective.

Soft Label: Hinton et al. [31] proposed a knowledge transfer method by regarding a teacher model prediction as a 'soft label' with temperature scaling. We minimize cross-entropy between $P^T(y|x)$

and $P^S(y|x)$, the word output probabilities of LM^T and LM^S given input text $\mathbf{x}$ respectively:

$$\mathcal{L}^{\text{KD}}_{\text{soft-label}}(\mathbf{x}) = -\sum_{i=1}^{|\mathbf{x}|}\sum_{z\in Z} P^T(y_i = z|\mathbf{x}) \log P^S(y_i = z|\mathbf{x}) \tag{6}$$

where Z is the word vocabulary. Following Hinton et al. [31], we divide the softmax logits of LM^T and LM^S by a temperature parameter $\tau = 2.0$. Note that for soft label KD, we reuse the LM head, instead of learning an additional distillation head.

L2 Regression: Following Ba and Caruana [3] which used feature regression for KD, we calculate the L2 distance between $\mathbf{s}(\mathbf{x})$ and $\mathbf{t}(\mathbf{x})$, the last hidden states of LM^T and LM^S given input text $\mathbf{x}$:

$$\mathcal{L}^{\text{KD}}_{\text{Regression}}(\mathbf{x}) = \sum_{i=1}^{|\mathbf{x}|} \|\mathbf{s}(\mathbf{x})_i - \mathbf{t}(\mathbf{x})_i\|_2^2 \tag{7}$$

Neuron Selectivity Transfer (NST): NST [32] is a KD method that transfers heatmap like spatial activation patterns of teacher neurons to student neurons. We transfer the sequential activation patterns of $\mathbf{t}(\mathbf{x}) \in \mathbb{R}^{|\mathbf{x}|\times d}$ to $\mathbf{s}(\mathbf{x}) \in \mathbb{R}^{|\mathbf{x}|\times d}$, where $\mathbf{t}(\mathbf{x})$ and $\mathbf{s}(\mathbf{x})$ are the last hidden states of LM^T and LM^S given input text $\mathbf{x}$, and d is the hidden state dimension (# neurons). Following Huang and Wang [32], we use the squared maximum mean discrepancy (MMD) [24] with kernel trick to measure the distance between the activation patterns of student neurons $\{\mathbf{s}(\mathbf{x})_{\cdot i}\}_{i=1}^d$ and teacher neurons $\{\mathbf{t}(\mathbf{x})_{\cdot j}\}_{j=1}^d$:

$$\begin{aligned}\text{MMD}^2(\mathbf{x}) = &\frac{1}{d^2}\sum_{i=1}^{d}\sum_{i'=1}^{d} k\left(\mathbf{s}(\mathbf{x})_{\cdot i}, \mathbf{s}(\mathbf{x})_{\cdot i'}\right) + \frac{1}{d^2}\sum_{j=1}^{d}\sum_{j'=1}^{d} k\left(\mathbf{t}(\mathbf{x})_{\cdot j}, \mathbf{t}(\mathbf{x})_{\cdot j'}\right) \\ &- \frac{2}{d^2}\sum_{i=1}^{d}\sum_{j=1}^{d} k\left(\mathbf{s}(\mathbf{x})_{\cdot i}, \mathbf{t}(\mathbf{x})_{\cdot j}\right)\end{aligned} \tag{8}$$

where we use Gaussian kernel for $k(\boldsymbol{s}, \boldsymbol{t}) = \exp\left(-\frac{\|\boldsymbol{s}-\boldsymbol{t}\|_2^2}{2\sigma^2}\right)$ with $\sigma = 1$. We transfer the teacher activation patterns to the student by minimizing squared MMD: $\mathcal{L}^{\text{KD}}_{\text{NST}}(\mathbf{x}) = \text{MMD}^2(\mathbf{x})$

Contrastive Representation Distillation (CRD): CRD [66] is a KD objective which maximizes the mutual information between the teacher and student representations with contrastive learning. Let's denote $\mathbf{s} \in S$ and $\mathbf{t} \in T$ as student and teacher representations given $\mathbf{x}$. We are given 1 positive pair (drawn from the joint distribution) for every N (batch size) negative pairs (drawn from the product of marginals; independent randomly drawn inputs from T and S). Following [66], we maximize the lower bound of mutual information between $\mathbf{s}$ and $\mathbf{t}$ by minimizing the following term:

$$\begin{aligned}\mathcal{L}^{\text{KD}}_{\text{CRD}}(\mathbf{x}) &= -\mathbb{E}_{q(\mathbf{s},\mathbf{t}|\text{postive})}[\log h(\mathbf{s},\mathbf{t})] - N \cdot \mathbb{E}_{q(\mathbf{s},\mathbf{t}|\text{negative})}[\log(1 - h(\mathbf{s},\mathbf{t}))] \\ h(\mathbf{s},\mathbf{t}) &= \frac{\exp\left(f_1(\mathbf{s})^\top f_2(\mathbf{t})\right)}{\exp\left(f_1(\mathbf{s})^\top f_2(\mathbf{t})\right) + \frac{N}{M}}\end{aligned} \tag{9}$$

where M is the cardinality of the dataset, f_1, f_2 are learned linear layers followed by $L2$ normalization, which map the student and teacher representations into a same feature space. Since a large N leads to a tight mutual information lower bound, following [66], we implement a memory buffer that stores the latent features of each data sample computed from previous batches. Therefore, during training we can efficiently retrieve a large number of negative samples from the memory buffer. Note that since CRD is based on contrastive learning, it is the only KD objective where student and teacher language models can take different inputs.

Vokenization: Vokenization [64] could be viewed as a knowledge distillation method, where token-level text-to-image retrieval results (called 'vokens') of a multi-modal retrieval model are used as labels for a student language model. For i-th input token $\mathbf{x}_i$, we calculate cosine similarity between i-th teacher language model hidden state $\mathbf{t}(\mathbf{x})_i$ and a video feature $\mathbf{v}$. Out of 30K pre-selected videos, we select a video that maximizes cosine similarity and use it as the voken for $\mathbf{x}_i$. By denoting the

voken of $\mathbf{x}_i$ as $\mathbf{voken}_i$, we formulate our vokenization-based KD objetive as:

$$\mathcal{L}_{\text{Voken}}^{\text{KD}}(\mathbf{x}) = -\sum_{i=1}^{|\mathbf{x}|} \log P_{\text{voken}}^{S}(y_i = \mathbf{voken}_i|\mathbf{x}) \tag{10}$$

where $P_{\text{voken}}^{S}(y|\mathbf{x})$ is the voken classification probabilities of LM^S given input text $\mathbf{x}$. We experiment with vokenization-based KD by retrieving vokens from images and videos (see Table 4 of Sec. 5.2).

4 Experimental Setup

4.1 Datasets

Video-Text Dataset. We use HowTo100M [51] for cross-modal pretraining of our teacher model (Sec. 3.2). HowTo100M has 1.22M videos totaling 136M video clips with total duration of 134,472 hours describing over 23K different visual tasks. There are 138M captions, 568M tokens with 633K distinct tokens. Please see appendix for video voken sampling.

Text Pre-training Dataset. To transfer the knowledge from our teacher language models to student language models (Sec. 3.3), we follow Tan and Bansal [64] to use English Wikipedia. For ablation studies (Sec. 5.2), we use Wiki103 [49], a widely used subset of English Wikipedia. There are 2.9B tokens and 120M sentences in English Wikipedia, and 111M tokens and 4.2M sentences in Wiki103.

Text Downstream Dataset. Following Tan and Bansal [64], we finetune our models on GLUE [68], SQuAD [57] 1.0 and SQuAD2.0 [56], and SWAG [74] to assess the pre-training performance. Since some smaller tasks in GLUE are reported as unstable in recent papers [64], we evaluate on the four largest tasks of GLUE: SST-2 [9], QNLI [57], QQP [33], and MNLI [69]. In addition, we also evaluate our models on the GLUE diagnostics [68] and PIQA [6] to measure its linguistic sub-category, world knowledge, and physical reasoning abilities.

4.2 Video Feature Representations

Following Miech et al. [51], we encode video features by concatenating features from a 2D frame-level image encoder and a 3D video encoder in channel dimension. Note that the parameters for 2D image encoder and 3D video encoder are not updated. The 2D image encoder outputs features for each frame individually. We experiment with different 2D image encoders: ResNet-152 [28] pretrained on ImageNet-1K [18] and CLIP [54] image encoder (ViT-B/32 [22]). For the 3D video encoder, we use 3D-ResNeXt-152 [70; 26; 34] trained on combination of ActivityNet [8], Kinetics [35], UCF-101 [61], and HMDB-51 [41]. See appendix for more details on video features.

4.3 Implementation Details

For the student distillation head, we use a two-layer MLP with ReLU activation. For both student and teacher language models, following previous works [46; 17; 64], we truncate input text that is longer than 128 tokens. We truncate videos features that are longer than 512 frames. We use an AdamW [39] optimizer with learning rate 2e-4 with weight decay [47] of 0.01. We reserve 10K samples of the HowTo100M dataset as validation data. We train the teacher model until it converges on validation data. For downstream tasks, we report the results on the validation sets. We train 3 epochs with a learning rate of 1e-4 and a batch-size of 32 for all downstream tasks. We use hinge loss margin $\alpha = 1.0$ for $\mathcal{L}_{\text{CT}}$ (Eq. 4). See more implementation details in appendix.

5 Results and Analysis

5.1 Primary Downstream Task Results

In the first block of Table 1, we include the image-based vokenization (Img-Voken) and their text-only pretrained baseline from Tan and Bansal [64]. Given our reproduced text-only baseline shows a similar average performance (78.9 vs 78.6), our student models distilled from NST+CRD are much better (82.6 vs 81.4). We discuss the comparison between video-based and image-based KD in detail in the following ablation study in comparison to vokenization (Table 4).

Table 1: Cross-modal knowledge distillation results of $\text{BERT}_{12L/768H}$ student language model on 7 downstream NLU tasks. In the first block, we include the image-based vokenization (Img-Voken) and its text-only pretrained baseline performance from Tan and Bansal [64]. In the second block, we compare our cross-modal KD method (NST+CRD) to video-based vokenization (Vid-Voken) and a text-only pretrained baseline. †EM refers to 'exact match'.

	SST-2 Acc	QNLI Acc	QQP Acc	MNLI Acc	SQuAD v1.1 EM†	SQuAD v2.0 EM	SWAG Acc	Avg.
$\text{BERT}_{12L/768H}$ [64]	89.3	87.9	83.2	79.4	77.0	67.7	65.7	78.6
+ KD (Img-Voken) [64]	92.2	88.6	88.6	82.6	78.8	68.1	70.6	81.4
$\text{BERT}_{12L/768H}$	89.0	88.0	86.2	79.2	77.2	68.0	65.0	78.9
+ KD (Vid-Voken) w/ ResNet	93.4	89.2	88.7	83.0	78.9	68.7	70.0	81.7
+ KD (Vid-Voken) w/ CLIP	94.1	**89.8**	89.0	83.9	79.2	68.6	71.6	82.3
+ KD (NST+CRD) w/ ResNet	94.2	89.3	89.7	84.0	79.0	**68.9**	71.8	82.4
+ KD (NST+CRD) w/ CLIP	**94.5**	89.6	**89.8**	**84.2**	**79.6**	68.7	**72.0**	**82.6**

In the second block of Table 1, we compare our proposed cross-modal KD method (NST+CRD) to video-based vokenization (Vid-Voken) and a non-KD baseline ($\text{BERT}_{12L/768H}$) which is only pretrained on text. We can see both cross-modal KD methods (i.e., KD and Vid-Voken) significantly outperform the text-only baseline across all 7 downstream tasks. We also experiment with different 2D frame encoders (Sec. 4.2): ResNet and CLIP. For both Vid-Voken and NST+CRD, we observe CLIP further improves the performance results over ResNet, indicating using a strong visual encoder helps the teacher training and thus benefits the knowledge distillation.

5.2 Ablation Studies

In this section, we conduct a comprehensive ablation study to show the effectiveness of our proposed methods. For all ablation experiments, we use $\text{BERT}_{6L/512H}$ architecture for student and teacher language models, we use use ResNet-152 for 2D frame encoder and 3D-ResNeXt-152 for 3D frame encoder, and we use Wiki103 [49] for student model training. We also include ablation experiments on the effect of teacher training objectives and additional distillation head in appendix.

Text-only Pretraining We saw our cross-modal KD improves the performance on downstream NLU tasks significantly (Sec. 5.1). Where does the improvement come from, video or text? To answer this question, we conduct text-only pretraining of $\text{BERT}_{6L/512H}$ on Wiki103, HowTo100M captions and compare them to a no-pretrain baseline. In Table 2, while both pretrained models improve the performance over the no-pretrain baseline, Wiki103-trained model outperforms HowTo100M-trained model (which has more tokens) significantly. Since HowTo100M captions are obtained by automatic speech recognition (ASR), they show considerable amount of grammar and transcription errors, compared to Wiki103 which is written in more formal language. This indicates that our KD methods improve NLU performance because of multimodal grounding, instead of just the larger corpus.

Table 2: Text-only pretraining results of $\text{BERT}_{6L/512H}$ pretrained on Wiki103, HowTo100M captions, and no-pretrain baseline.

Pre-trained on	SST-2	QNLI	QQP	MNLI
No-Pretrain	79.6	61.5	72.7	61.6
Wiki103 (Formal language)	**88.8**	**84.9**	**85.3**	**77.4**
HowTo100M (ASR captions)	83.3	78.5	83.7	71.5

KD Objectives Comparison. In Table 3, we compare different knowledge distillation objectives introduced in Sec. 3.4. The student models trained with NST [32] and CRD [66] show the best finetuning performance on downstream tasks. When combining NST and CRD, performance further improves with marginal additional computation cost, hence we propose to use NST+CRD for our cross-modal knowledge distillation.

Table 3: Ablation of knowledge distillation objectives.

	SST-2	QNLI	QQP	MNLI
$\text{BERT}_{6L/512H}$	88.8	84.9	85.3	77.4
+KD-Soft label	87.2	84.4	86.4	76.6
+KD-Regression	88.8	84.8	87.1	78.1
+KD-Vid Voken	89.7	85.5	86.5	77.8
+KD-NST	**91.1**	85.0	**87.4**	**78.4**
+KD-CRD	90.0	**85.5**	87.3	78.3
+KD-NST+CRD	**91.5**	**85.8**	**87.4**	**78.7**

Comparison to Vokenization. In Table 4, we compare NST [32] and Vokenization [64] in both image and video-level teacher model supervision. For video-level supervision, we provide our visual encoder with the combination of 2D and 3D video features (Sec. 4.2). For image-level supervision, we provide our visual encoder only with 2D features of the middle frame for each video clip. On image-level supervision (first block), Vokenization and NST show comparable performance. On video-level supervision (second block), NST outperforms Vokenization on 3 out of 4 tasks. The gap in the video domain might come from voken approximation error, where each image or video input is approximated with one of 30K predefined vokens. Since videos usually contain more diverse contents than images, the voken approximation error would be amplified in video-level supervision, whereas our NST distillation avoids this issue.

Table 4: Comparison between vokenization (Voken) and NST with image and video-level supervision.

	SST-2	QNLI	QQP	MNLI
$\text{BERT}_{6L/512H}$	88.8	84.9	85.3	77.4
+KD-Voken (Image)	89.3	84.4	86.0	77.5
+KD-NST (Image)	88.9	85.0	86.3	77.2
+KD-Voken (Video)	89.7	**85.5**	86.5	77.8
+KD-NST (Video)	**91.1**	85.0	**87.4**	**78.4**

5.3 Analyzing the Knowledge Learned from Video

In this subsection, we analyze the knowledge that our language models learn from video via cross-modal knowledge distillation. To measure linguistic knowledge and physical reasoning ability, we evaluate our models by finetuning on the GLUE diagnostics dataset [68] and the Physical Interaction: Question Answering (PIQA) dataset [6]. We also present visualizations in the appendix.

Linguistic Knowledge. GLUE diagnostics dataset [68] evaluates sentence understanding through natural language inference (NLI) problems. The dataset consists of sentence pairs labeled with their entailment relations (entailment, contradiction, or neutral) in both directions and tagged with a set of entailment labels. Each example in the dataset is labeled with 4 categories of linguistic phenomena: (1) lexical semantics, (2) predicate-argument structure, (3) logic, and (4) knowledge that includes common sense. In Table 5, we compare the baseline language model ($\text{BERT}_{6L/512H}$ pre-trained on Wiki103) to our NST-distilled model. We finetune the models on MNLI [69] that has the same format and test on GLUE diagnostics. We observe a large gain on the knowledge category (which involves common sense and external world knowledge) while there is no significant performance improvement on the lexical semantics, predicate-argument structure, and logic categories. This suggests that our model learns more of the external, grounded world knowledge from the teacher model and the video-text dataset.

Table 5: Finetuning performance on GLUE diagonastics dataset [68]. Each example is labeled with one of four linguistic categories.

	Lexicon	Predicate	Logic	Knowledge
$\text{BERT}_{6L/512H}$	53.0	64.2	44.5	44.0
+ KD-NST	53.3 (+0.3)	63.7 (-0.5)	44.8 (+0.3)	48.6 **(+4.6)**

Physical Reasoning. PIQA [6] is a question answering dataset evaluating physical interactions and commonsense reasoning. In Table 6, our $\text{BERT}_{6L/512H}$ distilled with NST significantly outperform the text-only pretrained baseline. The finding suggests (consistent with the GLUE diagnostics findings above) that video knowledge distillation also helps improve physical reasoning capabilities of the language model.

Table 6: Finetuning performance on PIQA [6], which measures physical commonsense reasoning.

	PIQA Accuracy
$\text{BERT}_{6L/512H}$	56.9
+ KD-NST	60.0 (+3.1)

6 Conclusion

We introduce VIDLANKD, a novel cross-modal knowledge distillation method to help ground language understanding. Our teacher model is first trained on a video-text dataset, and then transfers its knowledge to a student language model on a text dataset. Via new distillation objectives and video-text datasets, our method overcomes the limitations of the recent vokenization method. We empirically demonstrate that our VIDLANKD improves on several NLU tasks over a text-only language model and vokenization. Also, we conduct comprehensive analysis to show the effectiveness of each proposed component and the world and physical knowledge learned from videos.

References

[1] Rohan Anil, Gabriel Pereyra, Alexandre Passos, Robert Ormandi, George E Dahl, and Geoffrey E Hinton. 2018. Large scale distributed neural network training through online distillation. *arXiv preprint arXiv:1804.03235*.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *ICCV*.

[3] Lei Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? In *NeurIPS*.

[4] Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

[5] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience grounds language. In *EMNLP*.

[6] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *AAAI*, pages 7432–7439.

[7] Patrick Bordes, Eloi Zablocki, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari. 2019. Incorporating visual semantics into sentence representations within a grounded space. In *EMNLP-IJCNLP*.

[8] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970.

[9] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. In *SemEval*.

[10] Yevgen Chebotar and Austin Waters. 2016. Distilling knowledge from ensembles of neural networks for speech recognition. In *Interspeech*, pages 3439–3443.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR.

[12] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

[13] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Learning universal image-text representations. In *ECCV*.

[14] Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. 2018. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. In *AAAI*.

[15] Gordon Christie, Ankit Laddha, Aishwarya Agrawal, Stanislaw Antol, Yash Goyal, Kevin Kochersberger, and Dhruv Batra. 2016. Resolving language and vision ambiguities together: Joint segmentation & prepositional attachment resolution in captioned scenes. In *EMNLP*.

[16] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

[17] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

[20] Tuong Do, Thanh-Toan Do, Huy Tran, Erman Tjiputra, and Quang D Tran. 2019. Compact trilinear interaction for visual question answering. In *ICCV*, pages 392–401.

[21] 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. In *NeurIPS*.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

[23] David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

[24] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *JMLR*, 13(25):723–773.

[25] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. 2016. Cross modal distillation for supervision transfer. In *CVPR*, pages 2827–2836.

[26] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*.

[27] Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*, pages 770–778.

[29] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. 2019. Bag of tricks for image classification with convolutional neural networks. In *CVPR*, pages 558–567.

[30] Jack Hessel, Lillian Lee, and David Mimno. 2019. Unsupervised discovery of multimodal links in multi-image, multi-sentence documents. In *EMNLP-IJCNLP*, pages 2034–2045.

[31] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2014. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning Workshop*.

[32] Zehao Huang and Naiyan Wang. 2017. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*.

[33] Shankar Iyer, Nikhil Dandekar, and Kornel Csernai. 2017. First quora dataset release: Question pairs.

[34] Hirokatsu Kataoka, Tenga Wakamiya, Kensho Hara, and Yutaka Satoh. 2020. Would mega-scale datasets further enhance spatiotemporal 3d cnns? *arXiv preprint arXiv:2004.04968*.

[35] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

[36] Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. 2018. Learning visually grounded sentence representations. In *NAACL*.

[37] Douwe Kiela, Ivan Vulic, and Stephen Clark. 2015. Visual bilingual lexicon induction with transferred convnet features. In *EMNLP*. ACL; East Stroudsburg, PA.

[38] Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *ACL*.

[39] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *ICLR*.

[40] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. 2014. What are you talking about? text-to-image coreference. In *CVPR*, pages 3558–3565.

[41] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. Hmdb: a large video database for human motion recognition. In *ICCV*, pages 2556–2563. IEEE.

[42] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *ICLR*.

[43] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344.

[44] 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. In *EMNLP*.

[45] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer.

[46] 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*.

[47] Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

[48] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.

[49] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *ICLR*.

[50] 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*.

[51] 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*, pages 2630–2640.

[52] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *AAAI*, volume 34, pages 5191–5198.

[53] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*.

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

[55] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*.

[56] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *ACL*.

[57] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*.

[58] Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *EMNLP*.

[59] Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.

[60] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *ICML*.

[61] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CoRR*.

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

[63] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*.

[64] Hao Tan and Mohit Bansal. 2020. Vokenization: improving language understanding with contextualized, visual-grounded supervision. In *EMNLP*.

[65] Jiaxi Tang and Ke Wang. 2018. Ranking distillation: Learning compact ranking models with high performance for recommender system. In *SIGKDD*, pages 2289–2298.

[66] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive representation distillation. In *ICLR*.

[67] 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*.

[68] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*.

[69] Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL*.

[70] 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*.

[71] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*.

[72] Zheng Xu, Yen-Chang Hsu, and Jiawei Huang. 2018. Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks. In *ICLR Workshop*.

[73] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*.

[74] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP*.

[75] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. 2018. Deep mutual learning. In *CVPR*, pages 4320–4328.

[76] Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2019. Neural machine translation with universal visual representation. In *ICLR*.

[77] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *AAAI*.

[78] Luowei Zhou, Chenliang Xu, and Jason J Corso. 2017. Towards automatic learning of procedures from web instructional videos. In *AAAI*.

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See the supplemantry material.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See the supplemantry material.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See the supplemantry material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4.

(b) Did you mention the license of the assets? [Yes] See the supplemantry material.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See the supplemantry material.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]