Leveraging Visual Knowledge in Language Tasks: An Empirical Study on Intermediate Pre-training for Cross-modal Knowledge Transfer

Anonymous ACL submission

Abstract

Pre-trained language models are still far from human performance in tasks that need understanding of properties (e.g. appearance, measurable quantity) and affordances of everyday objects in the real world since the text lacks such information due to reporting bias. In this work, we study whether integrating visual knowledge into a language model can fill the gap. We investigate two types of knowledge transfer: (1) *text knowledge transfer* using image captions that may contain enriched visual knowledge and (2) *cross-modal knowledge transfer* using both images and captions with vision-language training objectives. On 5 downstream tasks that may need visual knowledge to solve the problem, we perform extensive empirical comparisons over the presented objectives. Our experiments show that visual knowledge transfer can improve performance in both low-resource and fully supervised settings.[1]

1 Introduction

Pre-trained language models (PTLMs) such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and T5 (Raffel et al., 2020) have shown impressive results in various conventional natural language understanding (NLU) tasks by capturing syntactic and semantic knowledge from the pre-training tasks of *masked language modeling* and *masked span infilling* tasks on massive text corpora.

Though yielding good performance on various NLU downstream tasks, these pre-training objectives suffer from a lack of out-of-domain knowledge that is not explicitly present in the pre-training corpus (Gururangan et al., 2020; Petroni et al., 2021; Schick and Schütze, 2020). Specifically, one type of knowledge that models often struggle with is the visual knowledge of common objects such as attributes (e.g. appearance, measurable quantity) and affordances. This is because this kind of knowledge is rarely explicitly described in the training text due to reporting bias. For example, as shown in Figure 1, people tend to report what interests them rather than general facts such as a shape or color of oranges they already know.

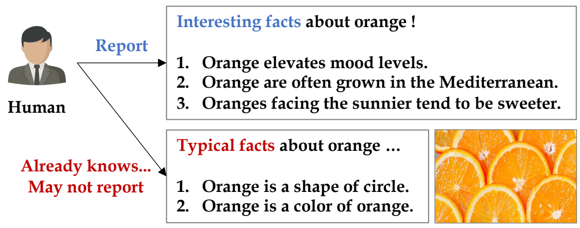

Figure 1: **Reporting Bias.** People tend to report what interests them rather than typical and general facts.

Towards better knowledge-enhanced PTLMs, recent works incorporate external knowledge bases (e.g., knowledge graph, dictionary) to inject entity knowledge into PTLMs (Zhang et al., 2019; Peters et al., 2019; Wang et al., 2021; Yu et al., 2021) or retrieve knowledge from external knowledge bases to solve the problem (Lin et al., 2019; Wang et al., 2020). However, these approaches still suffer from a lack of visual knowledge that is important to understand the real world.

In this paper, we conduct systematic experiments to understand whether such visual knowledge can be transferred into LMs, and if so, how to perform effective knowledge transfer. Specifically, we look into a series of analysis question as follows: (1) Can intermediate pre-training (Pruksachatkun et al., 2020) on image-caption pairs help transfer the knowledge? (2) What types of knowledge sources are more helpful? To answer questions, we explore various intermediate pre-training tasks (Pruksachatkun et al., 2020) on two different sources: text-only (*text knowledge transfer* from visual domains) and image-caption pairs (*cross-modal knowledge transfer*).

For the text knowledge transfer, we utilize text corpus from visual domain, e.g., image captions. We leverage two training objectives for the lan-

[1] Code and data have been uploaded and will be published.

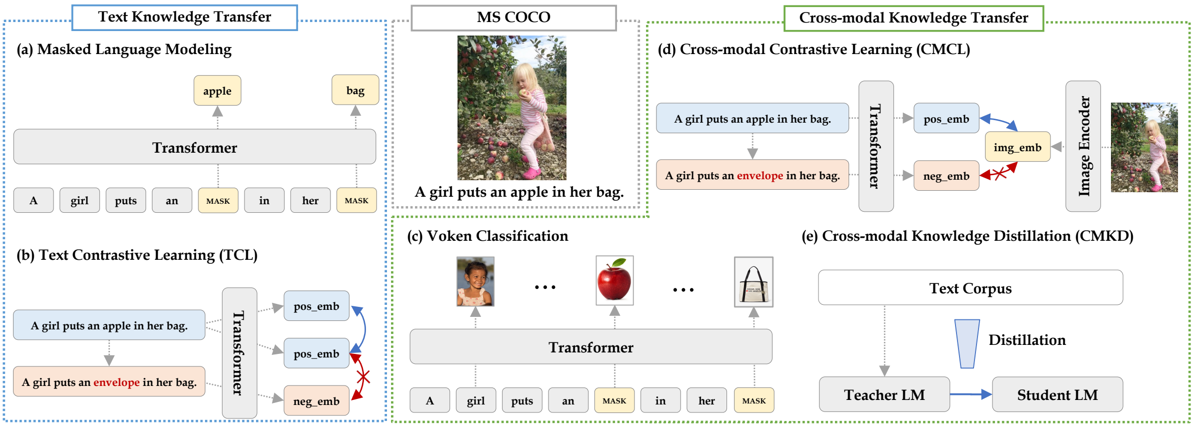

Figure 2: **Illustration of different methods for transferring visual knowledge into transformer-based language model.** In this example, we assume image-caption pair as an input. (a) *masked language model* (Devlin et al., 2018) on image captions. (b) *text contrastive learning* obtains positive example by dropout representation to learn better sentence representation while negative augmentation is optional. (c) *voken classification* employs token-level text-to-image retrieval to transfer visual knowledge. (d) *cross-modal contrastive learning* aims to train correct paring of images and captions. (e) *cross-modal knowledge distillation* transfers knowledge from the teacher model, which is trained by cross-modal contrastive learning, into student model.

guage model: (1) *masked language modeling* follows the domain adaptive pre-training scheme (Gururangan et al., 2020), assuming the corpus contains enriched visual knowledge or physical commonsense knowledge; (2) *text contrastive learning* augments the sentence representation with dropout to create positive samples while considering all others in the batch as negative samples for the contrastive learning (Gao et al., 2021), assuming training better sentence representations leads to better understanding of the corpus.

For the cross-modal knowledge transfer, we explore multiple methods to transfer visual-related knowledge to LMs: (1) *masked language modeling with visual clues* incorporates visual clues to capture dependencies between visual and linguistic contents (Su et al., 2019); (2) *voken classification* contextually aligns language tokens to their related images (called "vokens") to transfer visual knowledge into LMs (Tan and Bansal, 2020); (3) *cross-modal contrastive learning* aims to improve text representations by maximizing the agreement between correct image-text pairs versus random (in-batch) and adversarial negative pairs by contrastive learning between image and text modalities; and (4) *cross-modal knowledge distillation* transfers the knowledge from the teacher model, which is trained by cross-modal contrastive learning on image and text modalities, to the student language model using knowledge distillation.

We perform comprehensive comparisons on five downstream tasks that may require visual or physical commonsense knowledge, including PIQA (Bisk et al., 2020), Visual Paraphrasing (VP) (Lin and Parikh, 2015), CSQA (Talmor et al., 2018), OBQA (Mihaylov et al., 2018), and RiddleSense (Lin et al., 2021). Results suggest that: (1) Simple intermediate pre-training on captions can help improving performance on commonsense reasoning that needs physical or visual knowledge. (2) Cross-modal knowledge transfer approaches consistently improve the performance in a large margin when only few train examples are available. (3) Cross-modal contrastive learning shows that it is best for packaging visual knowledge into LMs.

2 Analysis Setup

In this work, we study how to transfer the visual knowledge into language models. For this study, we introduce our analysis setup: problem formulation, analysis questions, and knowledge corpora.

2.1 Problem Formulation

We focus on a pre-trained text encoder f_L and an image encoder f_V if images are available. f_L and f_V are initialized with pre-trained model and we continue to pre-train the models on different sources and tasks, which we call *intermediate pre-training*. After the intermediate pre-training, we fine-tune f_L on downstream NLU tasks. Existing NLU benchmarks have been trained against standard supervised learning paradigms that typically require a large number of question answering examples which need a large annotation efforts. However, in scenarios where the number of labeled examples is small, the model tends to overfit the training examples and shows poor generalization performance on test set. Here, we evaluate the intermediate pre-training objective's generalization ability on test set in both fully supervised and low-resource settings.

2.2 Analysis Questions

In this paper, we provide a comprehensive study for transferring the visual knowledge into LMs. Visual knowledge transfer can be done in two approaches, depending on the source to be trained: (1) *Text knowledge transfer* using the text corpus in the visual domain, e.g., image captions and (2) *cross-modal knowledge transfer* which passes visual knowledge about common objects to LMs by training over paired image and captions. By evaluating the model on 5 downstream datasets that require physical and visual commonsense knowledge, we explore following three research questions.

Q1: Can intermediate pre-training on external knowledge sources help transfer visual knowledge to augment text encoders? We investigate diverse intermediate pre-training methods with external knowledge sources including caption data to inject visual information from images and captions into LMs. We first analyze the performance of text and cross-modal knowledge transfer methods with a image-caption dataset, and we additionally study text knowledge transfer methods with other text corpora such as GenericsKB (Bhakthavatsalam et al., 2020), Wiki103 (Merity et al., 2016) and BookCorpus (Zhu et al., 2015a).

Q2: What types of knowledge sources are more helpful for visual knowledge transfer? As mentioned above, we have two categories to exploit visual information: (1) *text knowledge transfer* and (2) *cross-modal knowledge transfer*. Here, we explore which type of knowledge transfer is more useful to transfer the visual knowledge into LMs.

Q3: What intermediate pre-training objectives are effective for cross-modal knowledge transfer? We present three pre-training objectives for cross-modal knowledge transfer: (1) voken classification, (2) contrastive learning, and (3) knowledge distillation. Here, we want to present which strategy is best suited for cross-modal knowledge transfer. Furthermore, we study how to enhance cross-modal contrastive learning with adversarial negative samplings.

Dataset	# Train	# Dev	# Test	# choices
PIQA	14,113	1,838	2,000	2
VP	21,988	2,000	6,057	2
CSQA	8,500	1,221	1,241	5
OBQA	4,957	500	500	4
RiddleSense	3,510	1,021	1,202	5

Table 1: **Downstream task data statistics.** We create in-house test set for PIQA and CSQA, and in-house dev set for VP by splitting the train set.

2.3 Pre-training Data

To transfer the visual knowledge, we collect 250K image-caption pairs from MS COCO (Lin et al., 2014; Chen et al., 2015). MS COCO is a large scale dataset that contains images reflecting the composition of actual everyday scenes and corresponding captions which describe contextual reasoning between objects in the scene. We only use captions for text knowledge transfer while we use both images and captions for cross-modal knowledge transfer. As an ablation study, we explore other text corpora such as GenericsKB (Bhakthavatsalam et al., 2020), Wiki103 (Merity et al., 2016) and BookCorpus (Zhu et al., 2015a).

2.4 Downstream Tasks and Datasets

For downstream benchmarks, we find tasks that can benefit from visual knowledge: multiple choice question answering tasks including PIQA (Bisk et al., 2020) which requires physical commonsense reasoning, CSQA (Talmor et al., 2018) for general understanding of commonsense reasoning, OBQA (Mihaylov et al., 2018) that needs elemenatry-level science knowledge, and RiddleSense (RS) (Lin et al., 2021) for complex understanding of figurative language, and binary classification task including Visual Paraphrasing (VP) (Lin and Parikh, 2015) that needs scene understanding. We use in-house test sets made from training sets for PIQA and CSQA since test set is not provided to public. We list the data statics in Table 1. Moreover, We additionally test on GLUE (Wang et al., 2018) to evaluate the general text understanding.

2.5 Evaluation Protocol

We evaluate the models in both fully supervised and low-resource settings. For both settings, we consider accuracy for 5 different classification tasks and get average performance over tasks to check the final performance. In fully supervised setting, we evaluate models with 3 different random seeds and report the average accuracy. In a low-resource setting, we consider the size of train data to 64 or 128. For each experiment, we run over 5 different sub-samples and show the average accuracy.

3 Method

In this section, we introduce the following two approaches to integrate visual knowledge into LMs: (1) *text knowledge transfer*; and (2) *cross-modal knowledge transfer*. Throughout this section, we assume the data is a collection of image-caption pairs $\{(x_i^v, x_i^l)\}_{i=1}^m$ and image encoder f_V and text encoder f_L are given.

3.1 Text Knowledge Transfer

For text knowledge transfer, we investigate following pre-training objectives: (1) *masked language modeling*; and (2) *text contrastive learning*.

Masked Language Modeling (MLM) Following BERT (Devlin et al., 2018), we select 15% of input tokens and replace them with `[MASK]`. Of the selected tokens, 80% are replaced, 10% are not changed and 10% are replaced by random vocabulary token. Here, we employ dynamic masking, which performs random masking and replacement during training to prevent the same masking for the same examples (Liu et al., 2019). MLM objective is the cross-entropy loss for masked token predictions :

$$\ell_{\text{MLM}}(x_i^l) = -\log p(x_i^l | x^{\text{masked}}), \quad (1)$$

where x_i is the i-th token and x^{masked} is a mask.

Text Contrastive Learning (TCL) Contrastive learning aims to learn representations by pulling positive pairs closer and pushing negative pairs apart. Here, we employ the contrastive framework with cross-entropy objective and in-batch negatives (Chen et al., 2020a; Gao et al., 2021). Given a text encoder f_L, and a caption x_i^l, we first get text representations using the encoders $h_i^l = f_L(x_i^l)$. Following Gao et al. (2021), we augment identical positive sample h_i^{l+} by different dropout representations. The contrastive loss is defined as follows:

$$\ell_i^l = -\log \frac{e^{\text{sim}(h_i^l, h_i^{l+})/\tau}}{\sum_{j=1}^N e^{\text{sim}(h_i^l, h_i^{l+})/\tau}}, \quad (2)$$

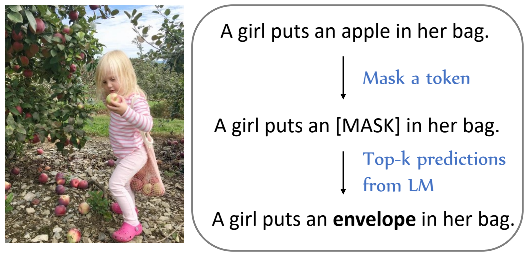

Figure 3: **LM perturbation.** We create adversarial negatives using language models.

where N is a batch size and sim$(\cdot)$ represents cosine similarity, i.e., sim$(u, v) = u \cdot v / \|u\| \|v\|$. τ represents a temperature parameter.

3.2 Cross-modal Knowledge Transfer

Language models might learn additional information from visual sources such as images and captions. So we include a variety of vision-based approaches and investigate the approaches whether they can benefit from visual sources. We introduce vision-based approaches as follows.

Voken Classification Vokenization (Tan and Bansal, 2020) employs token-level text-to-image retrieval to transfer visual knowledge. It aligns language tokens to their related images (called "vokens") to transfer visual knowledge into LMs, and call "voken classification". Given text x and a voken v_i for the i-th token, the loss is defined as

$$\ell_i^{\text{voken}} = -\log(p(v_i | x)). \quad (3)$$

Similar to masked language modeling, it classifies each token to a corresponding voken. Vokenization trains language models with the voken classification task and MLM.

Masked Language Modeling with Visual Clues VL-BERT (Su et al., 2019) adopts masked language modeling with visual clues in which models are given a caption with masked tokens and an image and predict the masked tokens using visual clues. VL-BERT is pre-trained on Conceptual Captions (Sharma et al., 2018) as an image-caption corpus, and BooksCorpus (Zhu et al., 2015b) and English Wikipedia as text-only corpora. It shows its effectiveness in many vision-language tasks. We investigate whether this model also succeed in NLP tasks and compare it with others.

Cross-modal Contrastive Learning (CMCL) To harness the visual knowledge from image-caption datasets, we adopt contrastive loss on image and text vectors. Given an image encoder f_V, a text encoder f_L, and an image-caption pair (x_i^v, x_i^l), we first get image and text representations using the encoders $h_i^v = f_V(x_i^v), h_i^l = f_L(x_i^l)$. Then the contrastive learning objective contains two loss functions: an image-to-text contrastive loss $\ell^{(v,l)}$ and and a text-to-image contrastive loss $\ell^{(l,v)}$. The image-to-text contrastive loss is defined as follows:

$$\ell_i^{(v,l)} = -\log \frac{e^{\text{sim}(h_i^v, h_i^l)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(h_i^v, h_j^l)/\tau}}, \tag{4}$$

where N is a batch size and $\text{sim}(\cdot)$ represents cosine similarity. This loss encourages a closer distance between representations of aligned image-caption pairs than unaligned pairs given an image and multiple captions. Similarly, the text-to-image contrastive loss $\ell^{(l,v)}$ is defined as follows:

$$\ell_i^{(l,v)} = -\log \frac{e^{\text{sim}(h_i^l, h_i^v)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(h_i^l, h_j^v)/\tau}}. \tag{5}$$

The final loss is defined as

$$L = \frac{1}{N} \sum_{i=1}^{N} (\ell_i^{(v,l)} + \ell_i^{(l,v)}). \tag{6}$$

CLIP (Radford et al., 2021) and ConVIRT (Zhang et al., 2020) also adopt contrastive learning, but we freeze the image encoder in training and use the trained text encoder for downstream tasks.

CMCL with Adversarial Negative Samples (ANS) As in-batch negatives in CMCL are not challenging enough for models to distinguish, we present adversarial negative sampling strategy to improve CMCL. Given an image-caption pair (x_i^v, x_i^l), we define a LM-perturbed sentence $x_i^{l^-}$, which is a hard negative where n is replaced with a different word n' from a probability distribution of PTLMs. We expect the l^- is syntactically correct and plausible sentence even the word n is replaced to n', while it does not semantically match to the corresponding image x_i^v. With such hard negative, we try to make more challenging task so that models can effectively learn from the task. For example, we choose a word 'girl' in the sentence 'A girl puts an apple in her bag.' in Figure 3. Then we mask the word with `[MASK]` token to do masked token predictions by PTLMs. Then we get top-k predictions from language models and replace the masked tokens with one of the predicted ones. To avoid false negative sentences which may have the same semantics as the original sentence, we introduce an additional filtering step: if the masked predictions are synonyms or hypernyms of the original tokens, we discard the predictions. We use WordNet (Miller, 1995) to find synonyms and hypernyms. The contrastive loss with hard negative is defined as follows:

$$-\log \frac{e^{\text{sim}(h_i^v, h_i^l)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(h_i^v, h_j^l)/\tau} + \sum_{k=1}^{M} e^{\text{sim}(h_i^v, h_j^{l^-})/\tau}}, \tag{7}$$

where M is the number of hard negative samples per positive pair. This formula is only for image-to-text contrastive loss $\ell^{(v,l)}$ and final loss is defined to same as equation (6).

CMCL with Positive Sample Augmentation (PSA) In ANS, we filter perturbed sentences where the masked predictions are synonyms or hypernyms of the original tokens. Instead of excluding these perturbed sentences, another option is to include them as additional positive samples l^+ to the paired images. We name this as positive sample augmentation (PSA). It also adopts LM-perturbed negative samples as in ANS.

Cross-modal Knowledge Distillation (CMKD) Cross-model knowledge distillation is to transfer knowledge between different modalities, e.g., image modality and text modality. In this category, CMKD is to transfer knowledge from a teacher model which is knowledgeable about visual information. VidLanKD (Tang et al., 2021) also utilizes a cross-modal knowledge distillation method to help with general language understanding. A teacher model is first trained using contrastive learning on a video-text dataset, and then it transfers its knowledge to a student language model using KD on a text corpus. Their contrastive learning loss (hinge loss) is defined as

$$\begin{aligned} L = \sum_{i}^{N} [&\max(0, \alpha - \text{sim}(h_i^v, h_i^l) + \text{sim}(h_i^{v'}, h_i^l)) \\ &+ \max(0, \alpha - \text{sim}(h_i^v, h_i^l) + \text{sim}(h_i^v, h_i^{l'}))], \end{aligned} \tag{8}$$

where v' and l' are a random image and caption text, respectively. α is the margin between the similarities of a positive pair and a negative pair. Instead of video datasets, we use a MS COCO dataset to train a teacher model and use two versions of contrastive learning, equations (6) and (8).

As another version of CMKD, we consider distilling visual knowledge from a pre-trained vision-

	Model	PIQA		VP		CSQA		OBQA		RiddleSense		Average	
		64	128	64	128	64	128	64	128	64	128	64	128
-	BERT-base	$52.6_{\pm0.9}$	$53.8_{\pm0.1}$	$85.9_{\pm1.1}$	$86.6_{\pm0.7}$	$35.8_{\pm0.7}$	$37.8_{\pm0.3}$	$31.3_{\pm1.2}$	$32.0_{\pm0.7}$	$24.7_{\pm0.1}$	$25.2_{\pm0.2}$	46.1	47.1
Caption	MLM	$53.1_{\pm0.2}$	$54.3_{\pm0.3}$	$86.5_{\pm0.3}$	$87.3_{\pm0.4}$	$35.7_{\pm0.3}$	$36.7_{\pm0.1}$	$33.4_{\pm0.6}$	$34.2_{\pm0.3}$	$26.3_{\pm0.1}$	$26.5_{\pm0.2}$	47.0	47.8
	TCL	$52.6_{\pm0.5}$	$52.9_{\pm0.6}$	$86.4_{\pm0.1}$	$88.0_{\pm0.1}$	$35.7_{\pm0.2}$	$36.1_{\pm0.3}$	$\mathbf{34.2}_{\pm1.4}$	$\mathbf{35.2}_{\pm0.7}$	$\mathbf{30.3}_{\pm0.5}$	$\mathbf{30.7}_{\pm0.4}$	47.8	48.5
	TCL + MLM	$53.6_{\pm0.7}$	$54.6_{\pm0.2}$	$84.2_{\pm0.2}$	$87.6_{\pm0.3}$	$33.6_{\pm2.2}$	$35.1_{\pm0.6}$	$31.8_{\pm2.3}$	$34.3_{\pm0.5}$	$20.6_{\pm0.0}$	$20.6_{\pm0.0}$	44.7	46.4
	TCL + ANS	$50.0_{\pm0.7}$	$50.5_{\pm0.6}$	$67.3_{\pm0.4}$	$68.2_{\pm0.7}$	$26.8_{\pm1.2}$	$27.5_{\pm0.5}$	$33.4_{\pm1.1}$	$35.0_{\pm1.0}$	$26.1_{\pm1.7}$	$26.5_{\pm1.8}$	40.7	41.5
	TCL + PSA + ANS	$51.1_{\pm0.1}$	$51.2_{\pm0.4}$	$66.0_{\pm0.0}$	$66.0_{\pm0.0}$	$22.7_{\pm0.9}$	$22.9_{\pm0.1}$	$30.2_{\pm3.1}$	$31.8_{\pm0.4}$	$23.5_{\pm1.2}$	$25.2_{\pm1.5}$	38.7	39.4
Caption-Image Pairs	VL-BERT-base	$53.1_{\pm0.6}$	$53.9_{\pm0.4}$	$\underline{88.5}_{\pm0.3}$	$88.4_{\pm0.5}$	$36.2_{\pm0.7}$	$36.8_{\pm0.8}$	$33.4_{\pm1.2}$	$34.6_{\pm1.2}$	$26.1_{\pm0.8}$	$26.1_{\pm0.9}$	47.7	48.5
	Vokenization	$50.5_{\pm0.5}$	$51.1_{\pm0.4}$	$68.8_{\pm1.6}$	$78.1_{\pm1.9}$	$19.2_{\pm1.4}$	$21.5_{\pm0.8}$	$31.2_{\pm2.7}$	$33.2_{\pm2.2}$	$17.1_{\pm0.5}$	$16.7_{\pm0.7}$	37.3	40.1
	VidLanKD	$55.0_{\pm0.4}$	$55.6_{\pm0.5}$	$86.7_{\pm0.5}$	$\underline{88.5}_{\pm0.5}$	$37.1_{\pm1.0}$	$38.6_{\pm0.5}$	$31.8_{\pm1.3}$	$32.6_{\pm1.0}$	$24.4_{\pm0}$	$24.4_{\pm0}$	47.0	47.9
	VidLanKD variant	$\underline{55.3}_{\pm0.3}$	$\underline{55.2}_{\pm0.4}$	$87.4_{\pm0.1}$	$88.2_{\pm0.6}$	$\underline{37.3}_{\pm1.2}$	$\underline{38.9}_{\pm0.5}$	$32.4_{\pm2.1}$	$32.2_{\pm1.1}$	$24.4_{\pm0.0}$	$24.4_{\pm0.0}$	47.3	47.7
	CMKD (VL-BERT-large)	$54.7_{\pm0.5}$	$54.5_{\pm0.2}$	$86.5_{\pm0.8}$	$88.4_{\pm0.4}$	$36.7_{\pm0.4}$	$38.5_{\pm0.4}$	$29.8_{\pm0.8}$	$31.7_{\pm0.2}$	$25.2_{\pm0.1}$	$25.2_{\pm0.0}$	46.5	47.6
	CMCL	$54.7_{\pm0.4}$	$55.1_{\pm0.1}$	$87.9_{\pm0.3}$	$\mathbf{88.9}_{\pm0.2}$	$36.3_{\pm0.3}$	$38.4_{\pm0.4}$	$31.1_{\pm1.1}$	$32.8_{\pm0.9}$	$25.0_{\pm0.2}$	$25.4_{\pm0.4}$	47.0	48.1
	CMCL + ANS	$\mathbf{55.4}_{\pm0.1}$	$\mathbf{55.7}_{\pm0.2}$	$88.1_{\pm0.9}$	$\mathbf{88.9}_{\pm0.7}$	$\mathbf{37.5}_{\pm0.8}$	$\mathbf{39.0}_{\pm0.2}$	$32.2_{\pm0.7}$	$32.0_{\pm0.6}$	$\underline{27.4}_{\pm0.0}$	$27.5_{\pm0.1}$	$\underline{48.1}$	$\underline{48.6}$
	CMCL + PSA + ANS	$\mathbf{55.4}_{\pm0.2}$	$55.1_{\pm0.2}$	$\mathbf{88.8}_{\pm1.0}$	$88.2_{\pm0.2}$	$37.0_{\pm0.3}$	$38.1_{\pm0.3}$	$\underline{34.1}_{\pm0.4}$	$\underline{34.8}_{\pm0.9}$	$26.7_{\pm0.4}$	$\underline{28.8}_{\pm0.7}$	**48.4**	**49.0**

Table 2: **Performance (accuracy) in low-resource setting.** We test models on diverse datasets with low-resource learning (64 and 128 training samples). We use captions in the MS COCO dataset for text knowledge transfer methods and images and captions for cross-modal knowledge transfer methods. We get average performance on 64 and 128 training samples. **Bold** and underlined numbers refer to the best and second-best performance, respectively.

language model, VL-BERT, which is knowledgeable about grounded language. We adopt masked language modeling on Wikitext103 (Merity et al., 2016), a subset of English Wikipedia, in the knowledge distillation step. For knowledge distillation, we adopt Neuron Selectivity Transfer (NST) (Huang and Wang, 2017), which proves the effectiveness in VidLanKD (Tang et al., 2021).

4 Experimental Settings

For all the approaches, we use `bert-base-uncased` (Devlin et al., 2018) as text encoder f_L and ResNeXt101 (Xie et al., 2017) as an image encoder f_V. For text knowledge transfer, (1) MLM follows the exact setting of codebase in huggingface[2] which uses dynamic masking strategy to conduct language modling task. (2) TCL conducts contrastive learning with f_L. We choose the best checkpoint by the best spearman correlation on STSb (Cer et al., 2017). For cross-modal knowledge transfer, (1) CMKD explores VL-BERT, Vokenization, and VidLanKD approaches. Here, we use VL-BERT-large model to do CMKD. Vokenization uses a checkpoint from their official codebase[3] and VidLanKD trains a teacher model by two versions of contrastive learning (equations (6) and (8)) on MS COCO dataset. We set $\alpha = 1$ in VidLanKD (equation (8)). (2) CMCL conducts contrastive learning with f_L and f_V. Here, we set $\tau = 0.05$ (equations (4) and (5)). (3) CMCL with ANS chooses three noun words or verb words to do masked prediction and

[2] `https://github.com/huggingface/transformers/tree/master/examples/pytorch/language-modeling`

[3] `https://github.com/airsplay/vokenization`

	Model	PIQA	VP	CSQA	OBQA	RiddleSense	Average
-	BERT-base	$62.5_{\pm1.3}$	$93.1_{\pm0.4}$	$53.2_{\pm1.2}$	$52.2_{\pm0.5}$	$38.9_{\pm0.9}$	59.9
Caption	MLM	$63.8_{\pm0.9}$	$93.5_{\pm0.1}$	$52.6_{\pm0.3}$	$\underline{53.9}_{\pm1.1}$	$39.3_{\pm1.4}$	**60.6**
	TCL	$62.1_{\pm0.5}$	$93.5_{\pm0.4}$	$49.0_{\pm0.5}$	$\mathbf{54.1}_{\pm1.0}$	$\mathbf{41.2}_{\pm0.3}$	$\underline{60.1}$
	TCL + MLM	$62.3_{\pm0.7}$	$93.2_{\pm0.3}$	$49.0_{\pm0.4}$	$49.0_{\pm0.8}$	$\underline{40.5}_{\pm0.5}$	58.8
	TCL + ANS	$60.1_{\pm1.2}$	$93.3_{\pm0.1}$	$47.0_{\pm0.1}$	$50.2_{\pm0.9}$	$36.7_{\pm0.8}$	57.4
	TCL + PSA + ANS	$59.5_{\pm1.0}$	$92.4_{\pm0.3}$	$34.0_{\pm1.3}$	$44.6_{\pm1.4}$	$28.4_{\pm2.3}$	51.7
Caption-Image Pairs	VL-BERT-base	$63.8_{\pm1.5}$	$93.6_{\pm0.1}$	$50.3_{\pm1.1}$	$49.6_{\pm2.3}$	$39.1_{\pm1.0}$	59.2
	Vokenization	$58.4_{\pm5.1}$	$92.7_{\pm0.3}$	$45.0_{\pm0.2}$	$48.1_{\pm0.8}$	$33.5_{\pm0.7}$	55.5
	VidLanKD	$63.1_{\pm1.1}$	$93.7_{\pm0.4}$	$52.4_{\pm0.8}$	$50.6_{\pm3.9}$	$39.5_{\pm1.7}$	59.8
	VidLanKD variant	$\mathbf{64.1}_{\pm0.2}$	$\underline{93.8}_{\pm0.3}$	$\mathbf{53.6}_{\pm0.5}$	$47.9_{\pm4.3}$	$38.8_{\pm2.0}$	59.6
	CMKD (VL-BERT-large)	$63.8_{\pm0.0}$	$93.7_{\pm0.7}$	$\underline{53.3}_{\pm1.4}$	$48.7_{\pm3.0}$	$38.7_{\pm0.4}$	59.6
	CMCL	$62.7_{\pm0.1}$	$93.3_{\pm0.3}$	$50.8_{\pm0.9}$	$52.3_{\pm0.7}$	$37.6_{\pm1.0}$	59.2
	CMCL + ANS	$63.5_{\pm0.1}$	$93.3_{\pm0.3}$	$50.3_{\pm0.1}$	$52.9_{\pm0.3}$	$38.4_{\pm0.9}$	59.7
	CMCL + PSA + ANS	$\underline{63.9}_{\pm0.5}$	$\mathbf{94.3}_{\pm0.1}$	$50.9_{\pm0.3}$	$52.4_{\pm1.2}$	$39.0_{\pm0.3}$	$\underline{60.1}$

Table 3: **Performance (accuracy) in fully supervised setting. Bold** and underlined numbers refer to the best and second-best performance, respectively.

use top-5 predictions from f_L as replacement. We filter out synonyms and hypernyms of original words using WordNet (Miller, 1995). (4) CMCL with PSA includes the perturbed sentences with synonyms and hypernyms as additional positive samples. In CMCL, we adopt ResNeXt101 (Xie et al., 2017) as an image encoder f_V and BERT as a text encoder f_L. TCL and CMCL train with batch size 64, maximum sequence length 20, learning rate 1e-4 for 3 epochs. For fine-tuning on downstream tasks, we do grid search on learning rates {5e-5, 1e-4, 3e-4, 4e-4, 5e-4, 6e-4} and choose the best learning rate. We set maximum epochs to 30 in low-resource and 15 in fully supervised settings.

5 Results and Analysis

We analyze the main results of intermediate pre-training. Tables 2 and 3 show the main results of low-resource learning and fully supervised learning with the MS COCO captioning dataset, respectively. We train the models with a few training examples,

	Model	RTE	MRPC	STS-B	CoLA	SST-2	QNLI	QQP	Avg.
-	BERT-base	**70.0**	87.9	89.1	57.4	91.3	90.4	89.3	**82.3**
Caption	MLM	62.8	87.0	89.1	53.9	**92.6**	91.1	**90.9**	81.0
	TCL	58.4	83.1	88.2	55.5	91.9	**91.4**	**90.9**	79.9
	TCL + MLM	54.8	81.6	87.2	53.6	91.9	90.9	89.2	78.5
	TCL + ANS	56.3	83.9	87.0	51.5	91.3	91.2	89.4	78.6
	TCL + PSA + ANS	52.3	75.6	81.5	17.4	90.0	85.8	88.2	70.1
Caption-Image Pairs	VL-BERT-base	57.4	85.7	89.5	**58.1**	90.6	89.7	88.7	80.0
	Vokenization	53.0	87.0	83.3	51.3	91.4	89.2	88.5	77.7
	VidLanKD	67.5	87.8	89.4	57.7	90.7	90.3	88.6	81.7
	VidLanKD variant	68.5	87.9	**89.7**	54.9	91.1	90.5	88.6	81.6
	CMKD (VL-BERT-large)	68.5	**88.5**	89.3	55.4	90.9	89.7	88.6	81.6
	CMCL	63.5	82.5	89.5	51.1	90.4	90.0	88.4	79.3
	CMCL + ANS	69.6	86.8	89.4	56.1	90.7	90.5	88.6	81.7
	CMCL + PSA + ANS	69.8	86.2	89.0	55.3	90.4	90.5	88.6	81.6

Table 4: **Performance (accuracy) on GLUE benchmark. Bold** and underlined numbers refer to the best and second-best performance, respectively.

64 and 128, to understand the better initialization. We argue that if a model obtains better performance in the low-resource setup, then it is a faster learner and has better generalization on downstream tasks.

Can text intermediate pre-training help improve text encoders? Text intermediate pre-training using MLM and TCL on a caption corpus improves the performance on downstream tasks in both low-resource and fully supervised settings. In particular, TCL shows significant improvement on OBQA and RiddleSense over BERT. These results suggest that text intermediate pre-training on visual-related datasets helps performance on commonsense reasoning tasks.

Can cross-modal intermediate pre-training help transfer visual knowledge to augment text encoders? We observe that cross-modal intermediate pre-training is helpful in both fully supervised and low-resource settings (See Table 2 and 3). Specifically, CMKD with VidLanKD variant outperforms the baseline by 1.6% point on the PIQA dataset in fully supervised setting. CMCL also shows its effectiveness. However, we could find that it becomes more powerful when equipped with PSA and ANS. It suggests that data augmentation for positive and negative sampling is an important factor for CMCL. In low-resource setting, we find that cross-modal knowledge transfer helps better initialization and let models learn new tasks faster.

What intermediate pre-training objectives are effective for cross-modal knowledge transfer? Among various cross-modal knowledge transfer methods, we study which method is the most effective for cross-modal knowledge transfer. Overall, CMCL with PSA and ANS shows the best performance among all cross-modal methods. Interestingly, VL-BERT also shows better performance than BERT-base on all datasets in the low-resource setting. This suggesting that exploiting images in masked language modeling task help transfer the knowledge to language models.

What types of knowledge sources are most helpful? Here, we investigate whether using an image source in addition to a text source can further improve the model. To answer this question, we analyze methods from different types of sources: text-only and text-image pair sources. We focus on the methods that use the contrastive learning objective: TCL and CMCL. Note that these two methods share the same objective but CMCL trains on cross modalities which are images and captions while TCL only trains on captions Overall, TCL performs slightly better than CMCL in low-resource and fully supervised settings. Interestingly, additional negative samples (ANS) and positive samples in TCL decreases the performance while they help CMCL to improve the performance. We conjecture that perturbed sentences in ANS might not be semantically negative to the original sentence so models learn from wrong labels.

5.1 Ablation Study

How do models perform on general NLU tasks? Table 4 presents results on GLUE benchmark. In GLUE, text intermediate pre-training methods slightly underperform the original BERT-base. We conjecture that the intermediate pre-training on caption data might sacrifice knowledge of general language understanding.

Analysis on diverse text corpora Table 5 represents text approaches with different pre-training corpora: MS COCO captions (Lin et al., 2014; Chen et al., 2015), GenericsKB (Bhakthavatsalam et al., 2020), BooksCorpus (Zhu et al., 2015a), and WikiText103 (Merity et al., 2016). We notice that caption datasets are useful on OBQA and RiddleSense datasets while GenericsKB are the most helpful on PIQA datasets. Results are expected since GenericsKB contains a lot of everyday statements that contain various types of commonsense.

Different training sizes. We test different training sizes on PIQA in Fig. 4. In the experiment, we observe that CMCL consistently outperforms BERT on all training sizes. Additional negative sample (ANS) improves the CMCL on different training sizes, and positive sample augmentation boosts the performance of CMCL further. This sug-

	Model	PIQA			VP			CSQA			OBQA			RiddleSense		
		64	128	Full	64	128	Full	64	128	Full	64	128	Full	64	128	Full
-	BERT-base	$52.6_{\pm0.9}$	$53.8_{\pm0.1}$	$62.5_{\pm1.3}$	$85.9_{\pm1.1}$	$86.6_{\pm0.7}$	$93.1_{\pm0.4}$	$35.8_{\pm0.7}$	$37.8_{\pm0.3}$	**53.2**$_{\pm1.2}$	$31.3_{\pm1.2}$	$32.0_{\pm0.7}$	$52.2_{\pm0.5}$	$24.7_{\pm0.1}$	$25.2_{\pm0.2}$	$38.9_{\pm0.9}$
CP.	MLM	$53.1_{\pm0.2}$	$54.3_{\pm0.3}$	$63.8_{\pm0.9}$	$86.5_{\pm0.3}$	$87.3_{\pm0.4}$	**93.5**$_{\pm0.1}$	$35.7_{\pm0.3}$	$37.7_{\pm0.1}$	$52.6_{\pm0.3}$	$33.4_{\pm0.6}$	$34.2_{\pm0.3}$	$53.9_{\pm1.1}$	$26.3_{\pm0.1}$	$26.5_{\pm0.2}$	$39.3_{\pm1.4}$
	TCL	$52.6_{\pm0.5}$	$52.9_{\pm0.6}$	$62.1_{\pm0.5}$	$86.4_{\pm0.1}$	$88.0_{\pm0.1}$	**93.5**$_{\pm0.4}$	$35.7_{\pm0.2}$	$36.1_{\pm0.3}$	$49.0_{\pm0.5}$	**34.2**$_{\pm1.4}$	**35.2**$_{\pm0.7}$	**54.1**$_{\pm1.0}$	**30.3**$_{\pm0.5}$	**30.7**$_{\pm0.4}$	**41.2**$_{\pm0.3}$
GK.	MLM	$53.2_{\pm0.1}$	$53.6_{\pm0.4}$	**64.9**$_{\pm0.1}$	$86.2_{\pm0.9}$	$87.6_{\pm0.3}$	$93.0_{\pm0.3}$	$34.6_{\pm0.7}$	$35.3_{\pm1.3}$	$51.6_{\pm0.5}$	$31.7_{\pm0.9}$	$32.3_{\pm1.0}$	$53.1_{\pm0.9}$	$25.8_{\pm0.6}$	$26.3_{\pm0.1}$	$39.3_{\pm0.7}$
	TCL	**56.0**$_{\pm1.0}$	**56.4**$_{\pm0.2}$	$64.4_{\pm0.1}$	**88.9**$_{\pm0.7}$	$89.4_{\pm0.2}$	$93.3_{\pm0.5}$	$37.8_{\pm1.2}$	**38.7**$_{\pm0.5}$	$51.0_{\pm0.5}$	$31.7_{\pm0.9}$	$32.3_{\pm1.0}$	$52.6_{\pm0.8}$	$27.4_{\pm0.2}$	$28.1_{\pm0.7}$	$40.9_{\pm0.8}$
BC.	MLM	$54.1_{\pm0.3}$	$54.1_{\pm0.8}$	$63.3_{\pm0.6}$	$86.4_{\pm0.8}$	$87.5_{\pm0.5}$	$93.0_{\pm0.3}$	$29.8_{\pm0.8}$	$32.1_{\pm0.9}$	$50.8_{\pm0.3}$	$29.6_{\pm0.8}$	$31.4_{\pm0.7}$	$50.2_{\pm0.4}$	$22.6_{\pm0.0}$	$22.7_{\pm0.0}$	$36.7_{\pm1.3}$
	TCL	$52.4_{\pm0.1}$	$53.1_{\pm0.4}$	$63.1_{\pm0.3}$	$87.1_{\pm1.9}$	**89.7**$_{\pm0.1}$	$93.2_{\pm0.2}$	**38.0**$_{\pm0.5}$	$38.1_{\pm1.1}$	$51.5_{\pm0.1}$	$33.8_{\pm2.7}$	$34.0\ _{\pm2.1}$	$55.6_{\pm0.4}$	$28.9_{\pm0.4}$	$29.1_{\pm0.3}$	**41.2**$_{\pm2.3}$
WT.	MLM	$52.7_{\pm0.2}$	$53.0_{\pm0.3}$	$63.8_{\pm0.6}$	$85.3_{\pm2.8}$	$88.1_{\pm0.3}$	**93.5**$_{\pm0.1}$	$33.2_{\pm1.4}$	$34.6_{\pm0.5}$	$52.5_{\pm0.2}$	$32.4_{\pm2.3}$	$33.0_{\pm0.7}$	$52.3_{\pm0.3}$	$24.4_{\pm0.0}$	$24.4_{\pm0.0}$	$39.4_{\pm2.0}$
	TCL	$52.9_{\pm0.9}$	$53.4_{\pm0.4}$	$62.7_{\pm0.6}$	$67.3_{\pm0.6}$	$68.6_{\pm0.7}$	$93.3_{\pm0.3}$	$31.3_{\pm1.6}$	$32.4_{\pm0.7}$	$48.2_{\pm0.3}$	$31.5_{\pm3.5}$	$33.1_{\pm0.6}$	$53.0_{\pm0.0}$	$24.8_{\pm1.3}$	$24.8_{\pm0.6}$	$36.3_{\pm1.0}$

Table 5: **Results of text knowledge transfer methods with different corpora.** We pre-train text knowledge transfer methods, MLM ans TCL, with different corpora. CP is MS COCO captions, GK is GenericsKB, BC is BooksCorpus, and WT is WikiText. **Bold** and underlined numbers refer to the best and second-best performance, respectively.

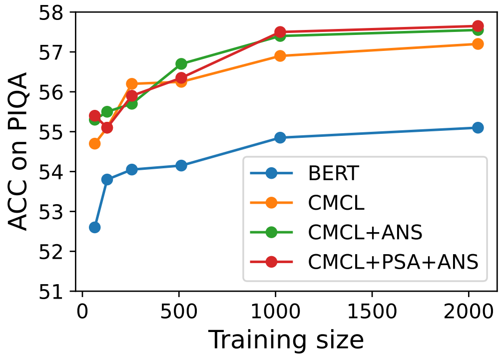

Figure 4: **Results on varying training sizes.** We test methods with different training sizes.

gests including perturbed sentences as positive and negative samples are useful to cross-modal knowledge transfer.

6 Related Work

Text Knowledge enhanced methods. Recently, huge efforts on integrating knowledge into PTLMs have been made. One typical form of knowledge is a knowledge graph. There have been efforts of using knowledge graph to inject entity and relation representations, which are pre-computed from external source, into PTLMs (Zhang et al., 2019; Peters et al., 2019; He et al., 2020). Some other works try to retrieve or generate the sub-graph from the graph to solve the problem (Lin et al., 2019; Wang et al., 2020). Another existing form of knowledge is extra large-scale corpus. Works that use such corpus present knowledge-related pre-training objectives such as concept order recovering (Zhou et al., 2021), entity category prediction (Yu et al., 2020) and source of knowledge prediction (Wang et al., 2021). They are mostly focused on injecting world knowledge presented in text, rather than physical and visual commonsense knowledge that can be found in images.

Cross-modal knowledge enhanced methods. There is a extensive line of works for a variety of vision-language tasks, such as VL-BERT (Su et al., 2019), VisualBert (Li et al., 2019), and Uniter (Chen et al., 2020b). These models aim to improve vision-language tasks, e.g., VQA (Goyal et al., 2017), and they are found to be not effective in improving language tasks (Tan and Bansal, 2020). Another line of works is to transfer visual knowledge to language models: Vokenization (Tan and Bansal, 2020) and VidLanKD (Tang et al., 2021). Vokenization employs token-level text-to-image retrieval to transfer visual knowledge to language models. For this, Vokenization introduces 30k vokens and matches each token into the limited voken space; it may have approximation errors. VidLanKD adopts contrastive learning to train a teacher model on video datasets and uses distillation approaches to distill visual knowledge from the teacher to a student model.

7 Conclusion

We study whether intermediate pre-training on visual knowledge can help transfer visual knowledge into LMs. We investigate text knowledge transfer and cross-modal knowledge transfer using images and captions. In our empirical analysis, we observe that intermediate pre-training on captions can help improving performance and cross-modal knowledge transfer approaches consistently improve performance. When the transfer methods are equipped with additional positive and negative samples, they show better performance. Future works include improving both commonsense reasoning and general language understanding.

References

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter E. Clark. 2020. Genericskb: A knowledge base of generic statements. *ArXiv*, abs/2005.00660.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7432–7439.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

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*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

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*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Bin He, Di Zhou, Jinghui Xiao, Xin Jiang, Qun Liu, Nicholas Jing Yuan, and Tong Xu. 2020. BERT-MK: Integrating graph contextualized knowledge into pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2281–2290, Online. Association for Computational Linguistics.

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

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

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. 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 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1504–1515.

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 *European conference on computer vision*, pages 740–755. Springer.

Xiao Lin and Devi Parikh. 2015. Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2984–2993.

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*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word 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 43–54, Hong Kong, China. Association for Computational Linguistics.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

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*.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Timo Schick and Hinrich Schütze. 2020. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8766–8774. AAAI Press.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Hao Tan and Mohit Bansal. 2020. Vokenization: improving language understanding with contextualized, visual-grounded supervision. *arXiv preprint arXiv:2010.06775*.

Zineng Tang, Jaemin Cho, Hao Tan, and Mohit Bansal. 2021. Vidlankd: Improving language understanding via video-distilled knowledge transfer. *arXiv preprint arXiv:2107.02681*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Peifeng Wang, Nanyun Peng, Filip Ilievski, Pedro Szekely, and Xiang Ren. 2020. Connecting the dots: A knowledgeable path generator for commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4129–4140, Online. Association for Computational Linguistics.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418, Online. Association for Computational Linguistics.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.

Donghan Yu, Chenguang Zhu, Yiming Yang, and Michael Zeng. 2020. Jaket: Joint pre-training of knowledge graph and language understanding. *arXiv preprint arXiv:2010.00796*.

Wenhao Yu, Chenguang Zhu, Yuwei Fang, Donghan Yu, Shuohang Wang, Yichong Xu, Michael Zeng, and Meng Jiang. 2021. Dict-bert: Enhancing language model pre-training with dictionary. *arXiv preprint arXiv:2110.06490*.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2020. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Wangchunshu Zhou, Dong-Ho Lee, Ravi Kiran Selvam, Seyeon Lee, and Xiang Ren. 2021. Pre-training text-to-text transformers for concept-centric common sense. In *International Conference on Learning Representations*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015a. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015b. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Dataset Properties

PIQA is a multiple-choice question answering task, which chooses the most appropriate solution for physical commonsense questions, which may need illustration or description of physical interaction in the real world. VP is to tell if two descriptions are describing the same scene or two different scenes. While they seem like purely textual tasks, they require visual common sense to answer. CSQA is a multiple-choice question answering task that requires commonsense reasoning to answer. It is built from ConceptNet (Speer et al., 2017). OBQA is a multiple-choice question answering task, which is modeled after open book exams on elementary-level core science questions. The task generally requires open book fact but also additional commonsense which can be learnt from scientific illustration. RiddleSense is a multiple-choice riddle-style question answering which requires complex commonsense reasoning ability and understanding of figurative language which may benefit from visual knowledge.