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., 2020a; 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.

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., 2020a) 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., 2020a) 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-

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

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

Interesting facts about orange!
1. Orange elevates mood levels.
2. Orange are often grown in the Mediterranean.
3. Oranges facing the sunnier tend to be sweeter.

Typical facts about orange
1. Orange is a shape of circle.
2. Orange is a color of orange.
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 (Gururangan et al., 2020b; Pruksachatkun et al., 2020b). 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.

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 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 elementary-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 statistics 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 the fully supervised setting, we evaluate models with 3 different random seeds and report the average accuracy. In the low-resource setting, we set the size of the 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 \( x^v \) and caption \( x^c \), a pair \( \{x^v, x^c\} \) \( m \) \( (m \) is the size of the pairs) and image encoder \( f^v \) and text encoder \( f^c \) are given. Note that we use the same text encoder.

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 each token to a corresponding voken. Vokenization trains language models with the voken classification task and MLM.

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 \( f^c \), and a caption \( x^c \), we first get text representations using the encoders \( h^t_i = f^c(x^c) \). Following Gao et al. (2021), we create identical positive sample \( h^{t+}_i \) by different dropout representations. The contrastive loss is defined as follows:

\[
\ell_{\text{MLM}}(x^c_i) = - \log p(x^c_i|x^{\text{masked}}),
\]

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

\[
\ell_{\text{TCL}} = - \log \frac{e^{\text{sim}(h^t_i, h^{t+}_j) + \tau}}{\sum_{j=1}^{N} e^{\text{sim}(h^t_i, h^t_j) + \tau}},
\]

where \( N \) is a batch size and \( \text{sim}() \) represents cosine similarity, i.e., \( \text{sim}(u, v) = u \cdot v / \|u\|\|v\| \). \( \tau \) 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 it “voken classification”. Given text \( x \) and a voken \( v_i \) for the \( i \)-th token, the loss is defined as:

\[
\ell_{\text{voken}} = - \log (p(v_i | x)).
\]

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 \((x_v^i, x_l^i)\), we first get image and text representations using the encoders \( h_v^i = f_v(x_v^i), h_l^i = f_L(x_l^i) \). Then the contrastive learning objective contains two loss functions: an image-to-text contrastive loss \( \ell^{(v,l)} \) and a text-to-image contrastive loss \( \ell^{(l,v)} \). The image-to-text contrastive loss is defined as follows:

\[
\ell_i^{(v,l)} = - \log \frac{e^{\text{sim}(h_v^i, h_l^i)/\tau}}{\sum_j^N e^{\text{sim}(h_v^i, h_l^j)/\tau}},
\]

where \( N \) is a batch size and \( \text{sim}(\cdot, \cdot) \) 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:

\[
\ell_i^{(l,v)} = - \log \frac{e^{\text{sim}(h_l^i, h_v^i)/\tau}}{\sum_j^N e^{\text{sim}(h_l^i, h_v^j)/\tau}}.
\]

The final loss is defined as

\[
L = \frac{1}{N} \sum_{i=1}^N (\ell_i^{(v,l)} + \ell_i^{(l,v)}).
\]

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-modal 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

\[
L = \sum_i^N \left[ \max(0, \alpha - \text{sim}(h_v^i, h_l^i) + \text{sim}(h_v^i, h_l^i)) \right. \\
+ \left. \max(0, \alpha - \text{sim}(h_v^i, h_l^i) + \text{sim}(h_v^i, h_l^i)) \right],
\]

where \( \alpha \) 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).
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 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.

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

4 Experimental Settings

For all the approaches, we use bert-base-uncased (Devlin et al., 2018) as text encoder $f_L$ and ResNetXt101 (Xie et al., 2017) as an image encoder $f_V$. We continue to pre-train the encoders in our experiments. For text knowledge transfer, (1) MLM follows the exact setting of codebase in huggingface\(^2\) which uses dynamic masking strategy to conduct language modeling 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. We use the VL-BERT and Vokenization checkpoints from their official codebases\(^3\). 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 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 ResNetXt101 (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, 4e-5, 4e-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
We argue that if a model obtains better performance when trained with a few training examples, it might benefit from cross-modal knowledge transfer. Overall, this helps better initialization and lets models learn new tasks faster.

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., 2014a; Chen et al., 2015), GenericsKB (Bhakthavatsalam et al., 2020), BooksCorpus (Zhu et al., 2015a), and WikiText103 (Merity et al., 2016). We sample 250k sentences from each corpus for a fair comparison. We notice that caption datasets are useful on OBQA and RiddleSense over BERT (p-value < 0.01). These results suggest that text intermediate pre-training on visual-related datasets helps performance on commonsense reasoning 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 Commonsense reading comprehension tasks over BERT (p-value < 0.01). 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 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 lets 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 suggests that exploiting images in masked language modeling task help transfer the knowledge to language models.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>RTE</th>
<th>MRPC</th>
<th>STS-B</th>
<th>CoLA</th>
<th>SST-2</th>
<th>QMIE</th>
<th>QQP Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-base</td>
<td>78.0</td>
<td>69.0</td>
<td>91.1</td>
<td>57.4</td>
<td>91.3</td>
<td>90.4</td>
<td>89.3</td>
</tr>
<tr>
<td>MLM</td>
<td>68.3</td>
<td>67.0</td>
<td>91.1</td>
<td>91.5</td>
<td>90.4</td>
<td>89.9</td>
<td>89.6</td>
</tr>
<tr>
<td>TCE</td>
<td>56.1</td>
<td>63.0</td>
<td>87.0</td>
<td>51.1</td>
<td>87.1</td>
<td>82.2</td>
<td>79.6</td>
</tr>
<tr>
<td>TCL + MLM</td>
<td>54.8</td>
<td>81.6</td>
<td>97.7</td>
<td>89.9</td>
<td>90.9</td>
<td>89.7</td>
<td>88.6</td>
</tr>
<tr>
<td>TCE + ANS</td>
<td>56.3</td>
<td>83.3</td>
<td>97.7</td>
<td>89.9</td>
<td>90.9</td>
<td>89.7</td>
<td>88.6</td>
</tr>
<tr>
<td>TCE + PSA + ANS</td>
<td>52.3</td>
<td>75.6</td>
<td>81.5</td>
<td>17.4</td>
<td>40.0</td>
<td>35.5</td>
<td>31.7</td>
</tr>
</tbody>
</table>

Table 5 presents results on GLUE benchmark. Bold and underlined numbers refer to the best and second-best performance, respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>RTE</th>
<th>MRPC</th>
<th>STS-B</th>
<th>CoLA</th>
<th>SST-2</th>
<th>QMIE</th>
<th>QQP Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-base</td>
<td>78.0</td>
<td>69.0</td>
<td>91.1</td>
<td>57.4</td>
<td>91.3</td>
<td>90.4</td>
<td>89.3</td>
</tr>
<tr>
<td>Tuckerization</td>
<td>77.4</td>
<td>69.0</td>
<td>91.1</td>
<td>91.5</td>
<td>90.4</td>
<td>89.9</td>
<td>89.6</td>
</tr>
<tr>
<td>VidLanKD</td>
<td>67.5</td>
<td>87.8</td>
<td>99.4</td>
<td>72.7</td>
<td>90.9</td>
<td>89.3</td>
<td>88.6</td>
</tr>
<tr>
<td>VidLanKD variant</td>
<td>64.5</td>
<td>85.0</td>
<td>98.7</td>
<td>78.9</td>
<td>91.1</td>
<td>89.5</td>
<td>87.5</td>
</tr>
<tr>
<td>CMKD (VL-BERT-large):</td>
<td>68.5</td>
<td>85.5</td>
<td>99.3</td>
<td>95.4</td>
<td>90.9</td>
<td>89.7</td>
<td>88.6</td>
</tr>
<tr>
<td>CMCL</td>
<td>63.5</td>
<td>82.8</td>
<td>99.5</td>
<td>91.1</td>
<td>90.4</td>
<td>90.0</td>
<td>89.4</td>
</tr>
<tr>
<td>CMCL + ANS</td>
<td>64.6</td>
<td>83.4</td>
<td>99.6</td>
<td>91.1</td>
<td>90.5</td>
<td>90.2</td>
<td>89.6</td>
</tr>
<tr>
<td>CMCL + PSA + ANS</td>
<td>62.8</td>
<td>86.2</td>
<td>99.0</td>
<td>95.5</td>
<td>90.4</td>
<td>90.5</td>
<td>89.6</td>
</tr>
</tbody>
</table>

Table 5 presents results on GLUE benchmark. Bold and underlined numbers refer to the best and second-best performance, respectively.

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 sample 250k sentences from each corpus for a fair comparison. 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 suggests 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; Phang 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; Calixto 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.

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>PIQA</th>
<th>VP</th>
<th>CSQA</th>
<th>OBQA</th>
<th>RiddleSense</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>64</td>
<td>128</td>
<td>Full</td>
<td>64</td>
<td>128</td>
</tr>
<tr>
<td>BERT-base</td>
<td>52.6</td>
<td>53.8</td>
<td>51.8</td>
<td>52.6</td>
<td>53.8</td>
</tr>
<tr>
<td>MLM</td>
<td>53.3</td>
<td>54.3</td>
<td>53.3</td>
<td>54.3</td>
<td>53.3</td>
</tr>
<tr>
<td>TCL</td>
<td>52.9</td>
<td>53.9</td>
<td>52.9</td>
<td>53.9</td>
<td>52.9</td>
</tr>
<tr>
<td>MLPTLM</td>
<td>53.2</td>
<td>54.2</td>
<td>53.2</td>
<td>54.2</td>
<td>53.2</td>
</tr>
<tr>
<td>TCL</td>
<td>56.1</td>
<td>56.1</td>
<td>56.1</td>
<td>56.1</td>
<td>56.1</td>
</tr>
<tr>
<td>CP</td>
<td>54.1</td>
<td>55.1</td>
<td>54.1</td>
<td>55.1</td>
<td>54.1</td>
</tr>
<tr>
<td>GK</td>
<td>53.0</td>
<td>54.0</td>
<td>53.0</td>
<td>54.0</td>
<td>53.0</td>
</tr>
<tr>
<td>BC</td>
<td>52.9</td>
<td>53.9</td>
<td>52.9</td>
<td>53.9</td>
<td>52.9</td>
</tr>
</tbody>
</table>

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


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.