Imagination-Augmented Natural Language Understanding

Anonymous ACL submission

Abstract

Human brains integrate linguistic and perceptual information simultaneously to understand natural language, and hold the critical ability to render imaginations. Such abilities enable us to construct new abstract concepts or concrete objects, and are essential in involving applicable knowledge to solve problems in low-resource scenarios. However, most existing methods for Natural Language Understanding (NLU) are mainly focused on the textual signals. They do not simulate human visual imagination ability, which hinders models from inferring and learning efficiently from limited data samples. Therefore, we introduce an **I**magination-**A**ugmented **C**ross-modal **E**ncoder (iACE) to solve natural language understanding tasks from a novel learning perspective—imagination-augmented cross-modal understanding. iACE enables visual imagination with the external knowledge transferred from the powerful generative model and pre-trained vision-and-language model. Extensive experiments on GLUE (Wang et al., 2018) and SWAG (Zellers et al., 2018) show that iACE achieves consistent improvement over visually-supervised pre-trained models. More importantly, results in extreme and normal few-shot settings validate the effectiveness of iACE in low-resource natural language understanding circumstances.

1 Introduction

Cognitive neuroscience studies reveal neural activation in vision-related brain areas when reading text (Just et al., 2004) and show a tight relationship between brain areas processing linguistic and visual semantic information (Popham et al., 2021). In addition, visual imagery improves comprehension during human language processing (Sadoski and Paivio, 1994). Such imagination empowers human brains with generalization capability to solve problems with limited supervision or data samples.

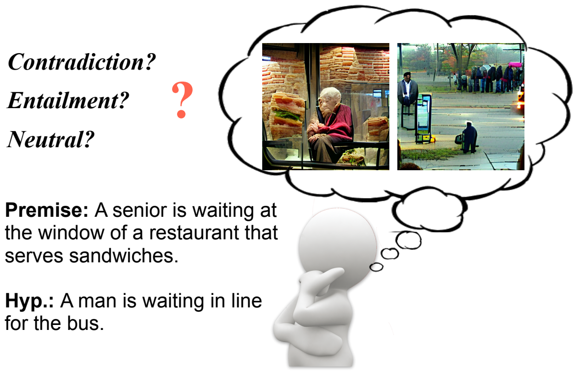

Contradiction?
Entailment?
Neutral?
?
Premise: A senior is waiting at the window of a restaurant that serves sandwiches.
Hyp.: A man is waiting in line for the bus.

Figure 1: Rendering visual imagination is an intuitive way to activate perception for linguistic understanding, e.g. natural language inference.

However, the field of Natural language Understanding has mainly been focused on building machines based solely on language, ignoring the inherently grounded imagination from the external visual world. These studies either learn text-only representations from language corpora (Devlin et al., 2019; Liu et al., 2019; Lan et al., 2020) or implicitly involve retrieved visual supervision in pre-trained language models (Tan and Bansal, 2020). Thus, their approaches appear limited in transferring the connection between language understanding and visual imagination to downstream tasks, which is essential to solving low-resource circumstances. In addition, these methods are limited to text-only augmentations, whereas visual imaginations leverage cross-modal augmentations to deal with low-resource situations.

Human brains are multi-modal, integrating linguistic and perceptual information simultaneously. Intuitively, the machines could achieve a higher-level understanding of natural language and better learning transference by imitating the procedure of human imagination behavior.

Inspired by this, we propose to understand language with the integration of linguistic and perceptual information via introducing imagination supervision into text-only NLU tasks. To imitate the imagination-augmented understanding process as shown in Figure 1 with text-only data, we devise a procedure with two steps: 1) pre-train a visually-supervised Transformer over paired text and images retrieved from large-scale language corpus and image set, and 2) construct the imagination with a generative model and fine-tune on downstream NLU datasets by learning the paired imagination and natural language in a cross-modal embedding. We show a detailed description of the cross-modal imagination process for a specific Natural Language Inference task in Figure 2. In this way, we utilize machine imagination to improve the performance of natural language understanding.

We adopt the few-shot learning setting to study the potential of using less human effort of annotation for our proposed iACE to learn the natural language with the help of imagination. Large margin performance gain in both extreme and normal few-shot settings demonstrate the effectiveness of iACE in solving problems with limited data samples. In full data setting of GLUE (Wang et al., 2018) and SWAG (Zellers et al., 2018), we observe the consistent performance gain of our proposed iACE over the visually-supervised approach (e.g., VOKEN (Tan and Bansal, 2020)) upon four language base models (e.g., BERT, RoBERTa).

In summary, the main contributions of our work are as follow:

- We propose to solve the text-only learning problem in natural language understanding tasks from a novel learning perspective: imagination-augmented cross-modal language understanding.
- To address the problem mentioned above, we devise iACE to generate imaginations in a cross-modal representation space to guide the fine-tuning of the visually supervised language models.
- Experimental results in the few-shot setting validate the consistent superiority of iACE over baselines in tackling the low-resource situation. In full settings, iACE maintains the improvement in GLUE and SWAG.

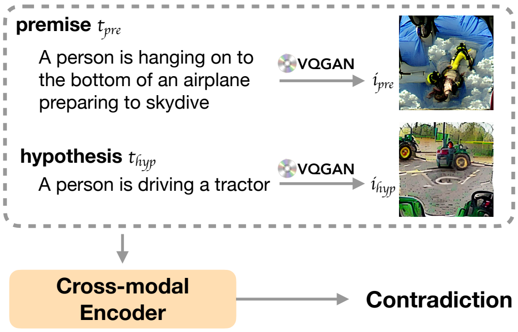

premise t_{pre}
A person is hanging on to the bottom of an airplane preparing to skydive
VQGAN
i_{pre}
hypothesis t_{hyp}
A person is driving a tractor
VQGAN
i_{hyp}
Cross-modal Encoder
Contradiction

Figure 2: A detailed view of our iACE framework fine-tunes on natural language inference task.

2 Related Work

Visually-aided Language Learning Previous research attempt to introduce visual information to improve language learning on various Natural Language Processing (NLP) scenarios, including but not limit to machine translation (Grubinger et al., 2006; Elliott et al., 2016), information retrieval (Funaki and Nakayama, 2015; Gu et al., 2018), semantic parsing (Christie et al., 2016; Shi et al., 2019), natural language inference (Xie et al., 2019), bilingual lexicon learning (Kiela et al., 2015; Vulic et al., 2016), natural language generation evaluation (Zhu et al., 2021), and language representation learning (Lazaridou et al., 2015; Collell et al., 2017; Kiela et al., 2018; Zablocki et al., 2019; Lu et al., 2019; Li et al., 2019; Sun et al., 2019; Huang et al., 2020; Luo et al., 2020; Chen et al., 2020; Li et al., 2020; Tan and Bansal, 2020; Radford et al., 2021). While most of these studies acquire visual information through retrieval from the web or large-scale image sets, a recent line of studies attempt to generate visual supervision from scratch. The visual information can either be provided in the form of representation (Collell et al., 2017; Long et al., 2021) or concrete images (Gu et al., 2018; Zhu et al., 2021). Though previous studies generate machine imagination, they only tackle specific tasks, such as machine translation (Long et al., 2021) or information retrieval (Gu et al., 2018). To the best of our knowledge, we are the first to utilize machine abstract imagination from large pretrained vision and language models to improve general NLU tasks. Recently, VOKEN (Tan and Bansal, 2020) incorporate retrieved token-level visual information into existing transformer models and achieve consistent improvement. iACE is different from this work for two aspects: 1) we explicitly encode visual imagination during fine-tuning. 2) we propose a novel model to borrow knowledge from imagination in both training and inference.

Few-shot Natural Language Understanding Natural Language Understanding (NLU) is a subfield in NLP that involves a broad range of tasks such as question answering, sentiment analysis, and textual entailment. Researchers have collected specific language corpus (Wang et al., 2018; Zellers et al., 2018; McCann et al., 2018; Xu et al., 2020) to train the machines on NLU learning. However, the general language understanding problem remains a challenge. Few-shot learning is a learning paradigm that aims to predict the correct class of instances with a relatively small amount of labeled training examples (Fink, 2004; Fei-Fei et al., 2006). It has been receiving increasing attention for its potential in reducing data collection effort and computational costs and extending to rare cases. To deal with data-scarcity in NLU problems, previous research introduces external knowledge (Sui et al., 2021), utilizes meta-learning (Geng et al., 2019; Bansal et al., 2020; Han et al., 2021) and adopts data augmentation to generate labeled utterances for few-shot classes (Murty et al., 2021; Wei et al., 2021). Recent studies (Radford et al., 2019; Brown et al., 2020) have shown that large-scale pre-trained language models are able to perform NLU tasks in a few-shot learning manner. The pre-trained multimodal models also display similar few-shot learning ability (Tsimpoukelli et al., 2021). Different from previous studies on pre-trained multimodal Transformers that target solving multimodal tasks, our study introduces imagination from the visual world into language models and aims at improving NLU tasks.

3 Our Approach

We illustrate how we solve the existing text-only learning problem in natural language understanding tasks as the Imagination-augmented Cross-modal Language Understanding (ICLU) problems in Section 3.1. Then we give a detailed illustration of our proposed iACE's architecture in Section 3.2. Finally, we describe the procedure and training protocol of the perceptual-enhanced linguistic understanding paradigm in Section 3.3.

3.1 Problem Definition

NLU is concerned with understanding the semantic meaning of the given utterances. Data pieces for NLU can be structured as $(x_{context}, \mathscr{X}, y)$, where $x_{context}$ represents the text context, $\mathscr{X} = \{x_1, x_2, ..., x_m, m \in \mathbb{N}\}$ denote a set of text snippets, and m denotes the number of text samples for a specific task. The model learns to predict the ground truth label y, which is either regression or a classification label. While NLU is usually regarded as a language-only task, we attempt to solve it from a cross-modal perspective by introducing the novel ICLU problem.

In our ICLU problem, data pieces are structured as $(x_{context}, i_{context}, \mathscr{X}, \mathscr{I}, y)$, in which $i_{context}$ represents the visual context related to the text context, and $\mathscr{I} = \{i_1, i_2, ..., i_n, n \in \mathbb{N}\}$ denotes the imagination set. The "imagination" refers to the images that are visualized from the text. Here, n is the number of visualized sentences for a specific task, which is the same as m by default.

To solve this problem, we devise a novel iACE to construct imagination from textual data and learn the bi-directional alignment between the imagination and text. Specifically, for each piece of text x_j in the sentence set $\mathscr{X}$, we first follow (Esser et al., 2020; Radford et al., 2021) and use a generative model to render a descriptive illustration i_j. The visualized imagination will later serve as the visual input in the ICLU problem.

3.2 Model Architecture

Overview Figure 3 provides an overview of the iACE framework. iACE consists of two modules: 1) the imagination generator G, 2) the imagination-augmented cross-modal encoder E_c. Given the textual sentence $x = \{w_1, w_2, ..., w_k, k \in \mathbb{N}\}$ (w_j denotes the j-th token in the sentence), G generates corresponding visual imagination i. The cross-modal encoder then encodes x and i as $\boldsymbol{t}$ and $\boldsymbol{v}$, respectively. iACE explicitly provides imagination supervision to the visually-supervised Transformer during fine-tuning on downstream NLU tasks.

Imagination Generator Previous studies introduce visual supervision through retrieval from the web or image sets. However, it is hard to find visuals that perfectly match the topics discussed in each text snippet, especially for the relatively complicated text input for the NLU tasks. Such misalignment between the input text and the retrieved visuals might hinder the model from general language understanding learning. Out of consideration for cross-modal feature alignment, we choose to render specific visualization corresponding to each piece of input text from scratch. Specifically, we construct imagination of the textual input with a large-scale vision and language model guided

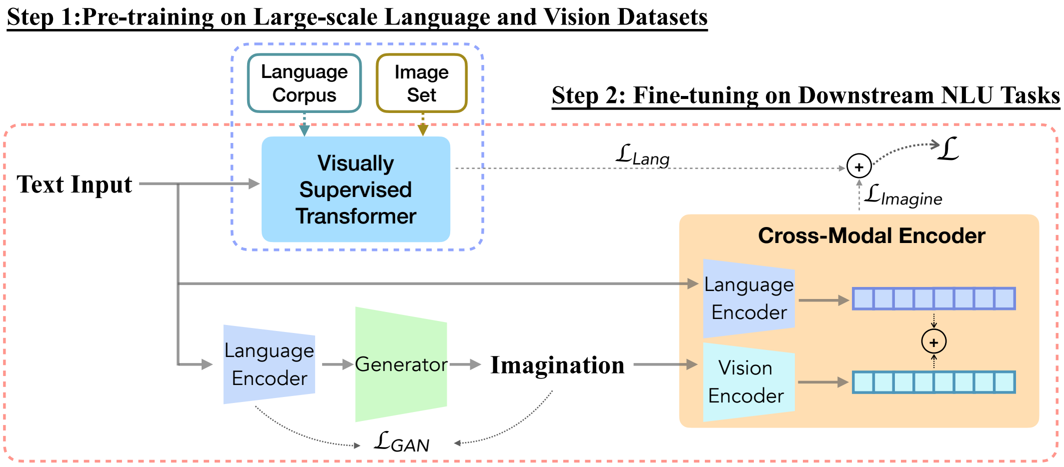


Figure 3: **Overview of iACE.** The generator G visualize imaginations close to the encoded texts by minimizing $\mathscr{L}_{GAN}$. The cross-modal encoder E_c learns imagination-augmented language representation. Two-step learning procedure consists of: 1) pre-train a Transformer with visual supervision from large-scale language corpus and image set, 2) fine-tune the visually supervised pre-trained Transformer and the imagination-augmented cross-modal encoder on downstream tasks.

generative framework - VQGAN+CLIP[1]. For each piece of input text x, we treat it as the prompt and use the VQGAN (Esser et al., 2020) model to render the imagination i with 128×128 resolution and 200-step optimization. At each optimization step, we use the CLIP (Radford et al., 2021) model to assess how well the generated image corresponds to the text. To be specific, CLIP encodes the input text x and the corresponding imagination i as $\boldsymbol{t}$ and $\boldsymbol{v}$, and the training objective is to minimize the distance between $\boldsymbol{t}$ and $\boldsymbol{v}$ in the cross-modal embedding space.

$$\mathscr{L}_{GAN} = 2[\arcsin(\frac{1}{2}\|\boldsymbol{t} - \boldsymbol{v}\|)]^2 \tag{1}$$

Cross-modal Encoder We adopt CLIP as the cross-modal encoder to encode the input text and the generated imaginations. CLIP (Radford et al., 2021) is trained on large-scale image-text pairs and is able to align visual and textual input in the embedding space. Specifically, we use the $ViT-B/32$ version of Vision Transformer as the image encoder, and Transformer (Vaswani et al., 2017) with the architecture modifications described in (Radford et al., 2019) as the text encoder. For each modality, the self-attention (SA) module is applied to model the regions of imagination or the words of the text as follow:

$$SA(F) = concat(softmax\frac{FW_j^Q FW_j^{K^{\mathrm{T}}}}{\sqrt{d_k}} FW_j^V, \ldots)W \tag{2}$$

where F denotes the set of regions of the imagination or the words of the textual sentence. W_j^Q, W_j^K, and W_j^V represents the weight in the j-th head for query, key and value respectively. d_k is the dimension of the embedding. W is the weight matrix for multiple heads.

To solve the ICLU problem, we learn the bi-directional relationship between the text input and the visualized imagination. We apply late fusion on the text feature $\boldsymbol{t}$ and visual feature $\boldsymbol{v}$ to construct the cross-modal feature. Given the set of visual features $S_{\boldsymbol{v}}$ and textual features $S_{\boldsymbol{t}}$, the fused embedding X_S can be given with:

$$X_S = [ReLU(W_t S_{\boldsymbol{t}} + b_t), ReLU(W_j S_{\boldsymbol{v}} + b_j)] \tag{3}$$

where W and b are of two separate fully connected layers to the visual and text embeddings. The fused embeddings X_S will go through two fully connected layers before we receive the final imagination-augmented language representation.

Visually-supervised Transformer We implement the visually-supervised Transformer language model proposed in Tan and Bansal (2020). The model architecture is a BERT-like pure-language-based masked language model.

3.3 Learning Procedure

We introduce a novel paradigm to better understand natural language by incorporating existing language models with visual imagination. As shown in Figure 3, the procedure consists of two steps: (1)

[1] https://github.com/nerdyrodent/VQGAN-CLIP

pre-train the visually-supervised Transformer, and (2) fine-tune the framework with imagination on downstream tasks.

Step 1: Visually-supervised Pre-training We pre-train a visually-supervised Transformer following the scheme proposed in VOKEN (Tan and Bansal, 2020), which extrapolates cross-modal alignments to language-only data by contextually mapping language tokens to the related images. In addition to masked language modeling, VOKEN proposed a voken classification task: given a set of tokens with masks, the model is asked to predict the best-matching image (the voken) for each tokens. The pre-training loss can be given as:

$$\mathscr{L} = -\lambda_1 \sum_{w_j \in \hat{s}} \log q_j(w_j|\check{s}) - \lambda_2 \sum_{w_j \in \hat{s}} \log p_j(v(w_j;s)|\check{s}) \tag{4}$$

Here s is the token set, $\hat{s}$ is the masked tokens, and $\check{s}$ is the unmasked tokens. The q_j and p_j represent the conditional probability distribution of the j-th token given the token w_j and voken $v(w_j;s)$ respectively, and λ_1 and λ_2 are the balance factor of the masked language modeling task and the voken-classification task. The cross-modal classification task enables the model to learn the matching between the tokens from the language corpus (e.g., wiki) and its most-related images from the image set (e.g., MSCOCO).

Step 2: Imagination-augmented Fine-tuning We use GLUE (Wang et al., 2018) and SWAG (Zellers et al., 2018) as the downstream datasets in the following sections. Our proposed iACE learns to minimize the cross-entropy loss below:

$$\mathscr{L}_{Imagine} = -\sum_{j=1}^{|D|}\sum_{k=1}^{K} y_k \log p_k(d_j(\boldsymbol{t};\boldsymbol{v})|D) \tag{5}$$

where j denotes the j-th data sample in dataset D, and K os the class number. The p_k represents the conditional probability distribution of d_j. During fine-tuning, the visually-supervised Transformer language model only relied on the textual input to make predictions. The loss are computed as:

$$\mathscr{L}_{Lang} = -\sum_{j=1}^{|D|}\sum_{k=1}^{K} y_k \log p_k(d_j(\boldsymbol{t})|D) \tag{6}$$

Notice that we use MSE loss for the regression task. The imagination-augmented loss and pure-language based loss are summed up with a balance factor λ in a jointly training schema as:

$$\mathscr{L} = \lambda \mathscr{L}_{Imagine} + (1-\lambda)\mathscr{L}_{Lang} \tag{7}$$

We use Adam Optimizer with a learning rate $1e-4$ for the GLUE benchmark and $2e-5$ for the SWAG dataset. We discuss more details in Section 4.

4 Experiments

4.1 Experimental Setup

Datasets & Metric We conduct experiments to evaluate the performance of our proposed method over SST-2 (Socher et al., 2013), QNLI (Rajpurkar et al., 2016), QQP (Iyer et al., 2017), MultiNLI (Williams et al., 2018), MRPC (Dolan and Brockett, 2005), STS-B (Agirre et al., 2007) from GLUE (Wang et al., 2018) Benchmark, and SWAG (Zellers et al., 2018) dataset. We construct few-shot setting subsets by taking 0.1%, 0.3%, and 0.5% of training instances as the Extreme Few-shot Setting, and 1%, 3%, and 5% as the Normal Few-shot Setting. We train the model with the subsets and evaluate its performance on the complete development set. We use accuracy as the default evaluation metric and compare such results in the following sections.

Baselines We choose BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) as the base language models, and apply our iACE framework on top of their small and base architectures for comparison. A recent study proposes a visually-supervised language model VOKEN (Tan and Bansal, 2020) that introduces visual supervision into language model pre-training by borrowing external knowledge from retrieved images of the tokens. In natural language understanding tasks, VOKEN achieved improvements over language-based baselines BERT and RoBERTa. Thus we also use VOKEN built upon these language-based models as a set of powerful baselines. In the following experiments, each model is first pre-trained with visual supervision introduced in (Tan and Bansal, 2020) upon the four base models (BERT$_{small}$, BERT$_{base}$, RoBERTa$_{small}$ and RoBERTa$_{base}$). Then the models will be fine-tuned on downstream tasks.

Notice that base models and VOKEN use pure-language training objectives during fine-tuning. Neither of them utilizes the visual signals inherent in the downstream language corpora. In contrast, our iACE explicitly introduces visual imagination supervisions into fine-tuning and inference stages.

	SST-2			QNLI			QQP			MNLI		
Extreme Few-shot	0.1%	0.3%	0.5%	0.1%	0.3%	0.5%	0.1%	0.3%	0.5%	0.1%	0.3%	0.5%
$VOKEN(Bert_{base})$	54.70	77.98	80.73	50.54	51.60	61.96	44.10	60.65	65.46	37.31	54.62	58.79
$iACE(Bert_{base})$	**77.98**	**80.96**	**81.42**	**51.64**	**58.33**	**64.03**	**49.36**	**63.67**	**71.17**	**40.07**	**56.49**	**59.57**
$VOKEN(Roberta_{base})$	70.99	71.10	77.86	54.37	62.23	65.78	62.32	67.25	70.18	48.59	49.76	58.23
$iACE(Roberta_{base})$	**75.34**	**78.66**	**83.60**	**54.79**	**65.03**	**65.83**	**65.43**	**68.11**	**70.77**	**48.94**	**52.74**	**59.39**
Normal Few-shot	1%	3%	5%	1%	3%	5%	1%	3%	5%	1%	3%	5%
$VOKEN(Bert_{base})$	81.40	86.01	84.75	64.17	77.36	80.19	72.55	78.37	80.50	60.45	62.73	72.35
$iACE(Bert_{base})$	**82.45**	**87.04**	**86.47**	**65.09**	**79.54**	**80.52**	**74.31**	**78.69**	**80.52**	**62.15**	**70.43**	**73.73**
$VOKEN(Roberta_{base})$	83.78	84.08	87.61	75.00	81.16	81.23	73.14	79.09	79.63	63.51	70.68	74.02
$iACE(Roberta_{base})$	**83.83**	**84.63**	**89.11**	**79.35**	**81.41**	**81.65**	**73.72**	**79.38**	**79.81**	**65.66**	**70.76**	**74.10**

Table 1: **Model-agnostic Improvement in Few-shot Setting.** iACE and VOKEN upon BERT and RoBERTa base size architecture are fine-tuned in Extreme Few-shot (0.1%, 0.3%, 0.5%) and Normal Few-shot setting (1%, 3%, 5%). For the few-shot setting, we use large and stable datasets from GLUE Benchmark. We compare accuracy on SST-2, QNLI, QQP, and MNLI and the average of accuracy and F1 score on QQP. **BEST** results are highlighted.

Implementation Details We train RoBERTa with the same configurations as a robustly optimized pre-training approach based on BERT of the same size. BERT$_{small}$ has 6 repeating layers, 512 hidden dimension. BERT$_{base}$ has 12 repeating layers, 768 hidden dimension.

The imagination of the texts is generated interactively by using VQGAN+CLIP, with 128 × 128 size, 500 iterations. We use pre-trained VQGAN (imagenet$_{f16}$) and CLIP (ViT-B/32). We leverage CLIP (ViT-B/32) as our language and vision model for premise and hypothesis, and imagination of them. The text and image dimension is 512. The dropout rate is set to 0.1. We use Cross-Entropy loss for our cross-modal classification. Each model was first pre-trained on 4 TITAN RX GPUs for 30 epochs with early stopping and a batch size of 32 and a sequence length of 126. The optimizer used is Adam with a learning rate of $2e-4$ and a weight decay of 0.01. The models are then fine-tuned on GLUE benchmark and SWAG dataset for 3 epochs with 32 batch size. We adopt the joint training strategy for our proposed iACE and visually supervised transformer during fine-tuning. The learning rate of the Adam optimizer is set as $1e-4$ and $2e-5$ for GLUE and SWAG, respectively.

4.2 Few-shot Learning Results

We claim that introducing imagination into language processing helps the existing language-based system tackle the low-resource situation. Thus, the automatically generated imagination helps reduce the human effort to annotate textual data. To verify this, we define two situations, a normal few-shot setting, and an extreme few-shot setting. For the normal few-shot setting, we keep 1%, 3%, and 5% of the training dataset for each task in GLUE Benchmark. For the extreme few-shot setting, we keep a lower number of the training dataset, which is reduced to 0.1%, 0.3%, and 0.5% of the training dataset. We train the models with the same configuration under these two settings and compare them with visually supervised transformer baselines to confirm the benefit that our proposed iACE brings to the few-shot situation.

Results of the few-shot setting are reported in Table 1. Following (Tan and Bansal, 2020), we only report the four largest and stable tasks in GLUE for the model-agnostic comparison. We report the accuracy for SST-2, QNLI, MNLI. For QQP and MRPC, we report the average of F1 and accuracy. For SWAG, we report the correlation. We observe that the imagination information remarkably helps with both the normal few-shot curriculum and extreme few-shot curriculum. We assume the imagination-augmented fine-tuning successfully transfers the language understanding from the large-scale vision and language model. Thus iACE achieves consistent performance gain and shows great superiority of generalization and transferring ability.

4.3 Ablation Studies

We conduct ablation studies over both the method side and data side to validate their contribution to our proposed iACE.

Method Design Ablation Two method variants of our imagination-augmented encoder are built as baselines to validate the importance of our

		SST-2			QNLI			QQP			MNLI			ALL
Base Model	**Method**	0.1%	1.0%	3.0%	0.1%	1.0%	3.0%	0.1%	1.0%	3.0%	0.1%	1.0%	3.0%	**Avg.**
BERT$_{base}$	Direction	49.01	79.59	87.15	51.31	52.55	66.90	56.74	31.58	31.59	32.73	61.54	70.72	55.95
BERT$_{base}$	Unify	48.96	77.98	86.92	50.54	52.02	67.20	55.29	56.93	79.09	39.05	63.29	70.86	62.34
BERT$_{base}$	iACE	77.98	82.45	87.04	51.64	65.09	79.54	49.36	74.31	78.69	40.07	62.15	70.43	68.23
RoBERTa$_{base}$	Direction	72.71	80.38	84.63	54.91	74.68	78.58	61.57	74.68	31.59	32.95	61.96	70.62	64.94
RoBERTa$_{base}$	Unify	75.11	80.04	88.07	53.62	74.64	78.47	64.94	74.85	76.84	51.12	65.42	70.74	71.15
RoBERTa$_{base}$	iACE	75.34	83.83	84.63	54.79	79.35	81.41	65.43	73.72	79.38	48.94	65.66	70.76	71.93

Table 2: **Method Design Ablation in Few-shot Setting.** We compare the results of two variants over 0.1%, 1.0%, 3.0% of SST-2, QNLI, QQP and MNLI dataset. Details of *Direction* and *Unify* are illustrated in Section 4.3.

		Extreme Few-shot (0.1%)				**Normal Few-shot (3.0%)**				**ALL**
Base Model	**Composition**	SST-2	QNLI	QQP	MNLI	SST-2	QNLI	QQP	MNLI	**Avg.**
BERT$_{base}$	Visual-Only	59.97	50.56	49.01	39.05	86.81	67.23	79.06	70.80	62.81
BERT$_{base}$	Visual+Textual (VT)	53.89	50.54	49.15	38.83	87.04	66.81	79.16	70.77	62.02
BERT$_{base}$	Bi-directional VT	77.98	51.64	49.36	40.07	87.04	79.54	78.69	70.43	66.84
RoBERTa$_{base}$	Visual-Only	75.11	54.18	65.01	47.22	84.17	79.88	76.88	70.56	69.12
RoBERTa$_{base}$	Visual+Textual (VT)	74.20	53.98	65.43	47.35	83.94	79.96	76.87	70.73	69.05
RoBERTa$_{base}$	Bi-directional VT	75.34	54.79	65.43	48.94	84.63	81.41	79.38	70.76	70.08

Table 3: **Imagination Composition Ablation in Few-shot Setting.** *Bi-directional VT* represents the full input for iACE. More details about *Visual Only* and *Visual+Textual* are illustrated in Section 4.3.

bi-directional cross-modal imagination design in iACE. The variants are built upon RoBERTa$_{base}$ and BERT$_{base}$ base models. Specifically, we develop variant *Direction* and *Unify*. *Direction* represent alignment between text input and imagination into a directional embedding as FUSE($\boldsymbol{t}_{sen1} - \boldsymbol{i}_{sen1}$, $\boldsymbol{t}_{sen2} - \boldsymbol{i}_{sen2}$). *Unify* encode the text and imagination, considering the direction from vision to language by encoding as FUSE($\boldsymbol{t}_{sent1}$, $\boldsymbol{t}_{sent2}$, $\boldsymbol{i}_{sent1}$, $\boldsymbol{i}_{sent2}$). While *iACE* consider direction from visoin to language and language to vision by encoding as the combination of FUSE($\boldsymbol{t}_{sent1}$, $\boldsymbol{i}_{sent2}$) and FUSE($\boldsymbol{i}_{sent1}$, $\boldsymbol{t}_{sent2}$).

As shown in Table 2, our bi-directional imagination and language learning achieve stable and best average performance. These results indicate that our bi-directional imagination method design obtain generalization and transferring ability. We assume iACE benefits from both learning from language to vision and learning from vision to language simultaneously.

Imagination Composition Ablation The composition of the imagination is essential for the performance. To further study the importance of full imagination, we ablate the data side by constructing a visual-only imagination denoted as *Visual Only* and a single directional imagination input denoted as *Visual+Textual*. *Visual Only* and *Visual+Textual* represent the imagination model use visual pairs ($\boldsymbol{i}_{sent1}, \boldsymbol{i}_{sent2}$) and one direction visual and textual pairs ($\boldsymbol{i}_{sent1}, \boldsymbol{t}_{sent2}$) as input respectively. Our full approach use *Bi-directional VT* which takes ($\boldsymbol{i}_{sent1}, \boldsymbol{t}_{sent2}$) and ($\boldsymbol{t}_{sent1}, \boldsymbol{i}_{sent2}$) as input.

Results are reported in Table 3 for Extreme Few-shot setting and normal few-shot setting. We observe *Bi-directional VT* data input achieve the most stable and the best average performance. Results show the importance of bi-directional imagination from all the textual input to construct an imagination-augmented cross-modal encoder.

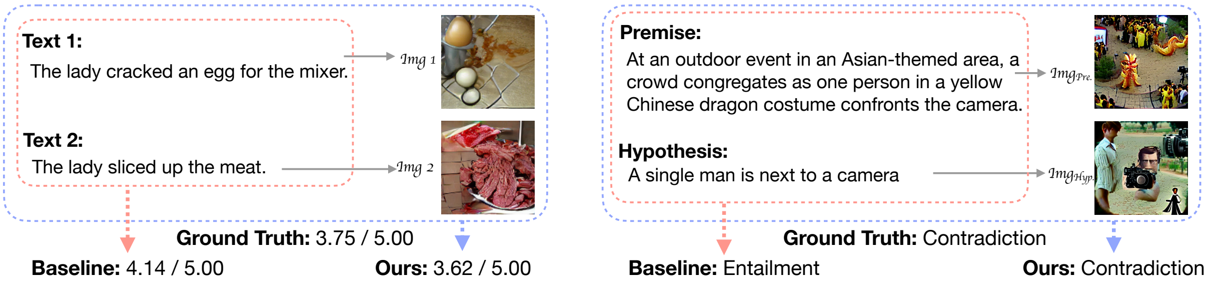
4.4 Model-agnostic Improvement

iACE is a model-agnostic training paradigm that could help existing models achieve consistent gain over GLUE and SWAG with both the few-shot setting and full data setting. To validate such model-agnostic effectiveness of our proposed novel paradigm in processing natural language, we compare the performance with two language models (BERT and RoBERTa) of two architectures ("6L/512H" and "12L/768H"), and a strong visually supervised pre-trained baseline VOKEN (Tan and Bansal, 2020).

Table 4 shows the metric comparison on GLUE and SWAG. The base models are trained with a masked language model. The VOKEN model is pre-trained with a masked language model with an

Base Model	Method	SST-2	QNLI	QQP	MNLI	MRPC	STS-B	SWAG	Avg.
BERT$_{small}$	VOKEN	89.7	85.0	87.3	78.6	78.2	80.4	57.6	79.5
BERT$_{small}$	iACE	**89.8**	**86.2**	**87.7**	**78.9**	**78.4**	**82.7**	**57.9**	**80.2**
BERT$_{base}$	VOKEN	**92.2**	88.6	88.6	82.6	83.5	86.0	70.6	84.6
BERT$_{base}$	iACE	91.7	**88.6**	**89.1**	**82.8**	**85.8**	**86.6**	**70.8**	**85.1**
RoBERTa$_{small}$	VOKEN	87.8	85.1	85.3	76.5	78.5	78.6	53.6	77.9
RoBERTa$_{small}$	iACE	**89.2**	**85.1**	**86.5**	**76.8**	**79.0**	**78.7**	**53.7**	**78.3**
RoBERTa$_{base}$	VOKEN	90.5	**89.2**	87.8	81.0	87.0	86.9	68.5	84.4
RoBERTa$_{base}$	iACE	**91.6**	89.1	**87.9**	**82.6**	**87.7**	**86.9**	**68.5**	**84.9**

Table 4: **Model-agnostic Improvement in Full Data Setting.** Results of iACE and VOKEN upon BERT and RoBERTa of small($6L/512H$) and base($12L/768H$) architecture are reported. The models are fine-tuned over GLUE Benchmark and SWAG with access to the full dataset. **BEST** results are highlighted.



Text 1: The lady cracked an egg for the mixer.

Text 2: The lady sliced up the meat.

Ground Truth: 3.75 / 5.00

Baseline: 4.14 / 5.00 **Ours:** 3.62 / 5.00

(a) STS-B

Premise: At an outdoor event in an Asian-themed area, a crowd congregates as one person in a yellow Chinese dragon costume confronts the camera.

Hypothesis: A single man is next to a camera

Ground Truth: Contradiction

Baseline: Entailment **Ours:** Contradiction

(b) SNLI

Figure 4: Case studies on the STS-B and SNLI tasks. The baseline models yield predictions solely based on the text input, while our approach takes both the text input and corresponding visualization into consideration. On both tasks, our iACE gives predictions that are more aligned with the ground truth.

additional voken-classification task as introduced visual supervision. iACE achieves model-agnostic improvement over the model that solely fine-tune based on textual information, including the pure-language-based model and visually supervised pre-trained model. The gain is consistently observed from different architectures of models.

4.5 Case Study

Figure 4 lists out our examples for the case study. We show the results from the natural language inference and sentence similarity task. We use examples from the STS-B and SNLI datasets. Our contextual imagination describes the textual input as expected and provides an external prediction reference.

For example (a), given the structurally diversified sentence and low n-grams overlaps but high semantic similarity, we observe the pure language-based model predicts the wrong label as well. While the imagination helps the model capture the semantic similarity between two textual inputs via comparing the cross-modal semantics with the imagination information. From example (b), we observe the pure language-based model predicts the wrong label based on the similar sentence structure and high n-grams overlaps. While the imagination helps the model capture the difference between the similar premise and hypothesis text.

5 Conclusion

We treat the text-only learning problem in Natural Language Understanding tasks as a cross-modal language understanding problem with generated imagination as supervisions. In this scenario, the task aims to bridge the gap between the human and the agent language understanding in both linguistic and perceptual procedures. To address the proposed problem, we devised a model-agnostic learning paradigm iACE. Specifically, we build the imagination of the downstream dataset using an interactive generative approach with guidance from a self-supervised pre-trained large-scale image and text model. Our proposed iACE surpassed baselines of two architecture sizes by a large margin in the few-shot setting. The improvement is consistently observed over pure-language baselines (BERT and RoBERTa) and visually supervised VOKEN on the GLUE and SWAG dataset. The results show the superiority of our iACE in language understanding and handling low-resource circumstances.

Ethical Statement

In this study, we only cover NLU datasets with English annotations. Such limitation is since the large-scale pre-trained multimodal models used in our studies, such as CLIP and VQGAN, are only trained on English corpus as of the date we conduct the experiments [2].

This study use CLIP and VQGAN to render images given the text prompt. Suppose there exists any bias in the training dataset for the large-scale pre-trained multimodal models used in our study. In that case, our "imagination" approach may face an issue of fairness since the visual generative model might be more likely to illustrate specific types of images that it has seen in the training data. Moreover, if the training dataset for CLIP or VQGAN contains any personal information, then our "imagination" approach may strike a threat on privacy leakage given certain triggers or prompts. Even though we did not witness such issues in our study, we should keep in mind that the aforementioned behaviors would impair iACE's effectiveness.

References

Eneko Agirre, Llu'is M'arquez, and Richard Wicentowski. 2007. Semantic textual similarity benchmark. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic. Computational Linguistics.

Trapit Bansal, Rishikesh Jha, and Andrew McCallum. 2020. Learning to few-shot learn across diverse natural language classification tasks. In *COLING*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Gordon A. Christie, Ankita Gajanan 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. *ArXiv*, abs/1604.02125.

Guillem Collell, Ted Zhang, and Marie-Francine Moens. 2017. Imagined visual representations as multimodal embeddings. In *AAAI*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*. Asia Federation of Natural Language Processing.

Desmond Elliott, Stella Frank, K. Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *ArXiv*, abs/1605.00459.

Patrick Esser, Robin Rombach, and Björn Ommer. 2020. Taming transformers for high-resolution image synthesis.

Li Fei-Fei, Rob Fergus, and Pietro Perona. 2006. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:594–611.

Michael Fink. 2004. Object classification from a single example utilizing class relevance metrics. In *NIPS*.

Ruka Funaki and Hideki Nakayama. 2015. Image-mediated learning for zero-shot cross-lingual document retrieval. In *EMNLP*.

Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, and Jian Sun. 2019. Induction networks for few-shot text classification. In *EMNLP*.

Michael Grubinger, Paul D. Clough, Henning Müller, and Thomas Deselaers. 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems.

Jiuxiang Gu, Jianfei Cai, Shafiq R. Joty, Li Niu, and G. Wang. 2018. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7181–7189.

Chengcheng Han, Zeqiu Fan, Dongxiang Zhang, Minghui Qiu, Ming Gao, and Aoying Zhou. 2021. Meta-learning adversarial domain adaptation network for few-shot text classification. In *FINDINGS*.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *ArXiv*, abs/2004.00849.

[2] As of Dec. 2021.

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

M. Just, S. Newman, T. Keller, A. McEleney, and P. Carpenter. 2004. Imagery in sentence comprehension: an fmri study. *NeuroImage*, 21:112–124.

Douwe Kiela, Alexis Conneau, A. Jabri, and Maximilian Nickel. 2018. Learning visually grounded sentence representations. In *NAACL*.

Douwe Kiela, Ivan Vulic, and Stephen Clark. 2015. Visual bilingual lexicon induction with transferred convnet features. In *EMNLP*.

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. *ArXiv*, abs/1909.11942.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *NAACL*.

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

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*, abs/1908.03557.

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*, abs/1907.11692.

Quanyu Long, Mingxuan Wang, and Lei Li. 2021. Generative imagination elevates machine translation. In *NAACL*.

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

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou. 2020. Univilm: A unified video and language pre-training model for multimodal understanding and generation. *ArXiv*, abs/2002.06353.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *ArXiv*, abs/1806.08730.

Shikhar Murty, Tatsunori B. Hashimoto, and Christopher D. Manning. 2021. Dreca: A general task augmentation strategy for few-shot natural language inference. In *NAACL*.

Sara F Popham, Alexander G Huth, Natalia Y Bilenko, Fatma Deniz, James S Gao, Anwar O Nunez-Elizalde, and Jack L Gallant. 2021. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature neuroscience*, 24(11):1628—1636.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Mark Sadoski and A. Paivio. 1994. A dual coding view of imagery and verbal processes in reading comprehension.

Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. 2019. Visually grounded neural syntax acquisition. In *ACL*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, A. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*.

Dianbo Sui, Yubo Chen, Binjie Mao, Delai Qiu, Kang Liu, and Jun Zhao. 2021. Knowledge guided metric learning for few-shot text classification. *ArXiv*, abs/2004.01907.

Chen Sun, Austin Myers, Carl Vondrick, Kevin P. Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7463–7472.

Haochen Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding via contextualized, visually-grounded supervision. In *EMNLP*.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *ArXiv*, abs/2106.13884.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Ivan Vulic, Douwe Kiela, Stephen Clark, and Marie-Francine Moens. 2016. Multi-modal representations for improved bilingual lexicon learning. In *ACL*.

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. *ArXiv*, abs/1804.07461.

Jason Wei, Chengyu Huang, Soroush Vosoughi, Yu Cheng, and Shiqi Xu. 2021. Few-shot text classification with triplet networks, data augmentation, and curriculum learning. In *NAACL*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *ArXiv*, abs/1901.06706.

Liang Xu, Xuanwei Zhang, Lu Li, Hai Hu, Chenjie Cao, Weitang Liu, Junyi Li, Yudong Li, Kai Sun, Yechen Xu, Yiming Cui, Cong Yu, Qianqian Dong, Yin Tian, Dian Yu, Bo Shi, Jun jie Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhen-Yi Yang, Kyle Richardson, and Zhenzhong Lan. 2020. Clue: A chinese language understanding evaluation benchmark. *ArXiv*, abs/2004.05986.

Éloi Zablocki, Patrick Bordes, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari. 2019. Incorporating visual semantics into sentence representations within a grounded space. *ArXiv*, abs/2002.02734.

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

Wanrong Zhu, Xin Eric Wang, An Yan, Miguel P. Eckstein, and William Yang Wang. 2021. Imagine: An imagination-based automatic evaluation metric for natural language generation. *ArXiv*, abs/2106.05970.