Large-Scale Adversarial Training for Vision-and-Language Representation Learning

Zhe Gan[1], **Yen-Chun Chen**[1], **Linjie Li**[1], **Chen Zhu**[2], **Yu Cheng**[1], **Jingjing Liu**[1]
[1]Microsoft Dynamics 365 AI Research, [2]University of Maryland, College Park
{zhe.gan,yen-chun.chen,lindsey.li,yu.cheng,jingjl}@microsoft.com
chenzhu@cs.umd.edu

Abstract

We present VILLA, the first known effort on large-scale adversarial training for vision-and-language (V+L) representation learning. VILLA consists of two training stages: (i) task-agnostic adversarial pre-training; followed by (ii) task-specific adversarial finetuning. Instead of adding adversarial perturbations on image pixels and textual tokens, we propose to perform adversarial training in the embedding space of each modality. To enable large-scale training, we adopt the "free" adversarial training strategy, and combine it with KL-divergence-based regularization to promote higher invariance in the embedding space. We apply VILLA to current best-performing V+L models, and achieve new state of the art on a wide range of tasks, including Visual Question Answering, Visual Commonsense Reasoning, Image-Text Retrieval, Referring Expression Comprehension, Visual Entailment, and NLVR^2.[1]

1 Introduction

Inspired by the success of BERT [11] on natural language understanding, there has been a surging research interest in developing multimodal pre-training methods for vision-and-language representation learning (*e.g.*, ViLBERT [35], LXMERT [58], and UNITER [10]). When finetuned on downstream tasks, these pre-trained models have achieved state-of-the-art performance across diverse V+L tasks, such as Visual Question Answering (VQA) [4, 15], Visual Commonsense Reasoning (VCR) [72], and Referring Expression Comprehension [69]. However, due to the immense capacity of large-scale pre-trained models yet limited amount of labeled data in downstream tasks, aggressive finetuning often falls into the overfitting trap [22]. *Adversarial training*, a method to combat adversarial attacks in order to create robust neural networks [57, 14], has recently shown great potential in improving the generalization ability of pre-trained language models [76, 22] and image classifiers [64]. A natural question that came to our mind: can we apply similar adversarial training techniques to V+L problems to improve model performance?

We propose VILLA (**Vi**sion-and-**L**anguage **L**arge-scale **A**dversarial training), which advocates the use of adversarial training for V+L representation learning. As illustrated in Figure 1, VILLA consists of two training stages: (i) *task-agnostic* adversarial pre-training (APT); followed by (ii) *task-specific* adversarial fine-tuning (AFT). Intuitively, if well-designed, multimodal pre-training tasks such as image-conditioned masked language modeling and image-text matching can resonate well with many downstream tasks that require visual grounding and reasoning abilities. This leads to our hypothesis that the improved generalization ability of pre-trained models learned during APT stage can be readily transferred to the AFT stage for diverse tasks. In other words, APT is able to uniformly lift model performance for all downstream tasks in a task-agnostic way, while AFT can further enhance the finetuned models by leveraging task-specific supervision signals.

[1] Code is available at `https://github.com/zhegan27/VILLA`.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

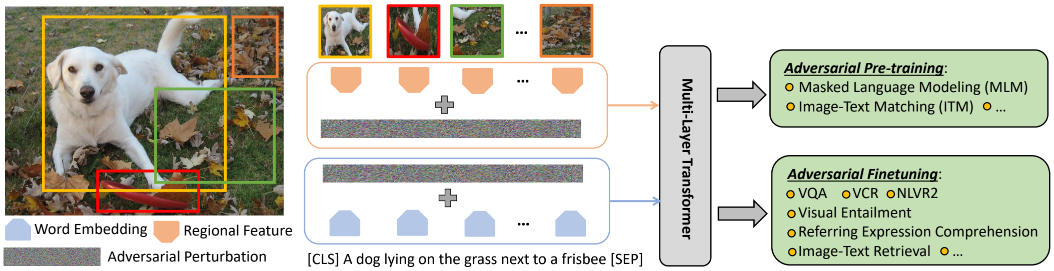

Word Embedding
Regional Feature
Adversarial Perturbation
[CLS] A dog lying on the grass next to a frisbee [SEP]
Multi-Layer Transformer
Adversarial Pre-training:
Masked Language Modeling (MLM)
Image-Text Matching (ITM)
...
Adversarial Finetuning:
VQA
VCR
NLVR2
Visual Entailment
Referring Expression Comprehension
Image-Text Retrieval
...

Figure 1: Overview of the proposed VILLA framework for vision-and-language representation learning.

To bring in more flexibility in generating adversarial examples for robust training, we propose to perform adversarial training on the embedding level for multi-modalities, instead of operating on image pixel and sub-word token level in conventional practice. For text modality, we add adversarial perturbations to word embeddings [41, 76, 22]. For image modality, most previous work observes that robustness is at odds with generalization, *i.e.*, trained models are able to resist adversarial attacks on clean images at the expense of performance [39, 65, 74]. Distinctive from these studies, we directly add adversarial perturbations to extracted image-region features [2], as our end goal is the final V+L model performance rather than crafting adversarial image examples. Experiments show that this strategy leads to large performance gain on clean inputs.

Adversarial training procedure is time-consuming and computationally expensive. To power efficient large-scale training, we adopt the recently proposed "free" adversarial training strategy [50, 73, 76], which obtains the gradients of parameters with almost no extra cost when computing the gradients of inputs. In addition to requiring adversarial perturbations to be label-preserving, we also introduce KL-divergence-based regularization to enforce the confidence level of the prediction to be close, characterized by the "dark" knowledge hidden in the probability vectors. This promotes higher smoothness of the training objective and has empirically proven as important regularization effective for further performance boost.

For evaluation, we mostly focus on UNITER [10], the current best-performing V+L model with state-of-the-art performance across many popular V+L benchmarks, and enhance UNITER with VILLA through comprehensive experiments on six V+L tasks: VQA [15], VCR [72], NLVR^2 [54], Visual Entailment [66], Referring Expression Comprehension [69], and Image-Text Retrieval [27]. VILLA is a generic framework that can be applied to any multimodal pre-training method. To demonstrate its versatility, we further apply it to LXMERT on VQA, GQA [21], and NLVR^2 tasks for generalizability test.

The main contributions are summarized as follows. (i) We present VILLA, the first known effort on adversarial pre-training and adversarial finetuning for V+L representation learning. (ii) Instead of operating on pixel and word token level, we propose to add adversarial perturbations in the embedding space of multi-modalities, and introduce a smoothness-inducing adversarial regularization term on top of the "free" adversarial training strategy. (iii) VILLA achieves new state of the art across six popular V+L tasks. In particular, by relying on standard bottom-up image features only [2], VILLA improves the single-model performance of UNITER-large from 74.02 to 74.87 on VQA, and from 62.8 to 65.7 on VCR. With ensemble, VQA performance is further boosted to 75.85.

2 Related Work

Multimodal Pre-training ViLBERT [35] and LXMERT [58] are the pioneering works in vision+language pre-training, where two Transformers are used to encode image and text modalities, respectively, then a third Transformer is built on top for multimodal fusion. Compared to this two-stream architecture, recent work such as VL-BERT [53], VisualBERT [31], B2T2 [1], Unicoder-VL [28] and UNITER [10] advocate a single-stream model design, where two modalities are directly fused in early stage. More recent studies leverage multi-task learning [36] to enhance finetuning and use detected image tags [32] to further enhance pre-training. Pixel-BERT [19] proposes to align text with image pixels instead of conventional bottom-up features. Multimodal pre-training has brought leaping advances in vision+language understanding tasks such as VQA and VCR, with great potential in extending to visual captioning [75, 63], visual dialog [43, 62], vision-language navigation [16, 40], as well as video-and-language representation learning [56, 55, 38, 29]. Recent work [7] also investigates the design of probing tasks to understand the knowledge learned in pre-training.

V+L Representation Learning Before multimodal pre-training dominated the scene, there had been a long line of studies on how to learn better V+L representations. Prominent work includes: (i) advanced attention mechanisms [37, 70, 47]; (ii) better multimodal fusion methods [12, 71, 25, 24]; (iii) multi-step reasoning [68, 20, 13]; (iv) incorporation of object relations [49, 44, 6, 30]; and (v) neural module networks for compositional reasoning [3, 23, 18, 9]. In principle, our proposed VILLA framework can be plugged into these "shallower" models. In this paper, we mainly focus on enhancing Transformer-based state-of-the-art models.

Adversarial Training Adversarial machine learning is an active research area [57, 14, 5]. Algorithms are developed to either attack existing models by constructing adversarial examples, or train robust models to defend against adversarial attacks. Among existing defense approaches, adversarial training (AT) is a general strategy to empower models with state-of-the-art robustness in different settings [59, 39, 65, 74, 46]. Existing research mostly focuses on AT for image classification, and the general notion is that robustness is often at odds with accuracy. Most recently, [64] shows that model accuracy on clean images can be improved if a separate auxiliary batch norm is used for adversarial examples. There are also some parallel studies on applying AT to language modeling [61] and natural language understanding [41, 76, 22]. Due to growing dominance of large-scale pre-training, very recent work has started to explore adversarial training in the pre-training stage [17, 8, 34]. VILLA is the first known effort that studies AT for V+L tasks and adds adversarial perturbations to both image and word embedding space. We also prove that AT can be effectively incorporated in both pre-training and fine-tuning stages. A more detailed discussion on related work is provided in Appendix.

3 Vision-and-Language Large-scale Adversarial Training

There are three key designs that encapsulate VILLA's unique strengths in improving performance and generalization of pre-trained V+L models : (i) Adversarial pre-training *and* fine-tuning; (ii) Adding perturbations in the embedding space; and (iii) Enhanced adversarial training algorithm.

3.1 Adversarial Pre-training and Finetuning

We first briefly review the *pretrain-then-finetune* paradigm that has become prevalent in V+L representation learning, then describe our proposed two-stage adversarial training framework.

Pre-training Let $\mathcal{D}_p$ denote a pre-training dataset, which consists of image-text pairs $(\boldsymbol{x}_{img}, \boldsymbol{x}_{txt})$. The goal in the pre-training stage is to learn universal image-text representations that are generalizable to different downstream tasks. Take one-stream models [10, 53] as an example. Image and text inputs are first represented as low-dimensional feature vectors $\boldsymbol{z}_{img} = g_{bu}(\boldsymbol{x}_{img})$ and $\boldsymbol{z}_{txt} = g_{emb}(\boldsymbol{x}_{txt})$, where $g_{bu}(\cdot)$ represents a fixed bottom-up image feature extractor [2], and $g_{emb}(\cdot)$ represents a learnable word embedding function. Then, a multi-layer Transformer [60] is applied on top to learn multimodal fusion. The above process can be abbreviated as $\tilde{\boldsymbol{z}}_{img}, \tilde{\boldsymbol{z}}_{txt}, \tilde{\boldsymbol{z}}_{cls} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_{img}, \boldsymbol{x}_{txt})$, where $\tilde{\boldsymbol{z}}_{img}$ and $\tilde{\boldsymbol{z}}_{txt}$ represent the contextualized representations of each image region and each textual token, respectively. Typically, V+L models employ a special `[CLS]` token whose embedding $\tilde{\boldsymbol{z}}_{cls}$ is considered as the joint V+L representation to be used for downstream tasks. $\boldsymbol{\theta}$ denotes all the learnable parameters including the word embedding matrix.

Let $\boldsymbol{y}$ denote the output supervision signal, which is different across different pre-training tasks. There are three typical pre-training tasks used in most V+L models: (i) Masked Language Modeling (MLM): some tokens in $\boldsymbol{x}_{txt}$ are replaced by special `[MASK]` tokens, and the goal is to predict the masked tokens $\boldsymbol{y}$ based on surrounding multimodal context; (ii) Masked Region Modeling (MRM): the features of some image regions in $\boldsymbol{x}_{img}$ are replaced by zero vectors, and the goal is to predict the masked image regions $\boldsymbol{y}$ given the remaining multimodal information (via cross-entropy loss, KL-divergence loss [35], or contrastive learning [55]); (iii) Image-Text Matching (ITM): both $\boldsymbol{x}_{img}$ and $\boldsymbol{x}_{txt}$ are kept intact, and the goal is to predict a binary label $\boldsymbol{y}$ to judge whether the input image and text are paired or not.

Finetuning Given a downstream task $\mathcal{T}_f$ and a supervised dataset $\mathcal{D}_f$ consisting of $(\boldsymbol{x}_{img}, \boldsymbol{x}_{txt}, \boldsymbol{y})$, the pre-trained model can be finetuned by introducing a small neural network $h(\cdot)$ on top of $\tilde{\boldsymbol{z}}_{cls}$ and minimizing the cross-entropy loss. $\boldsymbol{\theta}$ is initialized with pre-trained weights, and $\boldsymbol{y}$ now becomes a label. For example, in VQA, $\boldsymbol{y}$ corresponds to the ground-truth answer from a candidate pool, represented as a one-hot vector. In VCR [72], it is a four-way classification label.

In both pre-training and finetuning, by instantiating different $\boldsymbol{y}$, the training process can be uniformly abstracted as an empirical risk minimization problem:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x}_{img}, \boldsymbol{x}_{txt}, \boldsymbol{y}) \sim \mathcal{D}} [L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{img}, \boldsymbol{x}_{txt}), \boldsymbol{y})] \,. \tag{1}$$

Two-stage Adversarial Training Pre-training and finetuning are inherently connected. Independent of the tasks (*e.g.*, MLM, ITM for pre-training, or VQA for finetuning), model training requires the acquisition of essential reasoning skills that can catalyze multimodal fusion for cross-modality joint understanding. For example, in MLM, a masked token 'dog' can be predicted by looking at the image region that contains a dog; and in VQA, when asked whether there is a dog in an image, such visual grounding skills learned through pre-training can be readily applied. We hypothesize that: (i) by performing adversarial training in the pre-training stage, the improved generalization ability of a learned model can be beneficial to the finetuning stage; and (ii) in the subsequent finetuning stage, where task-specific training signals become available, adversarial finetuning can be applied again to further boost performance. Since pre-training and finetuning share the same mathematical formulation (Eqn. (1)), the same AT algorithm can be adopted in both stages.

3.2 Perturbations in the Embedding Space

For the image modality, since state-of-the-art V+L models typically use image features from pre-trained object detectors as input, we add adversarial perturbations in the feature space directly. Note that even though the main difference is simply the noise injecting space, our approach is distinctive from most previous work where perturbations are applied to the pixel space, which is more rigid than fine-grained embedding perturbation. On the other hand, unlike image pixels that are continuous-valued, discrete tokens in the text modality are more difficult to manipulate. It remains unclear how to craft label-preserving adversarial examples without changing the original semantic meaning of the sentence. But since we only care about the *ultimate effects* of adversarial training on downstream tasks, not intepretability of adversarial examples, we choose to add perturbations to the word embeddings following [76].

In pre-trained V+L models, positional embeddings are used to encode the location of image regions and sub-word tokens. Our adversaries only modify image and word embeddings, leaving other components of the multimodal features unchanged. Furthermore, due to the distinct characteristics of image and text modalities, we propose to add perturbations to one modality at a time. Specifically, we add adversarial perturbations $\boldsymbol{\delta}_{img}$ and $\boldsymbol{\delta}_{txt}$ such that the prediction becomes $\hat{\boldsymbol{y}} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_{img} + \boldsymbol{\delta}_{img}, \boldsymbol{x}_{txt})$ and $\tilde{\boldsymbol{y}} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_{img}, \boldsymbol{x}_{txt} + \boldsymbol{\delta}_{txt})$. To preserve original semantics, the norm of $\boldsymbol{\delta}_{img}$ and $\boldsymbol{\delta}_{txt}$ is controlled to be small. Also assumed is that model prediction should not change after the perturbation.

3.3 "Free" Multimodal Adversarial Training

Training Objective In VILLA, we use adversarial training as an effective regularization to improve model generalization, *i.e.,* to minimize the following objective:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x}_{img}, \boldsymbol{x}_{txt}, \boldsymbol{y}) \sim \mathcal{D}} \Big[\mathcal{L}_{std}(\boldsymbol{\theta}) + \mathcal{R}_{at}(\boldsymbol{\theta}) + \alpha \cdot \mathcal{R}_{kl}(\boldsymbol{\theta}) \Big] \,, \tag{2}$$

where $\mathcal{L}_{std}(\boldsymbol{\theta}) = L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{img}, \boldsymbol{x}_{txt}), \boldsymbol{y})$ is the cross-entropy loss on clean data, $\mathcal{R}_{at}(\boldsymbol{\theta})$ is the label-preserving AT loss, and $\mathcal{R}_{kl}(\boldsymbol{\theta})$ is a finer-grained adversarial regularization term. Specifically,

$$\mathcal{R}_{at}(\boldsymbol{\theta}) = \max_{||\boldsymbol{\delta}_{img}|| \le \epsilon} L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{img} + \boldsymbol{\delta}_{img}, \boldsymbol{x}_{txt}), \boldsymbol{y}) + \max_{||\boldsymbol{\delta}_{txt}|| \le \epsilon} L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{img}, \boldsymbol{x}_{txt} + \boldsymbol{\delta}_{txt}), \boldsymbol{y}) \,, \tag{3}$$

where L is the cross-entropy loss on adversarial embeddings. Frobenius norm is used to constrain $\boldsymbol{\delta}_{img}$ and $\boldsymbol{\delta}_{txt}$. For optimization, [39] demonstrated that the outer minimization in Eqn. (2) can be solved by SGD, while the inner maximization in Eqn. (3) can be solved reliably by PGD, a standard method for large-scale constrained optimization. Take $\boldsymbol{\delta}_{img}$ for example: PGD takes the following step (with step-size α) in each iteration:

$$\boldsymbol{\delta}_{img,t+1} = \Pi_{||\boldsymbol{\delta}_{img}|| \le \epsilon} (\boldsymbol{\delta}_{img,t} + \alpha g(\boldsymbol{\delta}_{img,t}) / ||g(\boldsymbol{\delta}_{img,t})||_F) \,, \tag{4}$$

Algorithm 1 "Free" Multi-modal Adversarial Training used in VILLA.

```
Require: Training samples $\mathcal{D} = \{(\boldsymbol{x}_{img}, \boldsymbol{x}_{txt}, \boldsymbol{y})\}$, perturbation bound $\epsilon$, learning rate $\tau$, ascent steps $K$, ascent step size $\alpha$
Initialize $\boldsymbol{\theta}$
for epoch $= 1 \dots N_{ep}$ do
    for minibatch $B \subset X$ do
        $\boldsymbol{\delta}_0 \leftarrow \frac{1}{\sqrt{N_{\boldsymbol{\delta}}}} U(-\epsilon, \epsilon),\ \boldsymbol{g}_0 \leftarrow 0$
        for $t = 1 \dots K$ do
            Accumulate gradient of parameters $\boldsymbol{\theta}$ given $\boldsymbol{\delta}_{img,t-1}$ and $\boldsymbol{\delta}_{txt,t-1}$
                $\boldsymbol{g}_t \leftarrow \boldsymbol{g}_{t-1} + \frac{1}{K}\mathbb{E}_{(\boldsymbol{x}_{img}, \boldsymbol{x}_{txt}, \boldsymbol{y}) \in B}[\nabla_{\boldsymbol{\theta}}(\mathcal{L}_{std}(\boldsymbol{\theta}) + \mathcal{R}_{at}(\boldsymbol{\theta}) + \mathcal{R}_{kl}(\boldsymbol{\theta}))]$
            Update the perturbation $\boldsymbol{\delta}_{img}$ and $\boldsymbol{\delta}_{txt}$ via gradient ascend
                $\tilde{\boldsymbol{y}} = f_{\boldsymbol{\theta}}(\boldsymbol{x}_{img}, \boldsymbol{x}_{txt})$
                $\boldsymbol{g}_{img} \leftarrow \nabla_{\boldsymbol{\delta}_{img}}[L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{img}+\boldsymbol{\delta}_{img}, \boldsymbol{x}_{txt}), \boldsymbol{y}) + L_{kl}(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{img}+\boldsymbol{\delta}_{img}, \boldsymbol{x}_{txt}), \tilde{\boldsymbol{y}})]$
                $\boldsymbol{\delta}_{img,t} \leftarrow \Pi_{\|\boldsymbol{\delta}_{img}\|_F \le \epsilon}(\boldsymbol{\delta}_{img,t-1} + \alpha \cdot \boldsymbol{g}_{img}/\|\boldsymbol{g}_{img}\|_F)$
                $\boldsymbol{g}_{txt} \leftarrow \nabla_{\boldsymbol{\delta}_{txt}}[L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{img}, \boldsymbol{x}_{txt}+\boldsymbol{\delta}_{txt}), \boldsymbol{y}) + L_{kl}(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{img}, \boldsymbol{x}_{txt}+\boldsymbol{\delta}_{txt}), \tilde{\boldsymbol{y}})]$
                $\boldsymbol{\delta}_{txt,t} \leftarrow \Pi_{\|\boldsymbol{\delta}_{txt}\|_F \le \epsilon}(\boldsymbol{\delta}_{txt,t-1} + \alpha \cdot \boldsymbol{g}_{txt}/\|\boldsymbol{g}_{txt}\|_F)$
        end for
        $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \tau \boldsymbol{g}_K$
    end for
end for
```

where $g(\boldsymbol{\delta}_{img,t}) = \nabla_{\boldsymbol{\delta}_{img}} L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{img} + \boldsymbol{\delta}_{img}, \boldsymbol{x}_{txt}), \boldsymbol{y})$ is the gradient of the loss w.r.t. $\boldsymbol{\delta}_{img}$, and $\Pi_{||\boldsymbol{\delta}_{img}|| \le \epsilon}$ performs a projection onto the ϵ-ball.

To further enhance the above AT algorithm, $\mathcal{R}_{kl}(\boldsymbol{\theta})$ is defined as

$$\begin{aligned}\mathcal{R}_{kl}(\boldsymbol{\theta}) = &\max_{||\boldsymbol{\delta}_{img}|| \le \epsilon} L_{kl}(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{img} + \boldsymbol{\delta}_{img}, \boldsymbol{x}_{txt}), f_{\boldsymbol{\theta}}(\boldsymbol{x}_{img}, \boldsymbol{x}_{txt})) \\ &+ \max_{||\boldsymbol{\delta}_{txt}|| \le \epsilon} L_{kl}(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{img}, \boldsymbol{x}_{txt} + \boldsymbol{\delta}_{txt}), f_{\boldsymbol{\theta}}(\boldsymbol{x}_{img}, \boldsymbol{x}_{txt}))\,, \end{aligned} \tag{5}$$

where $L_{kl}(p, q) = \mathrm{KL}(p||q) + \mathrm{KL}(q||p)$, p, q denote the two probability distributions, and $\mathrm{KL}(\cdot)$ denotes the Kullback-Leibler Divergence. Compared to Eqn. (3) that promotes label-preserving adversarial attack, Eqn. (5) further advocates that the confidence level of the prediction, characterized by the probability vector over the simplex Δ_n (n is the number of classes), should also be close. Similar techniques are used in Virtual AT [42], TRADES [74], and UDA [67]. However, previous work mostly focuses on semi-supervised learning or trade-off between accuracy and robustness; in our work, we found that it is highly effective for boosting model generalization ability.

"Free" AT Strategy K-step PGD requires K forward-backward passes through the network, which is computationally heavy. Another limitation is that after K steps, only perturbations at the final step are used for model training. To enable AT for large-scale training and promote diverse adversaries, we follow FreeLB [76] to perform multiple PGD iterations to craft adversarial embeddings, and simultaneously accumulate the "free" parameter gradients $\nabla_{\boldsymbol{\theta}} L$ in each iteration. After that, the model parameters $\boldsymbol{\theta}$ are updated all at once with the accumulated gradients, effectively creating a K-times-larger "virtual" mini-batch. The full procedure is provided in Algorithm 1.

4 Experiments

4.1 Experimental Setting

Downstream Tasks To validate the effectiveness of VILLA, we apply it to existing V+L pre-trained models and conduct a comprehensive evaluation over a wide range of downstream tasks, including Visual Question Answering (VQA), Visual Commonsense Reasoning (VCR), Referring Expression (RE) Compression, Visual Entailment, Image-Text Retrieval, and NLVR^2. To validate the strength of VILLA in model pre-training and finetuning, we first incorporate it into state-of-the-art UNITER model in both stages for downstream evaluation and ablation analysis. And to demonstrate the versatility of VILLA, we further apply it to another V+L model LXMERT [58] with a different architecture design from UNITER (two-stream vs. one-stream) for generalizability test.

Method	VQA		VCR			NLVR^2		SNLI-VE	
	test-dev	test-std	Q→A	QA→R	Q→AR	dev	test-P	val	test
ViLBERT	70.55	70.92	72.42 (73.3)	74.47 (74.6)	54.04 (54.8)	-	-	-	-
VisualBERT	70.80	71.00	70.8 (71.6)	73.2 (73.2)	52.2 (52.4)	67.4	67.0	-	-
LXMERT	72.42	72.54	-	-	-	74.90	74.50	-	-
Unicoder-VL	-	-	72.6 (73.4)	74.5 (74.4)	54.4 (54.9)	-	-	-	-
12-in-1	73.15	-	-	-	-	-	78.87	-	76.95
VL-BERT$_{\text{BASE}}$	71.16	-	73.8 (-)	74.4 (-)	55.2 (-)	-	-	-	-
Oscar$_{\text{BASE}}$	73.16	73.44	-	-	-	78.07	78.36	-	-
UNITER$_{\text{BASE}}$	72.70	72.91	74.56 (75.0)	77.03 (77.2)	57.76 (58.2)	77.18	77.85	78.59	78.28
VILLA$_{\text{BASE}}$	**73.59**	**73.67**	**75.54 (76.4)**	**78.78 (79.1)**	**59.75 (60.6)**	**78.39**	**79.30**	**79.47**	**79.03**
VL-BERT$_{\text{LARGE}}$	71.79	72.22	75.5 (75.8)	77.9 (78.4)	58.9 (59.7)	-	-	-	-
Oscar$_{\text{LARGE}}$	73.61	73.82	-	-	-	79.12	80.37	-	-
UNITER$_{\text{LARGE}}$	73.82	74.02	77.22 (77.3)	80.49 (80.8)	62.59 (62.8)	79.12	79.98	79.39	79.38
VILLA$_{\text{LARGE}}$	**74.69**	**74.87**	**78.45 (78.9)**	**82.57 (82.8)**	**65.18 (65.7)**	**79.76**	**81.47**	**80.18**	**80.02**

(a) Results on VQA, VCR, NLVR^2, and SNLI-VE.

Method	RefCOCO+						RefCOCO					
	val	testA	testB	val^d	testA^d	testB^d	val	testA	testB	val^d	testA^d	testB^d
ViLBERT	-	-	-	72.34	78.52	62.61	-	-	-	-	-	-
VL-BERT$_{\text{BASE}}$	79.88	82.40	75.01	71.60	77.72	60.99	-	-	-	-	-	-
UNITER$_{\text{BASE}}$	83.66	86.19	78.89	75.31	81.30	65.58	91.64	92.26	90.46	81.24	86.48	73.94
VILLA$_{\text{BASE}}$	**84.26**	**86.95**	**79.22**	**76.05**	**81.65**	**65.70**	**91.93**	**92.79**	**91.38**	**81.65**	**87.40**	**74.48**
VL-BERT$_{\text{LARGE}}$	80.31	83.62	75.45	72.59	78.57	62.30	-	-	-	-	-	-
UNITER$_{\text{LARGE}}$	84.25	**86.34**	79.75	75.90	81.45	66.70	91.84	92.65	91.19	81.41	87.04	74.17
VILLA$_{\text{LARGE}}$	**84.40**	86.22	**80.00**	**76.17**	**81.54**	**66.84**	**92.58**	**92.96**	**91.62**	**82.39**	**87.48**	**74.84**

(b) Results on RefCOCO+ and RefCOCO. The superscript d denotes evaluation using detected proposals.

Method	RefCOCOg				Flickr30k IR			Flickr30k TR		
	val	test	val^d	test^d	R@1	R@5	R@10	R@1	R@5	R@10
ViLBERT	-	-	-	-	58.20	84.90	91.52	-	-	-
Unicoder-VL	-	-	-	-	71.50	90.90	94.90	86.20	96.30	99.00
UNITER$_{\text{BASE}}$	86.52	86.52	74.31	74.51	72.52	92.36	**96.08**	85.90	97.10	98.80
VILLA$_{\text{BASE}}$	**88.13**	**88.03**	**75.90**	**75.93**	**74.74**	**92.86**	95.82	**86.60**	**97.90**	**99.20**
UNITER$_{\text{LARGE}}$	87.85	87.73	74.86	75.77	75.56	94.08	96.76	87.30	**98.00**	**99.20**
VILLA$_{\text{LARGE}}$	**88.42**	**88.97**	**76.18**	**76.71**	**76.26**	**94.24**	**96.84**	**87.90**	97.50	98.80

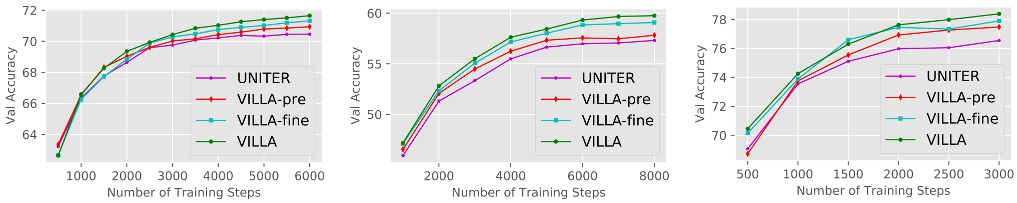
(c) Results on RefCOCOg and Flickr30k Image Retrieval (IR) and Text Retrieval (TR).

Table 1: Comparison with state-of-the-art pre-trained models on all the downstream tasks.

UNITER and LXMERT UNITER-base is a single-stream model, which has 12 layers, with 768 hidden units per layer and 12 attention heads; UNITER-large has 24 layers, with 1024 hidden units per layer and 16 attention heads. UNITER shares the same structure as BERT, except that the input now becomes a mixed sequence of two modalities. LXMERT is a two-stream model, which first performs self-attention through several layers on each modality independently (9 layers for text modality, and 5 layers for image modality), then fuses the outputs of both streams through another 5 layers (first cross-attention, then self-attention).

Implementation Details For UNITER experiments, we pre-train with the same four large-scale datasets used in the original model: COCO [33], Visual Genome (VG) [26], Conceptual Captions [52] and SBU Captions [45]. VILLA is applied to both MLM and ITM pre-training tasks. The original UNITER-base (12 layers) and UNITER-large (24 layers) models take 200k and 500k steps for pre-training, respectively. For fair comparison, when applying VILLA to UNITER-base, we run 100k steps of standard training, followed by 100k steps of adversarial training. When applying VILLA to UNITER-large, to save pre-training time,[2] we run 425k steps of standard training, followed by 75k steps of adversarial training.

[2] VILLA is K times computationally heavier than UNITER, where K is the number of adversarial training steps. We typically select adversarial learning rate from {1e-2, 1e-3}, adversarial training steps to 3, and α (Eqn. 2) from $1.0, 1.5, 2.0$. More implementation details are provided in Appendix.



(a) VQA (b) VCR (c) NLVR^2

Figure 2: The training curves of VILLA and UNITER on different tasks. For VQA, an internal val set is used.

Method	VQA	VCR (val)			NLVR^2	VE	Flickr30k IR			RefCOCO		Ave.
	test-dev	Q→A	QA→R	Q→AR	test-P	test	R@1	R@5	R@10	testA^d	testB^d	
UNITER (reimp.)	72.70	74.24	76.93	57.31	77.85	78.28	72.52	92.36	96.08	86.48	73.94	78.06
VILLA-pre	73.03	74.76	77.04	57.82	78.44	78.43	73.76	93.02	96.28	87.34	74.35	78.57
VILLA-fine	73.29	75.18	78.29	59.08	78.84	78.86	73.46	92.98	96.26	87.17	74.31	78.88
VILLA	73.59	75.54	78.78	59.75	79.30	79.03	74.74	92.86	95.82	87.40	74.48	**79.21**

Table 2: Ablation study on VILLA-pre (pre-training) and VILLA-fine (finetuning) with base model size.

Method	VQA	VCR (val)		
	test-dev	Q→A	QA→R	Q→AR
VILLA$_{\text{BASE}}$ (txt)	73.50	75.60	78.70	59.67
VILLA$_{\text{BASE}}$ (img)	73.50	**75.81**	78.43	59.68
VILLA$_{\text{BASE}}$ (both)	**73.59**	75.54	**78.78**	**59.75**
VILLA$_{\text{LARGE}}$ (txt)	74.55	78.08	82.31	64.63
VILLA$_{\text{LARGE}}$ (img)	74.46	78.08	82.28	64.51
VILLA$_{\text{LARGE}}$ (both)	**74.69**	**78.45**	**82.57**	**65.18**

(a) Image vs. Text Modality.

Method	VQA	VCR (val)		
	test-dev	Q→A	QA→R	Q→AR
UNITER$_{\text{BASE}}$ (reimp.)	72.70	74.24	76.93	57.31
UNITER$_{\text{BASE}}$+FreeLB	72.82	75.13	77.90	58.73
VILLA$_{\text{BASE}}$-fine	**73.29**	**75.49**	**78.34**	**59.30**
UNITER$_{\text{LARGE}}$ (reimp.)	73.82	76.70	80.61	62.15
UNITER$_{\text{LARGE}}$+FreeLB	73.87	77.19	81.44	63.24
VILLA$_{\text{LARGE}}$-fine	**74.32**	**77.75**	**82.10**	**63.99**

(b) FreeLB vs. VILLA.

Table 3: Ablation study on adding perturbations to different modalities and on the VILLA algorithm.

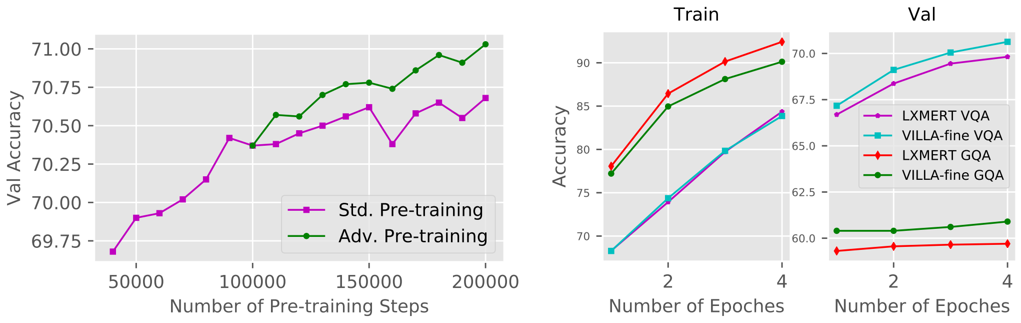
4.2 Results and Ablation Analysis

Downstream Task Evaluation Table 1 summarizes the results of VILLA applied to UNITER on all evaluation tasks. Compared with existing pre-trained V+L models, our VILLA method achieves new state of the art across all the benchmarks. Specifically, VILLA-base model outperforms UNITER-base by +0.76 on *VQA*, +2.4 on *VCR* for Q→AR, +1.45 on *NLVR*^{2}, +0.75 on *SNLI-VE*, +2.22/+0.70 on *Flickr30k* for Image/Text Retrieval (R@1), and +0.99 on average for the three *RE* datasets.

Similar universal performance lift is also observed in VILLA-large. It is highly encouraging to see that VILLA-large brings an absolute +2.9 points performance gain over UNITER-large for VCR on the Q→AR metric. Compared to the others, VCR is a relatively more challenging task, which requires commonsense reasoning and understanding complex social dynamics that is implicitly encoded in the image. Another significant boost is over the well-studied VQA benchmark, from 74.02 to 74.87. With ensemble, the performance of VILLA-large is further lifted to 75.85.

Pre-training vs. Finetuning To understand the effects of adversarial training on pre-training and finetuning, we conduct an ablation study with UNITER-base and summarize the results in Table 2. UNITER (reimp.) denotes our re-implementation of the UNITER-base model with standard training. VILLA-pre and VILLA-fine apply adversarial training to only the pre-training or finetuning stage, respectively. Averaged over the six evaluation tasks, VILLA-pre and VILLA-fine brings +0.51 and +0.82 points performance gain. By combining the two, +1.15 points gain is achieved. Figure 2 further provides the training curves of each task, which illustrate growing performance gaps between AT-enhanced models and the original UNITER, as the number of training steps increases. Interestingly, on VQA, though in early epochs UNITER achieves better performance than VILLA, VILLA catches up quickly after a few hundred of steps, which demonstrates the beneficial regularization effect of adversarial training. More training curves on other tasks can be found in Appendix.

To further understand the importance of adversarial pre-training, we use VQA as the probing task, and compare the performance of standard and adversarial pre-training at each intermediate model



(a) Standard vs. adversarial pre-training.

(b) Adversarial training as a regularizer.

Figure 3: For (a), we use VQA as probing, and compare the performance of standard and adversarial pre-training. For (b), we plot the training curves of standard and adversarial finetuning using LXMERT as backbone.

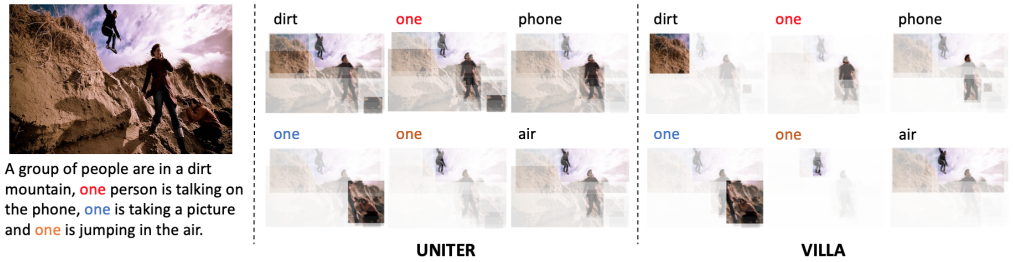

A group of people are in a dirt mountain, one person is talking on the phone, one is taking a picture and one is jumping in the air.
dirt
one
phone
one
one
air
UNITER
dirt
one
phone
one
one
air
VILLA

Figure 4: Visualization of text-to-image attention, comparing VILLA against UNITER.

Model	Visual Coreference (Flickr30k)					Visual Relation (Visual Genome)					Ave.
	scene	clothing	animals	instruments	vehicles	on	standing in	wearing	holding	covering	
UNITER$_{\text{BASE}}$	0.151	0.157	0.285	0.244	0.194	0.154	0.107	0.311	0.200	0.151	0.195
VILLA$_{\text{BASE}}$	**0.169**	**0.185**	**0.299**	**0.263**	**0.202**	**0.201**	**0.120**	**0.353**	**0.241**	**0.192**	**0.223**

Table 4: Probing analysis of the attention heads in pre-trained UNITER and VILLA models.

checkpoint (using standard finetuning to both pre-trained models). Results are presented in Figure 3a. As shown, once adversarial training is activated, VILLA-pre starts outperforming UNITER, and the performance gap increases as the number of pre-training steps grows.

Image vs. Text Modality To gain insights on the effects of adversarial examples in different modalities, we conduct experiments by adding perturbations to either image or text modality, and use VQA and VCR for ablation tests. Results are summarized in Table 3a. Conventionally, adversarial training in the image domain hurts model accuracy on clean images. However, in our setting, we observe that adding perturbations to image features alone can boost final model performance significantly. Our initial intuition was that adding perturbations to both modalities might increase the diversity of adversarial examples, hence bringing more benefits. However, ablation results show that adding perturbations on one modality is already gaining significant improvement.[3] The boost on VCR is larger than VQA, which we hypothesize is due to the higher complexity in VCR task, which adding more adversaries to model training can effectively help.

FreeLB vs. VILLA To compare with prior work FreeLB, we conduct an additional ablation study also on VQA and VCR, two representative and challenging V+L tasks. Table 3b shows that VILLA achieves consistently better performance than FreeLB over both benchmarks, thanks to the additional fine-grained adversarial regularization term. For example, FreeLB brings little performance boost on VQA, while VILLA achieves considerable improvement over the baseline.

Probing Analysis Pre-trained models are expected to learn intricate knowledge about multimodality correlations, such as visual coreference (*i.e.*, region-phrase alignment) and visual relation (*i.e.*, region-

[3] We also tried adding adversarial perturbations to both modalities simultaneously instead of alternatively. Empirically, we observe that they obtained similar performance.

Method	VQA		GQA		NLVR^2		Meta-Ave.
	test-dev	test-std	test-dev	test-std	dev	test-P	
LXMERT	72.42	72.54	60.00	60.33	74.95	74.45	69.12
LXMERT (reimp.)	72.50	72.52	59.92	60.28	74.72	74.75	69.12
VILLA-fine	**73.02**	**73.18**	**60.98**	**61.12**	**75.98**	**75.73**	**70.00**

Table 5: Results on LXMERT with VILLA-fine (finetuning).

Data split	MUTAN	BUTD	BUTD+CC	Pythia	Pythia+CC	BAN	BAN+CC	UNITER	VILLA
Original	59.08	61.51	62.44	64.08	64.52	64.97	65.87	70.35	**71.27**
Rephrasing	46.87	51.22	52.58	54.20	55.65	55.87	56.59	64.56	**65.35**

Table 6: Results on VQA-Rephrasings. Both UNITER and VILLA use the base model size. Baseline results are copied from [51].

region interaction). To provide a more direct measurement on how well our adversarial pre-trained model captures such multimodal signals, we conduct a probing analysis following [7]. We consider five most common visual coreference types in Flickr30k Entities [48] and top five visual relations in Visual Genome [26] (listed in Table 4), and calculate the attention weights between region and phrase (or between regions) learned by pre-trained models. Results show that VILLA presents higher attention weights across all the ten categories (0.223 vs. 0.195 on average), indicating a higher probability of identifying those relations. Figure 4 further provides a visualization of text-to-image attention, where VILLA exhibits more accurate and sharper multimodal alignment.

Results on LXMERT VILLA is a generic framework that can be readily applied to any V+L models. To demonstrate its generalization ability, we conduct additional experiments using LXMERT as the backbone. Since adversarial pre-training is highly time-consuming, we only focus on adversarial finetuning for LXMERT.[4] We use VQA, GQA and NLVR^2 as the evaluation tasks, the same as LXMERT. Results in Table 5 show that VILLA-fine instantly provides +0.88 average performance boost across the three tasks. The training curves are provided in Figure 3b. Compared to LXMERT, VILLA-fine achieves higher accuracy on validation set and lower accuracy on training set for both VQA and GQA, clearly demonstrating its regularization effect in preventing overfitting of large-scale pre-trained models.

Robustness In order to test adversarial robustness, we need to perform adversarial attacks to existing V+L models. This V+L attack problem is largely unexplored in the literature. For example, how to reliably back-propagate the gradients from the multimodal Transformer to the CNN backbone to generate image adversaries is non-trivial. How to craft textual adversaries that align with the visual context is also challenging. In this work, we mainly focus on improving model's generalization performance on clean data, leaving a more thorough investigation of adversarial attack and robustness as important future work.

As a proxy for robustness evaluation, we conduct additional experiments on the VQA-Rephrasings dataset [51] to test the robustness of existing V+L models to paraphrases. For fair comparison, we have re-trained both UNITER and VILLA on the VQA training set only. Results are summarized in Table 6, where 'Original' and 'Rephrasing' denote the test set with original questions and their rephrasings, respectively. UNITER has already lifted up the performance by a large margin, and VILLA facilitates further performance boost.

We provide additional experimental results, more details about the probing analysis, and additional visualization examples in Appendix.

5 Conclusion

In this paper, we present VILLA, an advanced adversarial training (AT) framework for better vision-and-language representation learning. By performing AT in both pre-training and finetuning stages, and by adding adversarial perturbations to the embedding space, VILLA achieves consistent performance boost on all the benchmarks evaluated. As AT is time-consuming, for future work, we plan to study how to accelerate AT so that it can be more feasible for large-scale pre-training in practice.

[4] Code is available at `https://github.com/zhegan27/LXMERT-AdvTrain`.

Broader Impact

Our research advances vision-and-language representation learning by incorporating adversarial training in both pre-training and finetuning stages. By utilizing the enormous amount of image-text data available on the web for pre-training, VILLA can absorb multimodal clues to capture multi-channel signals from the world, towards a smarter AI system. Furthermore, VILLA can provide instant performance boost in finetuning stage, which will help accelerate future studies in this field. However, in order to train models to learn such capabilities, our method also calls for a high demand on computational resources due to large-scale training, which could be costly both financially and environmentally. As part of our research effort, we will release our pre-trained models to facilitate future research, to empower others' scientific exploration and save environmental cost.

Acknowledgments and Disclosure of Funding

We appreciate anonymous reviewers for their constructive feedback. This work was fully supported by Microsoft, and was not funded by any other agencies.

References

[1] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054*, 2019.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[3] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *CVPR*, 2016.

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

[5] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

[6] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *CVPR*, 2019.

[7] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. *arXiv preprint arXiv:2005.07310*, 2020.

[8] Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. *arXiv preprint arXiv:2003.12862*, 2020.

[9] Wenhu Chen, Zhe Gan, Linjie Li, Yu Cheng, William Wang, and Jingjing Liu. Meta module network for compositional visual reasoning. *arXiv preprint arXiv:1910.03230*, 2019.

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[12] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[13] Zhe Gan, Yu Cheng, Ahmed EI Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. Multi-step reasoning via recurrent dual attention for visual dialog. *arXiv preprint arXiv:1902.00579*, 2019.

[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

[16] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. *arXiv preprint arXiv:2002.10638*, 2020.

[17] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. *arXiv preprint arXiv:1901.09960*, 2019.

[18] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *ICCV*, 2017.

[19] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

[20] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018.

[21] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

[22] Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437*, 2019.

[23] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *ICCV*, 2017.

[24] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NeurIPS*, 2018.

[25] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016.

[26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

[27] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, 2018.

[28] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *arXiv preprint arXiv:1908.06066*, 2019.

[29] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020.

[30] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. *arXiv preprint arXiv:1903.12314*, 2019.

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

[32] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. *arXiv preprint arXiv:2004.06165*, 2020.

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

[34] Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*, 2020.

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

[36] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. *arXiv preprint arXiv:1912.02315*, 2019.

[37] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*, 2016.

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

[39] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[40] Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. Improving vision-and-language navigation with image-text pairs from the web. *arXiv preprint arXiv:2004.14973*, 2020.

[41] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*, 2016.

[42] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *PAMI*, 2018.

[43] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. *arXiv preprint arXiv:1912.02379*, 2019.

[44] Will Norcliffe-Brown, Stathis Vafeias, and Sarah Parisot. Learning conditioned graph structures for interpretable visual question answering. In *NeurIPS*, 2018.

[45] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *NeurIPS*, 2011.

[46] Tianyu Pang, Xiao Yang, Yinpeng Dong, Kun Xu, Hang Su, and Jun Zhu. Boosting adversarial training with hypersphere embedding. *arXiv preprint arXiv:2002.08619*, 2020.

[47] Gao Peng, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. *arXiv preprint arXiv:1812.05252*, 2018.

[48] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015.

[49] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *NeurIPS*, 2017.

[50] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *NeurIPS*, 2019.

[51] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *CVPR*, 2019.

[52] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

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

[54] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.

[55] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*, 2019.

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

[57] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

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

[59] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[61] Dilin Wang, Chengyue Gong, and Qiang Liu. Improving neural language modeling via adversarial training. *arXiv preprint arXiv:1906.03805*, 2019.

[62] Yue Wang, Shafiq Joty, Michael R Lyu, Irwin King, Caiming Xiong, and Steven CH Hoi. Vd-bert: A unified vision and dialog transformer with bert. *arXiv preprint arXiv:2004.13278*, 2020.

[63] Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, and Ming Zhou. Xgpt: Cross-modal generative pre-training for image captioning. *arXiv preprint arXiv:2003.01473*, 2020.

[64] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan Yuille, and Quoc V Le. Adversarial examples improve image recognition. *arXiv preprint arXiv:1911.09665*, 2019.

[65] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *CVPR*, 2019.

[66] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

[67] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. 2019.

[68] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.

[69] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016.

[70] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019.

[71] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *ICCV*, 2017.

[72] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019.

[73] Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. In *NeurIPS*, 2019.

[74] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019.

[75] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*, 2019.

[76] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Thomas Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for language understanding. *arXiv preprint arXiv:1909.11764*, 2019.