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

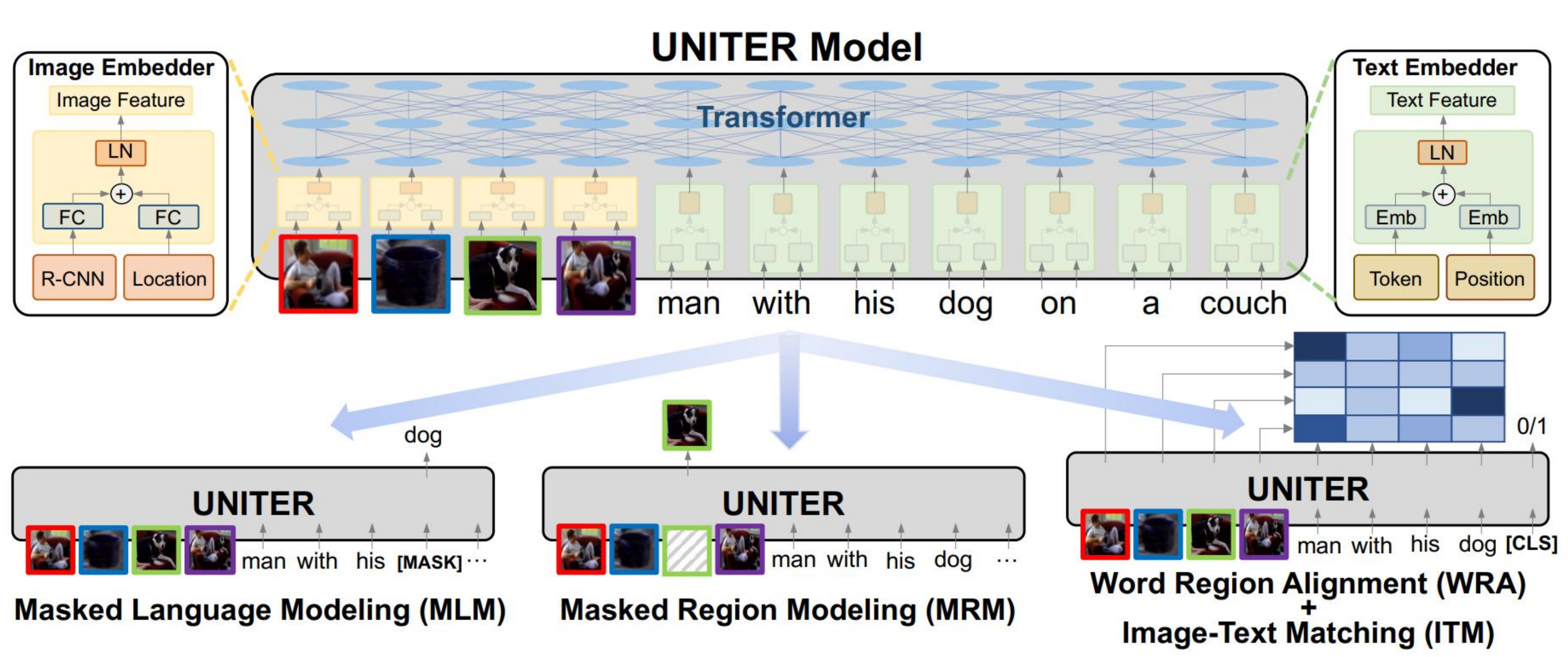
Zhe Gan[1], Yen-Chun Chen[1], Linjie Li[1], Chen Zhu[2], Yu Cheng[1], Jingjing Liu[1]

[1]Microsoft Dynamics 365 AI Research, [2]University of Maryland, College Park

Motivation & Contribution

- ***Multimodal pre-training***, such as ViLBERT, LXMERT and UNITER, has made tremendous progress in Vision-and-Language (V+L) research
- However, aggressive finetuning of pre-trained models often falls into the ***overfitting trap***
- ***Adversarial training*** has shown great potential in improving the generalization ability of BERT for language understanding tasks
- ***Our Contribution***: the first known effort to study large-scale adversarial training for V+L tasks

Algorithm and Backbone (UNITER)



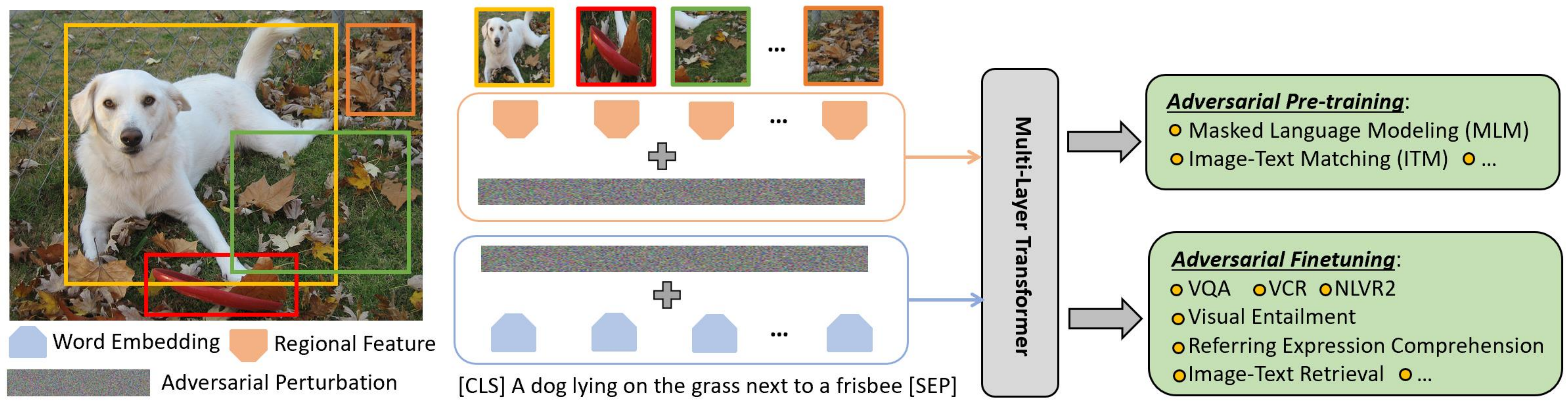
Algorithm 1 "Free" Multi-modal Adversarial Training used in VILLA.

Require: Training samples $\mathcal{D} = \{(\boldsymbol{x}_{img}, \boldsymbol{x}_{txt}, \boldsymbol{y})\}$, perturbation bound ϵ, learning rate τ, ascent steps K, ascent step size α

Initialize $\boldsymbol{\theta}$
for epoch $= 1 \ldots N_{ep}$ **do**
  for minibatch $B \subset X$ **do**
    $\boldsymbol{\delta}_0 \leftarrow \frac{1}{\sqrt{N_\delta}} U(-\epsilon, \epsilon),\ \boldsymbol{g}_0 \leftarrow 0$
    for $t = 1 \ldots K$ **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

Code is available at
https://github.com/zhegan27/VILLA

The Proposed VILLA Framework



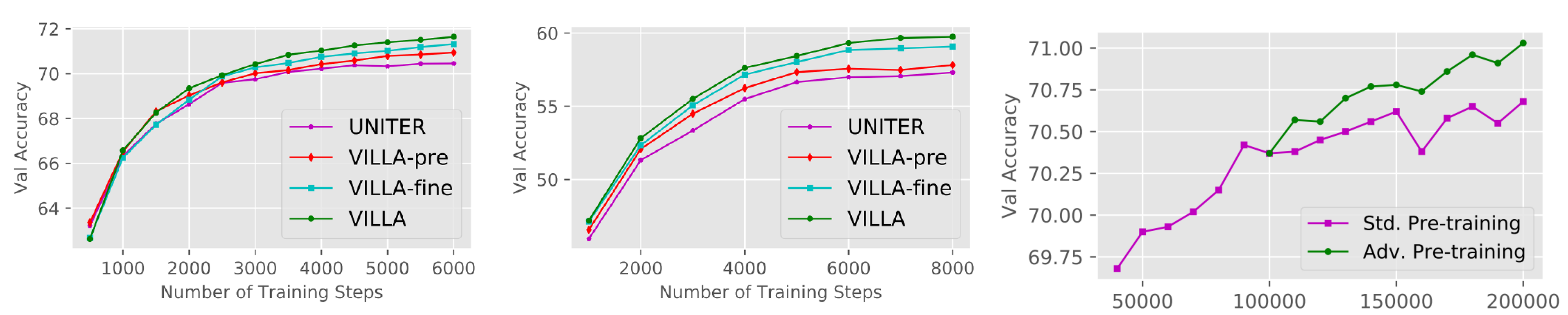
- ***Adversarial pre-training and finetuning***
- ***Perturbations in the embedding space***
- ***Enhanced adversarial training algorithm***

Experimental Results

- New state of the art on a wide range of V+L tasks (see paper for details)

Task	VQA	VCR	NLVR2	VE	RefCOCOg	RefCOCO+	Flickr30k IR	Flickr30k TR	VQA-Rep.
UNITER	74.02	62.8	79.98	79.38	75.77	66.70	75.56	87.30	64.56
VILLA	74.87	65.7	81.47	80.02	76.71	66.84	76.26	87.90	65.35

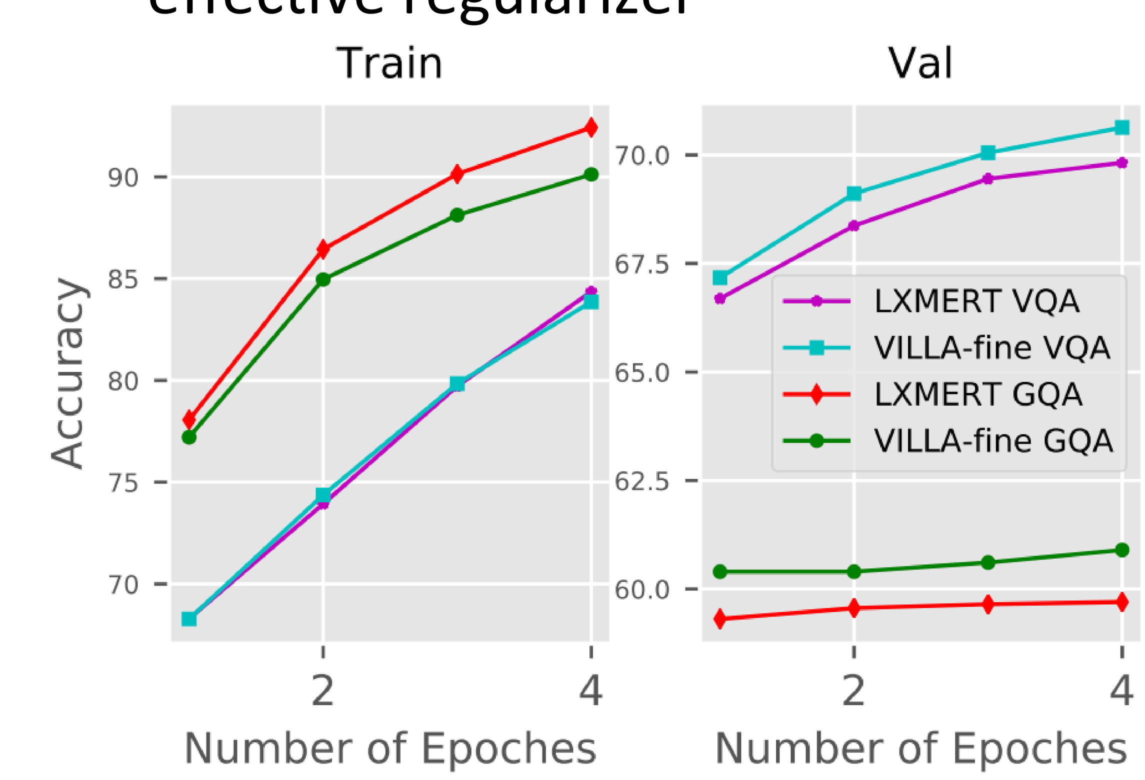
- Both adversarial pre-training (**VILLA-pre**) and finetuning (**VILLA-fine**) contribute to performance boost



(a) VQA (b) VCR

- Adversarial training on image or text modality alone is already effective
- VILLA captures richer visual coreference and visual relation knowledge than UNITER
- VILLA learns more accurate and sharper attention maps than UNITER
- VILLA is more robust to paraphrases than UNITER

- Adversarial training serves as effective regularizer



- VILLA can be readily extended to other pre-trained V+L models, such as LXMERT