SANGHANI CENTER FOR ARTIFICIAL INTELLIGENCE & DATA ANALYTICS
VIRGINIA TECH.

Emergent Corpus Pre-training Benefits Vision Language Modeling

Makanjuola Ogunleye, Chase Vickery, Ismini Lourentzou

Department of Computer Science, Virginia Tech

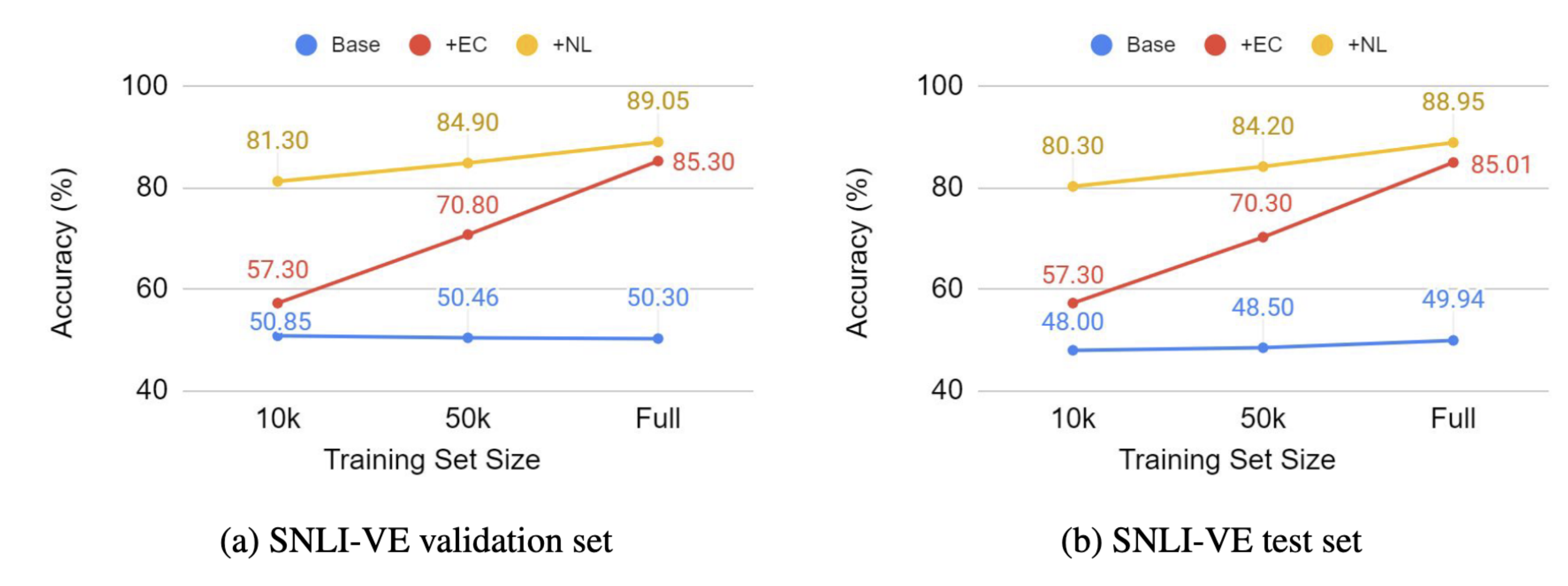
Perception & LANguage Lab
plan-lab.github.io

Introduction

TL;DR: *We pre-train OFA [4], a Vision Language Model on a corpus of Emergent Communication tokens and show that this pre-trained model improves performance on VL downstream tasks such as Visual Referring Expression and Visual Question Answering.*

- Prior Vision Language Modelling (VLM) approaches focused on building larger models with more parameters.
- Despite significant progress, VLMs still struggle to generalize effectively in low-resource scenarios.
- Emergent Communication (EC) is explored as a potential solution to this limitation. In EC, AI agents can develop languages (EC tokens) for tasks.
- We present experiments where Vision Language models are pre-trained using these EC tokens and then fine-tuned for various tasks, including VRE, VQA, and Visual Entailment (VE).
- Our experiments show significant gains in performance. VRE accuracy improves by **108.6%**, while we note an impressive gain of **69%** on the VE task.

Results on Visual Entailment



(a) SNLI-VE validation set

(b) SNLI-VE test set

Figure 1. Visual Entailment (VE) accuracy. Ablation with varying training set sizes and different pre-training/fine-tuning methods: Base, +EC Pretraining, and +NL Pretraining, corresponding to training from scratch with no pre-training, EC pre-training followed by natural language (NL) fine-tuning, and NL pre-training followed by NL fine-tuning, respectively.

Experimental Design

Vision Language Model

- We experiment with a unified vision-language (VL) model that can accommodate various VL tasks.
- Specifically, we employ OFA [4], a Transformer architecture that is suitable for both generation and classification tasks.

Pretraining and Finetuning

- We pre-train OFA entirely from scratch separately on the unified NL dataset and on the unified EC dataset.
- Each pretrained model (EC and NL) is fine-tuned independently on the three downstream vision-language tasks (VRE, VQA, and VE).
- The baseline is an OFA model with randomly initialized weights trained from scratch on each task without any pretraining.

Method Overview

Referential Game: In the referential game, a speaker agent sees an image and generates a message based on the image. The listener receives the message, and selects the correct image from a list of distractors.

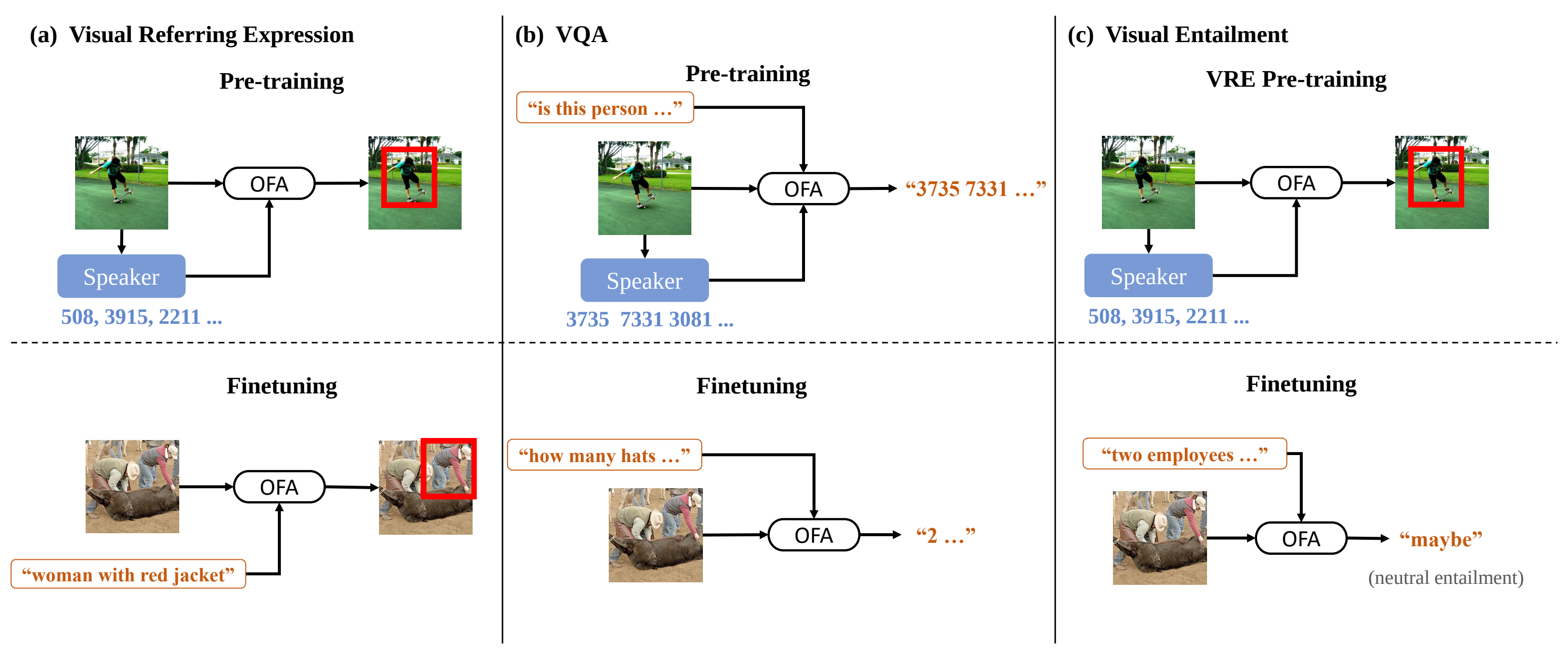


Figure 2. Method overview: (a) VRE pretraining uses EC tokens for learning instead of natural language. VRE fine-tuning trains with natural language input. (b) VQA pretraining uses EC text as a target answer while fine-tuning is performed on natural language answers. (c) Visual Entailment adopts the VRE pre-trained model for fine-tuning in order to explore EC pretraining transfer.

Qualitative Analysis

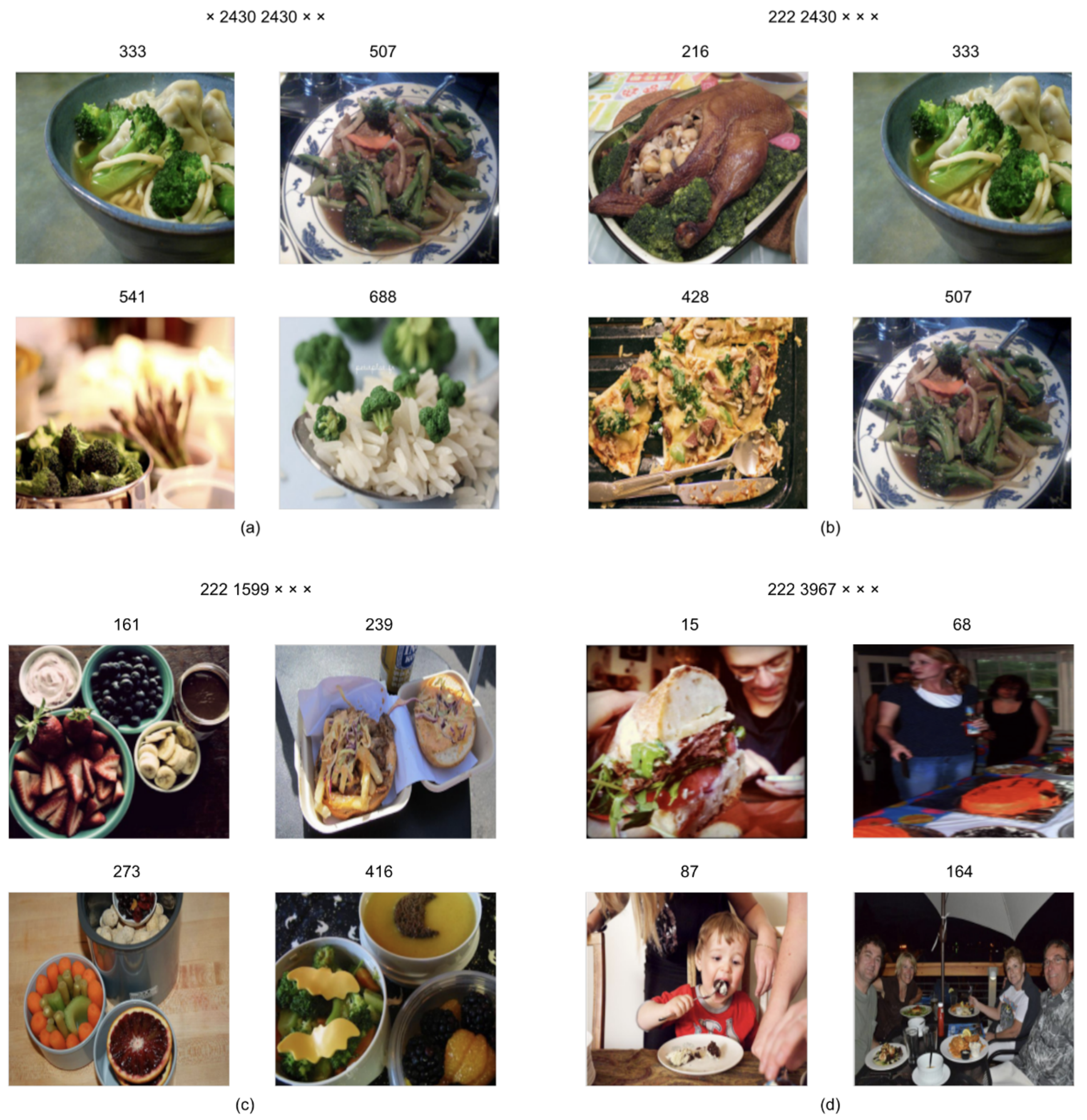


Figure 3. EC sequences of length 25, with EC tokens relating to food. (a) The repeated 2430 token is related to broccoli. (b) Token 222 is associated with the wider category of food. (c) Changing tokens after 222 changes what kind of food is described. (d) Bigram 222 3967 is still associated with food, but also people near to or eating it.

Benchmark Tasks

Visual Referring Expression

Caption: *"Boat in the water"*

Region: *<230.79,121.75,423.66,463.06>*

Visual Question Answering

Question: *What is the color of the boat?*

Answer: ***White***

Visual Entailment

Caption/Truth: *Yachts in harbor, city in background*

Hypothesis: *There are no boats in the water in front of city skyline*

Answer: ***Contradiction***

Results on VRE and VQA

- Pre-training a VLM with EC and finetune on VRE achieves more than double the accuracy (**108%** increase) compared to training from scratch.
- Comparatively we also observe a **11.5%** gain in performance on the VQA task between EC pretraining and training from scratch.

		Base	+EC	+NL
VRE	val	10.03	23.77	40.15
	testA	13.88	28.89	45.90
	testB	9.72	18.84	31.34
VQA	val	49.33	50.61	56.66
	test-dev	40.80	45.49	51.26

Table 1. **Base (No Pretraining)** - Fine-tuning OFA on natural language RefCOCO+ (VRE), VQAv2 (VQA), and SNLI-VE (VE) train sets without pretraining. **+EC Pretraining** - Fine-tuning the EC pre-trained model and **+NL Pretraining** - Fine-tuning the NL pre-trained model, both fine-tuned on natural language RefCOCO+ (VRE), VQAv2 (VQA), and SNLI-VE (VE) train sets.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *CVPR*, 2015.

[2] 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.

[3] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014.

[4] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICLR*, 2022.

[5] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual Entailment Task for Visually-Grounded Language Learning. In *NeurIPS Visually Grounded Interaction and Language (ViGIL) Workshop*, 2018.

[6] 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.

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

Github: https://makanjuola.github.io/

Poster Session

mogunleye@vt.edu