SUPPLEMENTARY MATERIALS

Anonymous authors
Paper under double-blind review

1 IMPLEMENTATION DETAILS

1.1 EMERGENT COMMUNICATION GAME

We adapt the public code[1] from Li et al. (2020b) and mostly follow their default setups. For training, we use a batch size of 256, and each batch element contains one input images and other 255 distractor images. Since the Conceptual Captions dataset has more than 2.8 million images, random sampling a batch of data is computationally costly. So for each batch, we first sample 50,000 images from the whole dataset, then sample input and distractor pairs from this subset. We use an Adam optimizer with learning rate 10^{-3}. We use a soft version of Gumbel-softmax with temperature 1, and have tried hard Gumbel-softmax and found it not further helpful for downstream performance. Each game training only takes less than 12 hours using one GeForce RTX 2080 GPU.

1.2 LANGUAGE MODELING

We use the public script[2] from Papadimitriou & Jurafsky (2020) to pre-process Wikipedia corpora of different languages, using the default setup of culling to 50,000 vocabulary size. We hand-pick downstream languages to make sure they represent different linguistic families.

We use the language modeling script[3] from Huggingface (Wolf et al., 2019) for both pre-training and fine-tuning.

We have tried grid search for the pre-training learning rate $(10^{-3}, 5 \times 10^{-4}, 10^{-4})$ and batch size $(4, 32)$, which checkpoint to transfer $(1000, 2000, 3000)$, as well as the fine-tuning learning rate $(10^{-4}, 5 \times 10^{-5}, 10^{-5})$ and batch size $(8, 32)$. We find that for all three source corpora (**es**, **ec**, **paren-zipf**), it works best to pre-train with learning rate 5×10^{-4} and batch size (32), transfer using the checkpoint with 3000 training steps, and fine-tune with learning rate 10^{-4} and batch size 8. For training from scratch, we have tried grid search for the learning rate $(10^{-3}, 5 \times 10^{-4}, 10^{-4}, 5 \times 10^{-5})$ and batch size $(4, 32)$, and find that learning rate 10^{-4} and batch size 8 work best for different downstream languages. An pre-training experiment can finish within one hour using one GeForce RTX 3090 GPU, while a fine-tuning or training-from-scratch experiment can finish within one hour using one GeForce RTX 2080 GPU.

1.3 IMAGE CAPTIONING

We use the pre-processed detection features[4] of Conceptual Captions from the codebase of Li et al. (2020a).

For both pre-training and fine-tuning, we use a public codebase[5] for image captioning based on FAIRSEQ (Ott et al., 2019), and mostly follow their default setups. Pre-training on Conceptual Captions takes 8 GeForce RTX 3090 GPU for around two days. Fine-tuning takes 1 GeForce RTX 2080 GPU for one hour.

[1] https://github.com/cambridgeltl/ECNMT/tree/master/ECPRETRAIN
[2] https://github.com/toizzy/tilt-transfer/tree/master/corpora/create_wiki_corpus
[3] https://github.com/huggingface/transformers/blob/v4.4.2/examples/language-modeling/run_clm.py
[4] https://github.com/microsoft/Oscar/blob/master/VinVL_DOWNLOAD.md
[5] https://github.com/krasserm/fairseq-image-captioning

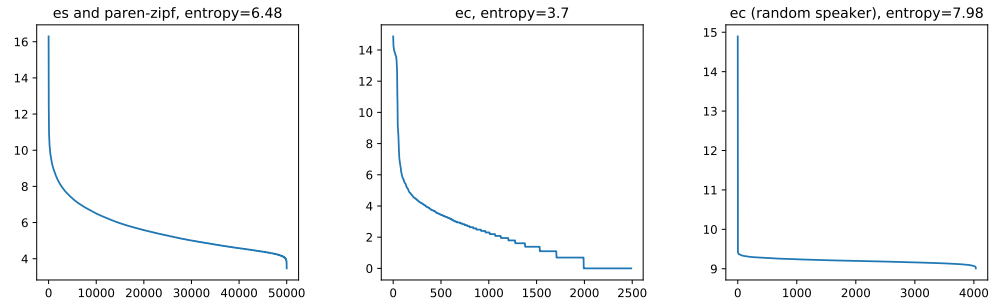

Figure 1: Unigram distributions of (1) **es** and **paren-zipf**, (2) **ec**, and (3) **ec** with **random speaker**.

Figure 2: The validation CIDEr (Vedantam et al., 2015) score across different fine-tuning epochs, when using 5,000, 50,000, or the all samples of MS-COCO training samples.

2 Additional Results

2.1 Language Unigrams

As shown in Figure 1, the **es** and **paren-zipf** corpora have a larger vocabulary size (5000) and a larger entropy (6.48). While **ec** is set with vocabulary limit 4,035, its corpus only uses around 2,500 words with smaller entropy (3.7). The **ec** corpus with **random speaker** almost has a large entropy (7.98).

2.2 Image Captioning

We visualize the fine-tuning process of image captioning experiments in Figure 2. Interestingly, we find that under different natural language resource conditions (5,000, 50,000, or all samples in the MS-COCO (Lin et al., 2014) training set) the training progress is different. Specifically, with 5,000 samples, EC or NL pre-training and training from scratch first learn similarly well, then gaps gradually appear with more training epochs. In contrast, when more han 50,0000 samples are used, the gap between pre-training methods and training from scratch is most significant when trained for only one epoch, and it starts to diminish with more training epochs. It suggests that even when downstream natural language resources are abundant, pre-training on an EC corpus might still help in a fast adaption setup.

References

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pp. 121–137. Springer, 2020a. 1

Yaoyiran Li, Edoardo Maria Ponti, Ivan Vulić, and Anna Korhonen. Emergent communication pretraining for few-shot machine translation. In *Proceedings of the 28th International Conference*

on Computational Linguistics, pp. 4716–4731, 2020b. 1

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 *European conference on computer vision*, pp. 740–755. Springer, 2014. 2

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*, 2019. 1

Isabel Papadimitriou and Dan Jurafsky. Learning music helps you read: Using transfer to study linguistic structure in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6829–6839, 2020. 1

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015. 2

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. 1