Enhancing Alignment

Region Descriptions for Contrastive Learning

- many papers use the Visual Genome dataset, consisting of images with region descriptions -> attractive source, as region descriptions are human annotated and highly curated -> focus on specfic regions of the image
- as mentioned in section TODO, we do not use Visual Genome because we encoutered problems when using it with Contrastive Learning
- Figure 1 shows accuracy on image-text contrast, which is image-text retrieval, when using data datasets in combination with Visual Genome and without

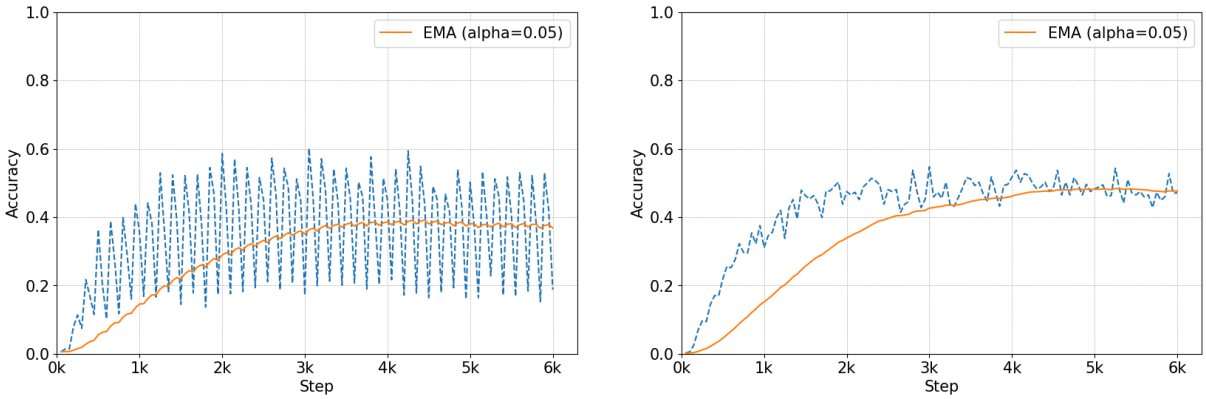

Figure 1: Training accuracy of Image-Text Contrast with Visual Genome (left) vs. without Visual Genome (right). Removing Visual Genome from the training data leads to a more stable training and a higher accuracy in the first 6k steps.

- probably specific to “smaller” models -> e.g. VLMo works well with Visual Genome -> increasingly larger models are able to exploit noisy

image-text pairs better TODO: cite

- comparison is only for the first 6k steps, as we stop experiments that show errors or do not seem promising, due to the high computational cost
- contrary to expectations, the accuracy of the model without Visual Genome continously increases after the first 6k steps (Figure 2), where is stagnates for a while

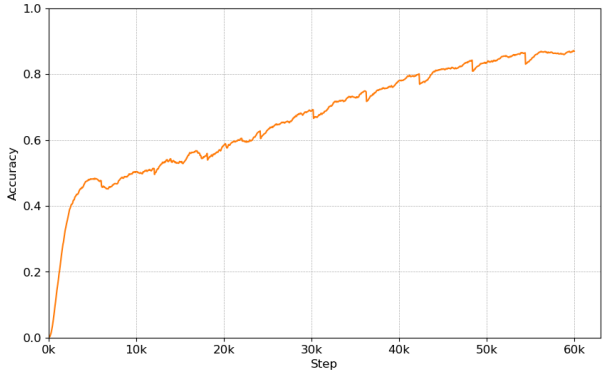

Figure 2:

- we assume reason is that region descriptions are too specific, i.e. focus on a specific part/region of the image, and do not capture the overall content of the image
- also, since the regions can be small, the caption will also be