10-701 Project Proposal

Roger Jin, Alexander Yu

1 Introduction and Problem Setup

A current topic of interest in Computer Vision is grounding, that is, linking image representations to the semantic content found in other modalities such as language. While object classification can be seen as a rudimentary form of grounding, vis-a-vis linking images to a set of known text labels, a more challenging problem is caption generation from images. The aim of this project is to use existing methods that generate representations from image and text to build a network that can jointly learn those representations. To that end, we will assess the impact of two variables on performance: 1) existing methods of generating representations and 2) architectures for the image to text model.

2 Background and Literature

The Contrastive Language Image Pretraining (CLIP) model is the primary inspiration for this project and treats caption generation as a zero-shot classification task [1]. CLIP jointly trains an image encoder and text encoder with text-image pairs. For each text-image pair, a classifier trains to associate the latent representation of the text with the latent representation of the image (hence its ”contrastive” nature). At test time, CLIP is provided with an image and a sentence fragment to complete. For example:

”A photo of a __________, a type of food”

Leveraging context from the incomplete sentence and information from the image allows CLIP to achieve roughly 40% zero-shot classification accuracy on the ImageNet dataset. While generation of captions from scratch is not CLIP's intended purpose, captions can be generated by passing CLIP's previous output sentence as the input fragment to a subsequent iteration.

3 Methods/Model

While CLIP treats the association between text and images as a classification task, we propose a model which associates text and images through regression. The steps to constructing this proposed model are as follows:

1. Train or find a pre-trained image autoencoder. Convolutional Res-Net Autoencoders are a promising option.

2. Train or find a pre-trained text autoencoder. Tranformers are a promising option [2].

3. Train a regression model (denoted the ”Transform Network”) which transforms the image's latent representation to the text's latent representation. While training the Transform Network, the text and image autoencoders' weights are frozen.

This allows for end-to-end caption generation, whereas CLIP is only able to fill in missing phrases in sentences from a predefined set of classes. Furthermore, our approach does not require that the image and text autoencoders be jointly trained; each autoencoder may be pre-trained or trained separately and the Transform Network should be trainable on a smaller set of text-image pairs. However, our proposed model loses CLIP's zero-shot capabilities. Figure 1 shows a graphical representation of the model.

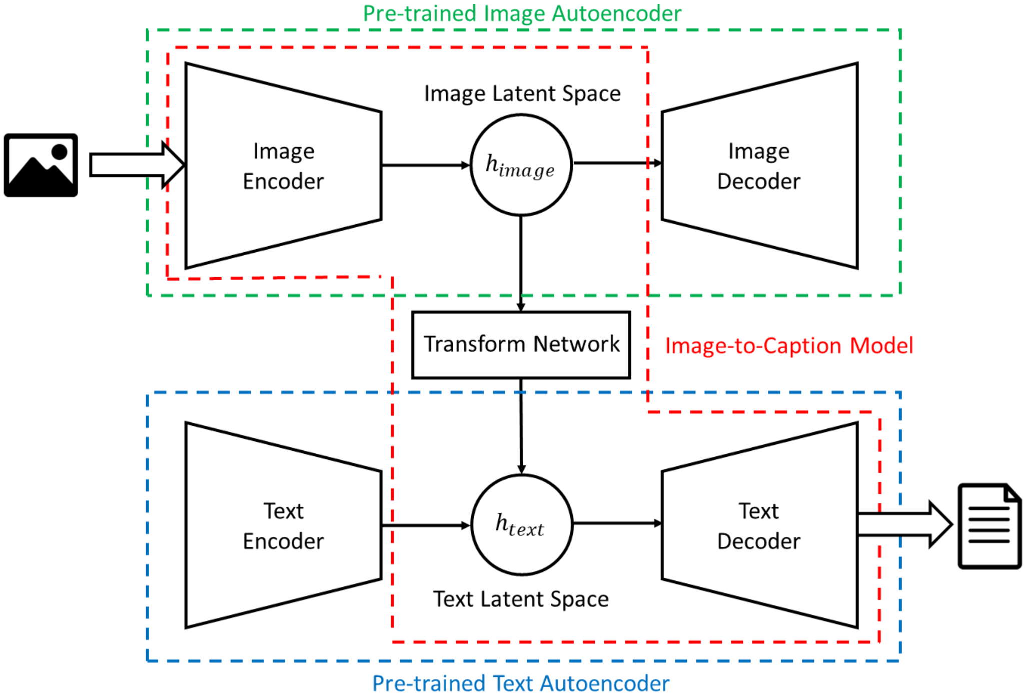

Figure 1: Overview of the proposed image to caption network. The image decoder and text encoder are used for training only and are not used at test time. Gradients will *not* propagate through the Transform Network into the image encoder.

4 Data

To train the image to text model, we will use the Common Objects in Context (COCO) dataset which contains roughly 200,000 annotated images with 5 captions per image [3]. All images contain at least one of 91 object types, each of which can be easily identified in an image by a human. Because the objects in its constituent images are readily identifiable, COCO is a useful benchmark in assessing the baseline capabilities of our text to image model; if we cannot even learn captions for common objects, it may be harder yet to learn captions for uncommon ones.

5 Evaluation Criteria

We will evaluate the quality of generated captions both quantitatively and qualitatively. In terms of quantitative evaluation, we will compare the learned text representation to the learned text representation of each of the ground truth captions on a holdout set of COCO. In terms of qualitative evaluation, we will generate a separate small test set containing objects not found in the COCO dataset and determine whether the generated captions are accurate. Since caption generation is, at its core, an exercise in grounding, testing on unseen examples is needed to interrogate the generality of the semantic content the model can link to images. For both the qualitative and quantitative evaluations, we will compare against existing pre-trained models such as CLIP as baselines.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, “Learning transferable visual models from natural language supervision,” in *International Conference on Machine Learning*, pp. 8748–8763, PMLR, 2021.

[2] C. S. Wickramasinghe, D. L. Marino, and M. Manic, “Resnet autoencoders for unsupervised feature learning from high-dimensional data: Deep models resistant to performance degradation,” *IEEE Access*, vol. 9, pp. 40511–40520, 2021.

[3] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: common objects in context,” *CoRR*, vol. abs/1405.0312, 2014.