Testing Egunean Behin Visual Question Answering Dataset with BLIP

Julen Etxaniz
University of the Basque Country (UPV/EHU)
jetxaniz007@ikasle.ehu.eus

Abstract

Egunean Behin is a popular Basque quiz game. The game consists on answering 10 daily multiple choice questions. Questions were translated to English because VQA models like BLIP are mainly trained on English questions. Three types of questions from the game were selected: figures, cubes and maze. All the images and questions were generated automatically. There are multiple questions for each image. Questions require counting figures, colors, cubes and understanding the dimensions of the pictures. Each question has one correct and two wrong answers. These can be used for multiple choice question answering.

1 Introduction

Egunean Behin is a popular Basque quiz game. The game consists on answering 10 daily multiple choice questions. Questions were translated to English because VQA models like BLIP (Li et al., 2022) are mainly trained on English questions. Three types of questions from the game were selected: figures, cubes and maze. All the images and questions were generated automatically. There are multiple questions for each image. Questions require counting figures, colors, cubes and understanding the dimensions of the pictures. Each question has one correct and two wrong answers. These can be used for multiple choice question answering.

2 Related Work

2.1 Vision-language Pre-training

Vision and language are two of the most fundamental methods for humans to perceive the world. An important goal of AI has been to build intelligent agents that can understand the world through vision and language inputs, and communicate with humans through language.

Vision-language pre-training has emerged as an effective approach to achieve this goal. Deep neural network models are pre-trained on large scale image-text datasets to improve performance on downstream vision-language tasks, such as image-text retrieval, image captioning, and visual question answering.

Models are commonly pre-trained before they are fine-tuned on each task. Fine-tuning involves additional training of the pre-trained model, using data from the downstream task. Without pre-training, the model needs to be trained from scratch on each downstream task, which leads to worse performance.

Despite the success of vision-language pre-training, existing methods have two major limitations related to models and training data.

From the model perspective, most existing pre-trained models are not flexible enough to adapt to a wide range of vision-language tasks. On the one hand, encoder-based models such as CLIP (Radford et al., 2021) and ALBEF (Li et al., 2021) are less straightforward to directly transfer to text generation tasks. On the other hand, encoder-decoder models like SimVLM (Wang et al., 2021) have not been successfully adopted for image-text retrieval tasks.

From the data perspective, most models are pre-trained on image and alt-text pairs that are automatically collected from the web. However, these web texts often do not accurately describe the images, making them a noisy source of supervision.

2.2 Bootstrapping Language-Image Pre-training

To address these limitations, BLIP: Bootstrapping Language-Image Pre-training (Li et al., 2022) introduces two contributions, one from each perspective.

On the one hand, Multimodal mixture of Encoder-Decoder (MED) is a new model architec-

ture that enables a wider range of downstream tasks than existing methods. An MED can operate either as a unimodal image or text encoder, or an image-grounded text encoder, or an image-grounded text decoder.

On the other hand, Captioning and Filtering (CapFilt) is a new dataset bootstrapping method for learning from noisy web data. A captioner model produces synthetic captions given web images, and a filter model removes noisy captions from both the original web texts and the synthetic texts.

BLIP achieves state-of-the-art performance on five vision-language tasks: image-text retrieval, image captioning, visual question answering, visual reasoning, and visual dialog. It also achieves state-of-the-art zero-shot performance on two video-language tasks: text-to-video retrieval and video question answering.

2.3 Visual Question Answering (VQA)

Visual Question Answering (Antol et al., 2015) is a popular vision and language task. Given an image and a question about the image, the task is to provide an accurate answer. VQA dataset is commonly used as a benchmark to evaluate VQA systems. Questions are generally open-ended but multiple choices are provided for some questions. If we compare it to image captioning, VQA requires a more detailed understanding of the image and more complex reasoning (Antol et al., 2015).

3 Material and Methods

This section first introduces the new model architecture MED and its pre-training objectives, then explains dataset bootstrapping and finally finetuning architecture.

3.1 Model Architecture

In order to pre-train a unified vision-language model with both understanding and generation capabilities, BLIP introduces a multimodal mixture of encoder-decoder (MED) model which can operate in three functionalities. The model architecture can be seen in Figure 1.

(1) Unimodal encoders are trained with an image-text contrastive (ITC) loss to align the image and text representations. The image encoder is a visual transformer (Dosovitskiy et al., 2020), which divides an input image into patches and encodes them as a sequence of embeddings. The text encoder is the same as BERT (Devlin et al., 2018), where a `[CLS]` token is appended to the beginning of the text input to summarize it.

(2) Image-grounded text encoder is trained with a image-text matching (ITM) loss to distinguish between positive and negative image-text pairs. It has a cross-attention layer between the self-attention layer and the feed forward layer for each transformer block. The output embedding of the `[Encode]` token is used as the multimodal representation of the image-text pair.

(3) Image-grounded text decoder is trained with a language modeling (LM) loss to generate captions for given images. It replaces the bi-directional self-attention layers in the text encoder with causal self-attention layers. A `[Decode]` token is used to as the beginning of a sequence.

In order to perform efficient pre-training and improve multi-task learning, the text encoder and text decoder share all parameters except for the SA layers.

3.2 Dataset Bootstraping

The learning framework can be seen in Figure 2.

3.3 Finetuning

On different downstream tasks, different paths of the pre-trained model are finetuned to achieve different objectives. The finetuning architecture for VQA can be seen in Figure 3.

4 Results

5 Conclusions

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

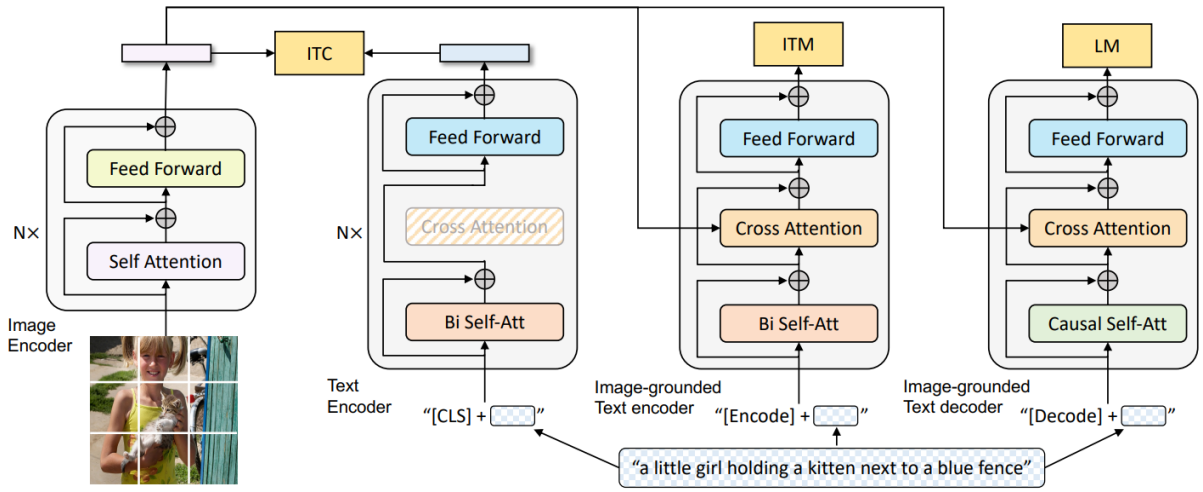

Figure 1: BLIP pre-training model architecture: multimodal mixture of encoder-decoder.

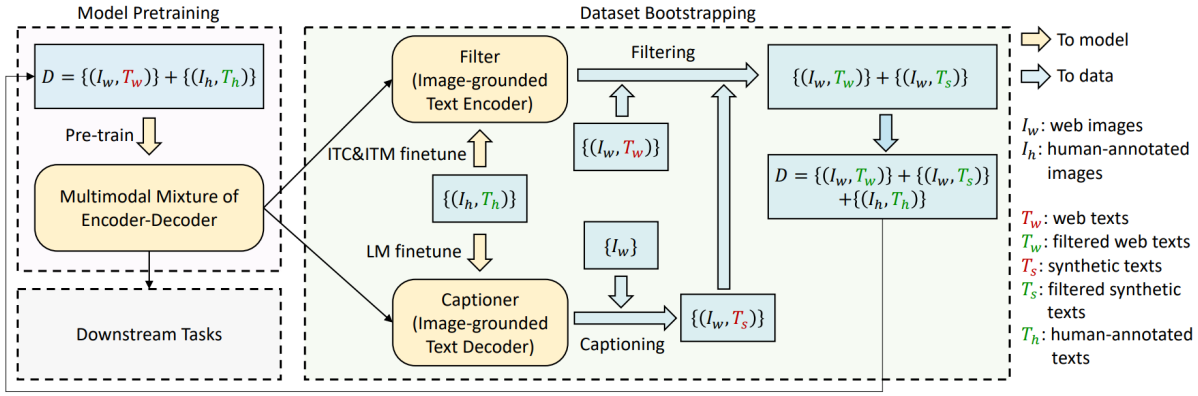

Figure 2: BLIP learning framework: a captioner to produce synthetic captions and a filter to remove noisy captions.

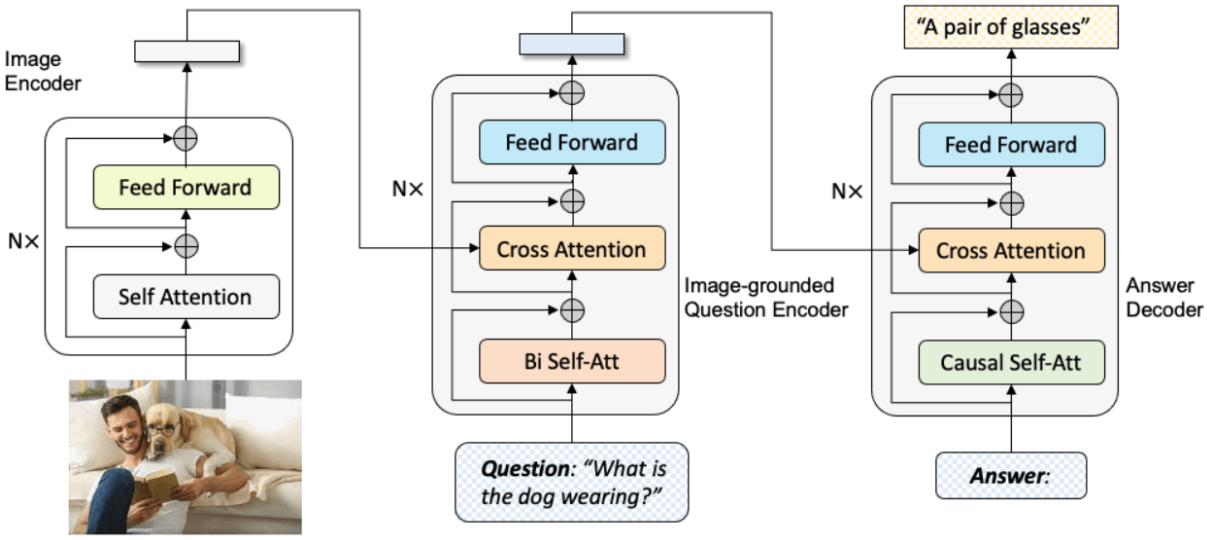

Figure 3: BLIP VQA finetuning architecture.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.