Testing Egunean Behin Visual Question Answering Dataset with BLIP (Bootstrapping Language-Image Pre-training)

github.com/juletx/egunean-behin-vqa

Julen Etxaniz
UPV/EHU
jetxaniz007@ikasle.ehu.eus

huggingface.co/spaces/Salesforce/BLIP

1. Egunean Behin VQA Dataset

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.

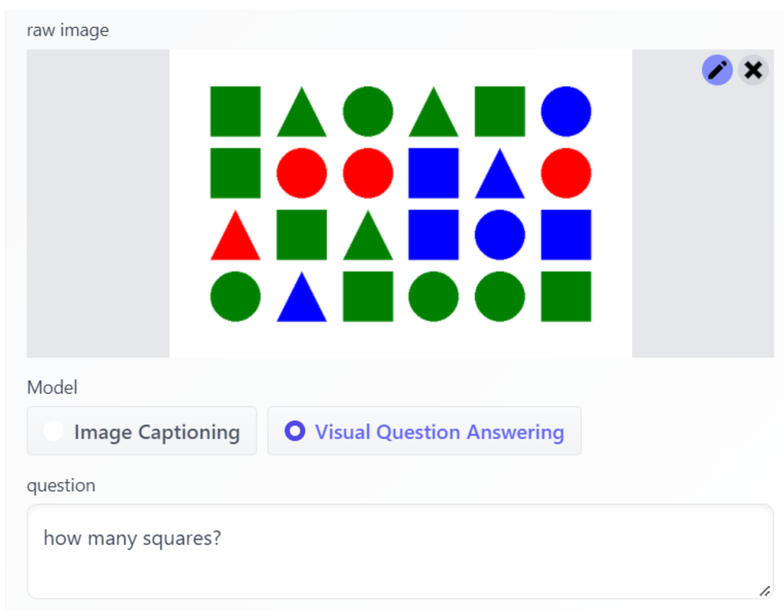
Three types of questions from the game were selected: figures, cubes and maze. All the images and questions were generated automatically.

There are multiple questions for each image. Questions require counting figure, 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.

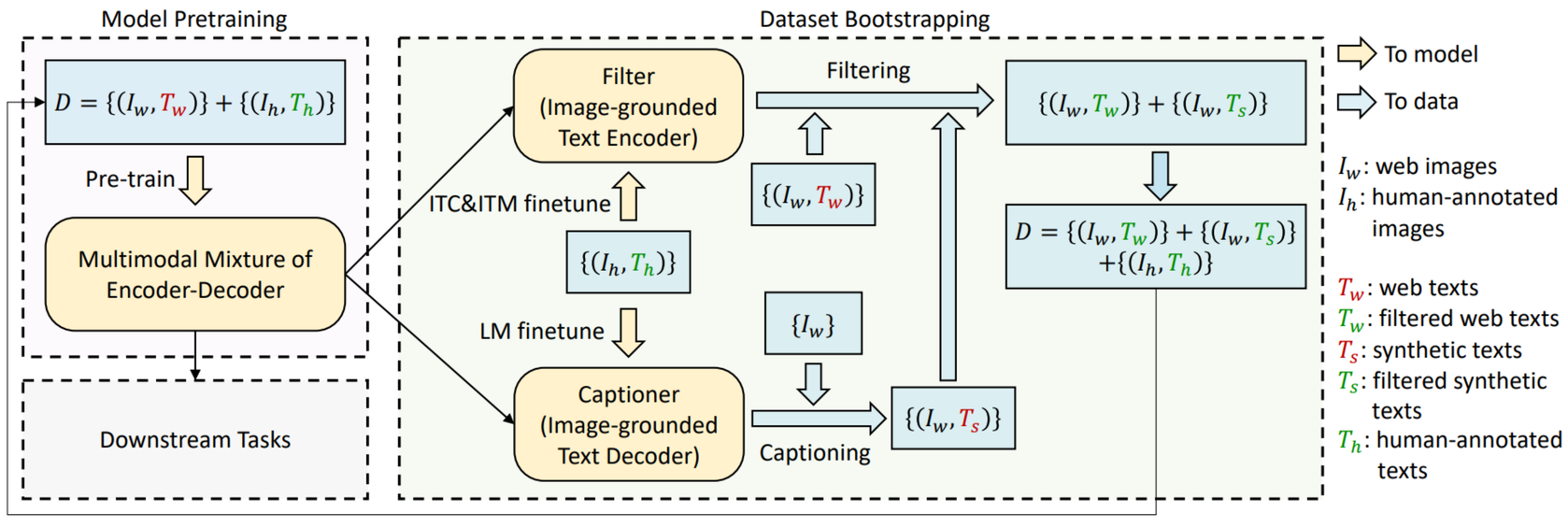
2. Testing

Testing BLIP in an out of domain dataset as open-ended questions with no options. Most answers are wrong, but close to the correct answers.

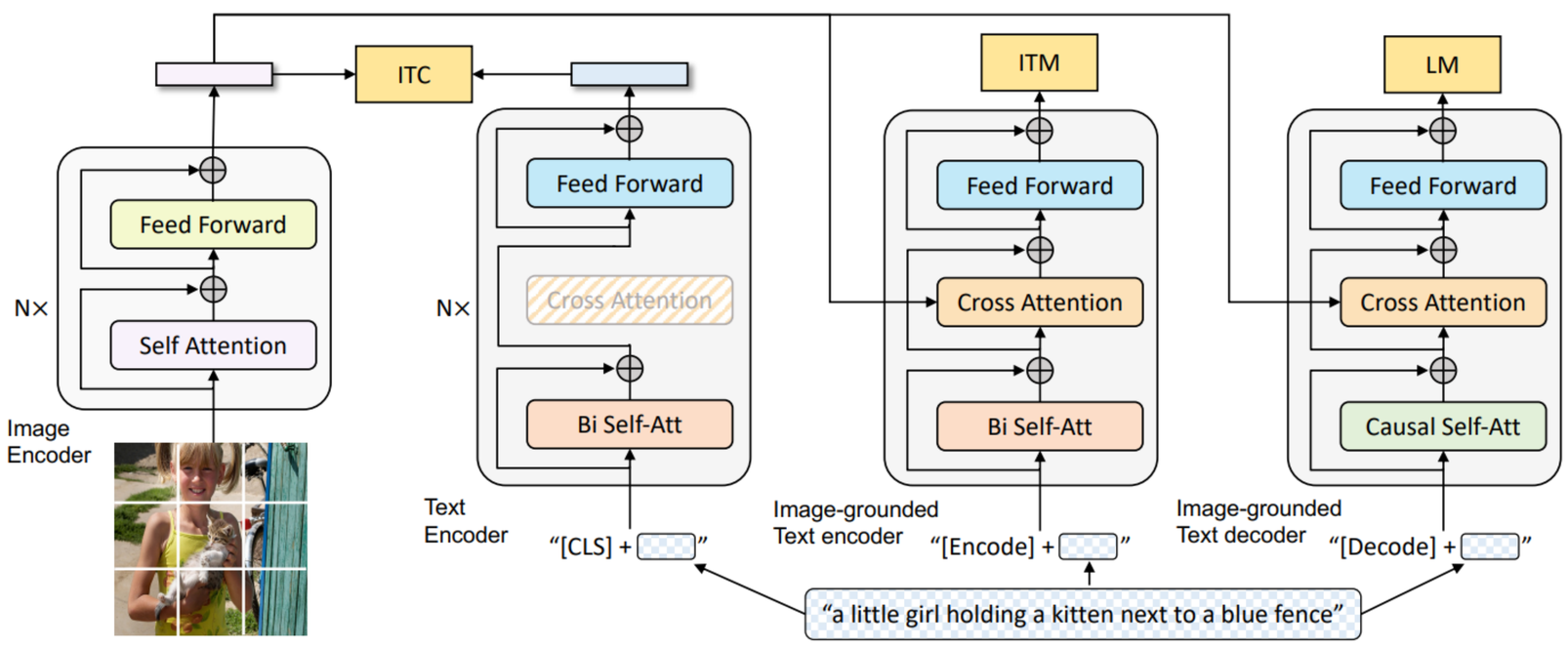


3. Framework

A captioner is used to produce synthetic captions for web images, and a filter to remove noisy image-text pairs. They are initialized from the same pre-trained model and finetuned individually on a small-scale human-annotated dataset.



4. Pretraining



Unified vision-language model which can operate in one of the 3 functionalities:

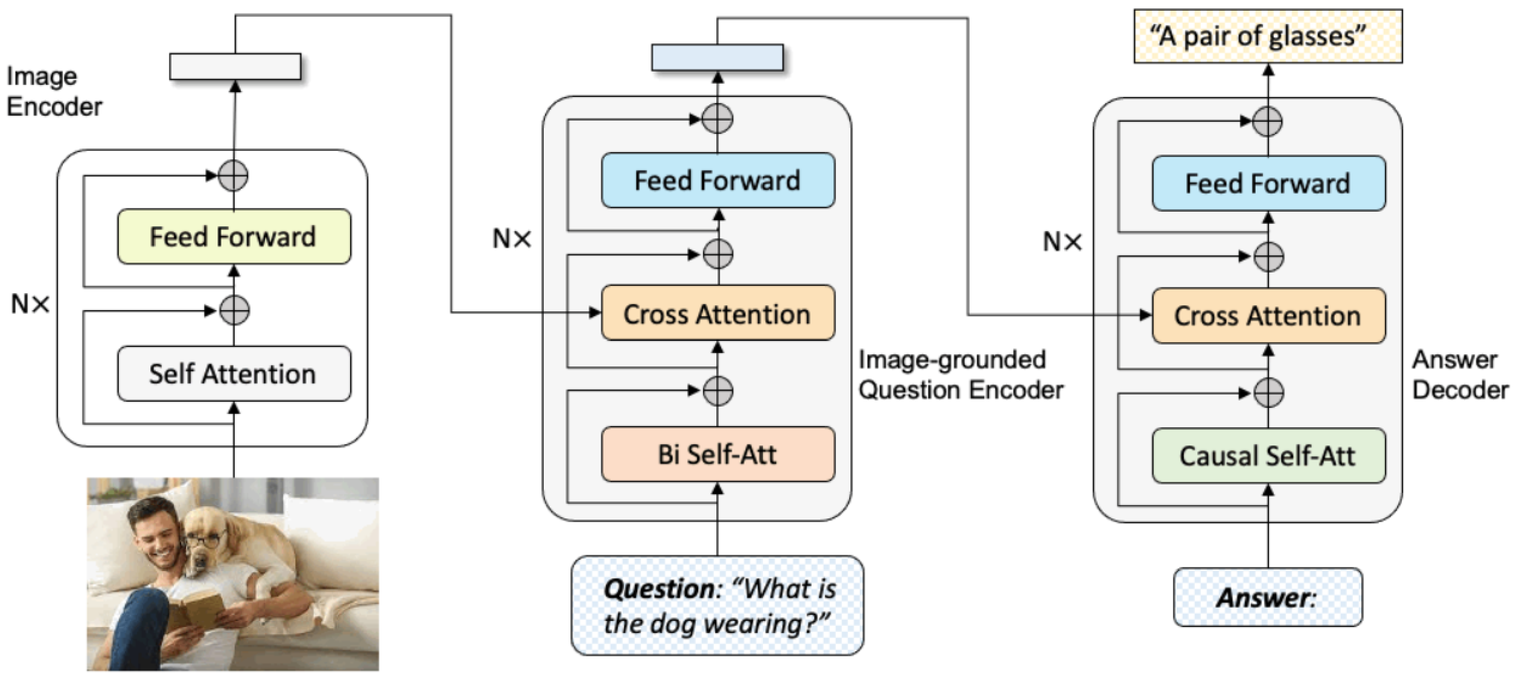
(1) Unimodal encoder is trained with an image-text contrastive (ITC) loss to align the vision and language representations.

(2) Image-grounded text encoder is trained with a image-text matching (ITM) loss to distinguish between positive and negative image-text pairs.

(3) Image-grounded text decoder is trained with a language modeling (LM) loss to generate captions given images.

5. Finetuning

During finetuning, rearrange the pre-trained model, where an image-question is first encoded into multimodal embeddings and then given to an answer decoder. VQA as an answer generation task, open-ended.



6. Results

Using 14M images, BLIP outperforms ALBEF by +1.64% on the test set. Using 129M images, BLIP achieves better performance than SimVLM which uses 13× more pre-training data.

Method	Pre-train #Images	VQA test-dev	VQA test-std
LXMERT	180K	72.42	72.54
UNITER	4M	72.70	72.91
VL-T5/BART	180K	-	71.3
OSCAR	4M	73.16	73.44
SOHO	219K	73.25	73.47
VILLA	4M	73.59	73.67
UNIMO	5.6M	75.06	75.27
ALBEF	14M	75.84	76.04
$\text{SimVLM}_{\text{base}}$†	1.8B	77.87	78.14
BLIP	14M	77.54	77.62
BLIP	129M	78.24	78.17
$\text{BLIP}_{\text{CapFilt-L}}$	129M	**78.25**	**78.32**