Visual Question Answering using Visual Transformers

Julen Etxaniz Aragoneses

Tutor: Gorka Azkune

1 Description

Visual Question Answering (VQA) [1] consists of predicting an answer given an image and a question. This is a multimodal task that requires language and vision understanding. I will use visual transformers [2] for this task.

2 Related work

Many previous approaches have achieved good results in VQA using visual transformers.

VisualBERT, a multimodal masked transformer, achieved 71% in VQA in 2019 [3].

SimVL, an encoder-decoder transformer model, got 78% in VQA in 2021 [4]. It achieved SOTA results in 6 VL benchmarks: VQA, Visual entailment, Image captioning… It also showed some zero-shot capabilities for those tasks.

BLIP achieved a slightly better performance than SimVL using much less pretraining images in 2022 [5]. It achieves state-of-the-art results on a wide range of vision-language tasks, such as image-text retrieval, image captioning and VQA.

3 Data

The dataset commonly used is VQA: https://visualqa.org/. VQA is a new dataset containing open-ended questions about images. These questions require an understanding of vision, language, and common-sense knowledge to answer. It has the following features.

- 265,016 images (COCO and abstract scenes)
- At least 3 questions (5.4 questions on average) per image
- 10 ground truth answers per question
- plausible (but likely incorrect) answers per question
- Automatic evaluation metric

As the dataset is quite big, I will have to use a smaller dataset because of the time and computation limitations.

4 Algorithm

I will use visual tranformers for this task. The architecture might be based on some of the previously mentioned models. However, the size of the transformer will be much smaller. Another option is to use a pre-trained multimodal model and fine-tuning it for this task.

5 Evaluation

I will use the automatic evaluation metric provided by VQA. I will also evaluate some examples qualitatively. The results are expected to be much worse than the previously mentioned ones, due to the limitations in the model and dataset sizes.

References

[1] Anton et al. VQA: visual question answering.

[2] Wu et al. Visual Transformers: Token-based Image Representation and Processing for Computer Vision

[3] Li et al. VisualBERT: A Simple and Performant Baseline for Vision and Language. 2019

[4] Wang et al. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. 2021

[5] Li et al. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. 2022