MACLEAN 2021 Workshop

Machine Learning for Earth Observation

Paper review

How to find a good image-text embedding for remote sensing visual question answering?

Background

Visual question answering (VQA) is a computer vision workflow that enables the accessibility of image information in human language. Through VQA, we can provide answers to both open-ended and closed-ended questions in natural language about an image such as how many cars are there? How has the forest changed? This presents itself as both an appealing and challenging opportunity for image analysts. With the exponential growth of remote sensing images in the last few years, VQA opens the pathway to extend the use of remote sensing as a tool for anyone (both experts and non-experts) to benefit from the data abundance. In fact, in the IEEE position paper, "*Towards a collective agenda on AI for Earth science data analysis*", VQA was highlighted as one of the six promising directions in the agenda of AI for Earth Science Data Analysis.

What is the problem under study and why is it important?

VQA for remote sensing follows two data processing streams: image analysis and text mining. These two stand-alone data processing streams are combined in a dedicated fusion step before predicting an answer. **In this paper**, the authors sought to assess the required level of complexity at the fusion step to understand the interplay between the features extracted from the questions and the image. This was implemented by comparing two fusion strategies with a baseline (elementwise) strategy used by Lobry et al, 2020 in the curation of the **RSVQA. Justifiably,** the element-wise approach was limited in interactions since it strictly requires the matching of image and question contents at the same index in the latent space before fusion. Examining the performance of other fusion methods that have been tested for VQA but not on multi-dimensional images like remote sensing images, is crucial to developing more dynamic VQA systems.

How was this problem approached?

In this work, three fusion strategies were tested and compared – i.) Element-wise fusion strategy, ii.) MCB [1](ruled by randomness) and iii.) MUTAN [2](breaks down the full fusion into projections specific to each modality based on the Tucker decomposition). Methodologically, a simple structure of having two feature extractors before the fusion operations and followed by a classification network was maintained following the work of Lobry et al, 2020. The image feature extraction was based on ResNet-152, pre-trained on ImageNet, and the Questions feature extraction was based on skip-thoughts, pre-trained on BookCorpus. To predict the answer, classification was done with one fully connected layer and one output layer that contains as many classes as there are answers based on Lobry et al, 2020. This was tested on both low resolution (Sentinel 2 consisting of 77,232 questions-answers-images triplets) and very high-resolution aerial images in the north coast of USA (consisting of 1,066,316 triplets paired with questions and answers on various tasks).

[1] Multimodal Compact Bilinear pooling

[2] Multimodal Tucker Fusion for Visual Question Answering

The question-answers-images triplets were split into various tasks -- presence/ absence classification, object counting. Overall, a total of 6 models were tested and an ablation study on the MCB fusion method on the low resolution to examine the sensibility of the input dimension of the performance with feature dimension d = {1200, 4000, 8000, 16000, 32000}. The feature dimension 8000 was found to be the most optimal dimension.

How does it relate or differ from existing works?

This work shows it relevancy by building on and extending the methods of Lobry et al, 2020 and others in the broad spectrum of VQA and specifically VQA for remote sensing images. In this case, it tested and prove the use of a new fusion strategy in the overall workflow of the VQA for remote sensing images developed by Lobry et al. In all, it proves that the VQA model performance increases with more model complexity both in low resolution and high-resolution images with a dependence on the questions type.

What I think of the paper.

This work presented an interesting and very specific research question to understand the complexity of fusion methods and their influence on the performance of prediction outputs. This is an important objective as it would allow for better interactive VQA products and possibly the development of a VQA engine (like a CBIR or Image search engine). Along with this, is the clear presentation of problem statements and objectives. Also, the paper followed guidelines for reproducibility by providing external link to the data and shows strong pointers to the background paper that builds up into the specifics of its research objectives.
However, the paper was specifically focused on combining modalities with equally sized embeddings in the vector space. It would have been interesting to see an evaluation of the performance of these fusion strategies on unequally sized embeddings. It would also have been great to see the paper on the *Paperswithcode* platform for other researchers who might be interested in examining the results and replicating it for other fusion strategies other than the ones mentioned in the paper.