Unifying Sketch, Text and Photo

Aim

To learn a common unified embedding space that can encode any modality. In this case, focus is placed on three main modalities - sketches, text and photo. This will be evaluated on scene retrieval tasks like:

a. Fine-Grained Sketch Based Image Retrieval (FG-SBIR)
b. Fine-Grained Text Based Image Retrieval (FG-TBIR)
c. Fine-Grained Sketch + Text Based Image Retrieval (FG-STBIR)

Motivation

One of the major drawbacks in multimodal learning is the ability to find ***paired data*** readily. In our case, finding a single dataset with paired annotations for all 3 modalities - sketch, image and text is a major challenge. Also, In current approaches to multimodal processing, ***separate feature extractors*** are used for each modality which requires us to re-tune the architecture for each combination. By designing a unified framework that can fuse information from all three modalities, we can use a single model for all of the above-mentioned tasks.

Challenges

Essentially, we need to build a model that will accept inputs with the following criteria:

a. First, it should be **permutation invariant** — the output of the model should not change under any permutation of the elements in the input modalities.
b. The model should be able to process input sets of any **size**.
c. The model should be able to handle any **number** of inputs

Proposed Plan

1. The challenges mentioned in the previous slide can be tackled by leveraging models that were designed to tackle set-input problems. In this work, components from the Set Transformer [1] architecture, which was designed to handle set-inputs, will be utilized in designing our novel architecture.
2. The Perceiver [2] architecture has proven exceptional at handling a variety of input modalities as well as multi-modal inputs. Attention mechanisms and Transformer blocks from this architecture will be heavily used in our novel architecture

Objectives

1. Identify and perform initial analysis of the relevant datasets to fully realize the problem of multi-modal scene understanding with image, text and sketch modalities.
2. Explore the SketchyCOCO dataset and filter it to make it friendly for retrieval.
3. Establish baselines for the three main retrieval tasks by using leveraging modality-specific feature extractors
4. Explore more nuanced architectures for Sketch+Text-based image retrieval and compare the results with that of a single input (sketch or text) based image retrieval model.
5. Design a single modality-agnostic architecture that can jointly model distributions of all three modalities.
6. Conduct experiments with the newly designed architectures to perform tasks like (i) fine-grained sketch-based image retrieval (FG-SBIR), (ii) fine-grained text-based image retrieval (FG-TBIR), (iii) fine-grained sketch+text based image retrieval (FG-STBIR)

Objectives 1 - 3 have been accomplished.

What has been accomplished so far

1. Obtain a dataset with sketch, image and text triplets by combining SketchyCOCO dataset and the original MSCOCO dataset.
2. Dataset filtering has been done to include only scene-level sketches
3. Preliminary experiments have been carried out using a vanilla architecture to implement the three main tasks

Dataset Information

The original SketchyCOCO dataset has 14000 sketches. The custom filtered dataset now has 10000+ scene sketches with captions. The dataset filtering was done by using only those sketches where the number of foreground images were >=1. Corresponding captions for the photos were obtained using the MSCOCO dataset.

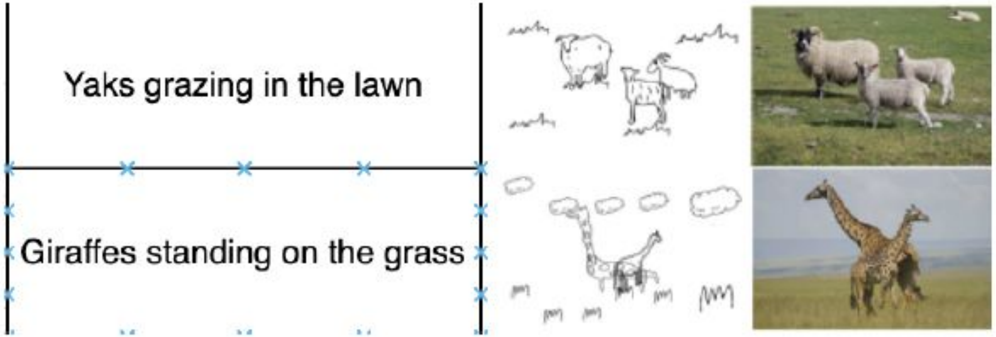

Figure 3.1: Examples of our fine-grained scene-level sketch dataset

Experimental results

Experiments	Top-1 accuracy(%)	Top-10 accuracy(%)
FG-SBIR	16.5	41.2
FG-TBIR	12.9	34.2
FG-STBIR	15	40

Observations

1. Between fine-grained **sketch** based IR and fine-grained **text** based IR, the former produces superior results compared to the latter. This could be because sketch works as a more efficient modality as it conveys more salient information about the inputs
2. Ideally, The performance of the model after combining the two modalities should be higher than its single modality counterparts. In the future, rigorous investigation will be performed to understand the drop in performance using visualization approached like GradCAM.
3. One of the main reasons why the obtained accuracies for all the three models is low is because the support set used during evaluation in this case is considerably larger than what was used in the benchmarks established by the original authors.

Future Plans

- In the future, we need to explore more nuanced architectures for Sketch+Text-based image retrieval and compare the results with that of a single input (sketch or text) based image retrieval model.
- A novel architecture that combines the benefits of Set Attention [1] and Perceiver [2] will be designed to explore modality-agnostic retrieval tasks.
- Frequent sanity checks in the form of unit testing, validation testing and integration testing have to be conducted to ensure bug-free code.
- Different training conditions have to be experimented with. Finally, a detailed ablation study with the new architecture must be carried out on multiple scene-level datasets.

References

1. Lee, J. , Lee, Y. , Kim, J. , Kosiorek, A. , Choi, S. , and Teh, Y. W. . Set transformer: A framework for attention-based permutation-invariant neural networks. In International Conference on Machine Learning, pages 3744–3753. PMLR, 2019.
2. Jaegle, A. , Gimeno, F. , Brock, A. , Vinyals, O. , Zisserman, A. , and Carreira, J. . Perceiver: General perception with iterative attention. In International Conference on Machine Learning, pages 4651–4664. PMLR, 2021.