Project Report: Attempts to Fitting A3DS Dataset into Issue Sensitive Captioning System

Ayodeji Olupinla 5522589
Christian Gerber 4157221
Wanzhao Zhang 5464375

1 Project Idea

As a starting point, we chose the Issue Sensitive Captioning System (ISIC) by Nie et al. (2020); which is an advanced RSA model for generating issue related captions of given pictures (Original codes and the paper: https://github.com/windweller/Pragmatic-ISIC). Two datasets were used in the experiment by Nie et al., the first one is the Caltech UC San Diego-Bird (CUB) dataset, which contains pictures of birds of different species. Each of the pictures has *property::value* structure attributes, for example *has_wing_color::brown*. Annotated captions can be generated from these attributes from both human and machine learning models. The second dataset used in the paper is MS COCO dataset, which covers different categories of everyday objects and scenes, compared to the CUB dataset, MS COCO has a broad coverage and is less controlled or detailed annotated.

However, even the detailed annotated CUB dataset is complicated and might not be the perfect dataset to test the model: Firstly, there are 312 attributes in total and they are arranged hierarchically; secondly, not all pictures have the same attributes and they have non-exhaustive captions. In order to see how the model works, we decided to use the 3D Shapes Dataset (A3DS) by Burgess & Kim (2018) (Description and download link: https://github.com/deepmind/3d-shapes/tree/master) and tried to fit it into the issue-sensitive captioning system. The first goal of our project is to reimplement the ISIC system using the A3DS dataset; the second goal is to adjust the rationality and entropy penalty to see how these hyperparameters can actually affect the quality of generated sentences (if we could successfully reimplement the model).

Before we really put our hands on the code provided by the author, we read the paper and expected a result of having high rationality, the generated captions would be more "random" than being "readable" and grammatical; having high entropy penalty will result in less detailed, short and simple caption only with the desired issue information and less details unrelated to the current issue.

2 Dataset

Instead of using the original A3DS dataset from Burgess & Kim (2018), we tried to use the reduced sandbox version of this dataset by Polina Tsvilodub (link to the project: https://github.com/polina-tsvilodub/3dshapes-language) because it has advantages compared to the original one: Firstly, the sandbox version requires smaller space on the hard drive; secondly, the numeric features are already paired to literal descriptions, which is easier to process (for example we can pair the numeric feature 0.4 of "WALL_HUE" to "light green"). Besides, the sandbox dataset also provides a list of sampled IDs, long and short captions for each item; a vocab file, a .json file as a dictionary mapping the similar items into different categories. We found these features would be very helpful for our project because "issue sensitive" is the most important point of out project. Finally, in the sandbox dataset, each item has a "standard caption", which could be used as an alternative of human evaluation for evaluating the generated sentences.

3 What we tried to do

3.1 The original code from Nie et al. (2020)

Before we started implementing the A3DS dataset. We tried to execute the original code provided by Nie et al. (2020). We roughly examined the code then we decided to run it with the CUB dataset. We failed to build the desired environment by following the instructions on the GitHub page because some of the packages are not available anymore and the link for downloading the dataset is also not available. After changing the settings, we built the conda environment for running the code and also downloaded the CUB and MS COCO dataset successfully. Firstly, we tried to execute the evaluation Python file multiple times with different generating options (S0, S1, S1_Q and S1_QH), a folder named "results" showed up with .json files but no caption was actually generated (no error message showed up at this stage). We also tried to execute the notebook file but the model stopped loading because "file not found". After trying to move the CUB training data under different folders and also trying to modify the path in the source code, the model still failed to load correctly. Without further instruction from the original GitHub page, we failed to execute the original code.

3.2 Our first attempt of using LSTM image captioner by Polina Tsvilodub as the S0 base caption

In Nie et al. (2020), the base caption system (S0) GVE-LRCN (link: https://github.com/salaniz/pytorch-gve-lrcn) is specifically based and trained on CUB and MS COCO dataset. Since we are using A3DS dataset, it might be unsuitable for the dataset we are using. In our first attempt, we adapted the decoder RNN model with the pre-trained weight provided by Polina Tsvilodub. In order to fit the A3DS data into the source code by Nie et al. (2020), we also tried to separate the data into batches by attributes and randomly as the original code did (at least we thought this is what class BirdDistractorDataset intended to do), but we failed to build a runnable code by trying to change a lot on the final caption generate (RSA part of the original code) and evaluation step.

3.3 Our second attempt of using the GVE-LRCN model used in the original code

We still tried to fit the A3DS dataset directly to both the base caption model and the RSA model but the "file not found" error kept happening again. Since the original code has a complex structure and we did not change the original code drastically, so this attempt also failed.

4 Conclusion

We underestimated the possible complexities of the source code and overestimated our actual ability to cope with complex codes with little extra instruction, also the unexpected cases of outdated packages and unavailable links. After trying to comprehend the original code and different ways of building runnable codes, we still failed to achieve the original goal of re-implement and reproducing the result of the paper. We put our remaining code in our GitHub repository.

References

Pragmatic Issue-Sensitive Image Captioning. Nie, A. Cohn-Gordon, R., and Potts, C. (2020). arXiv preprint arXiv:2004.14451.https://arxiv.org/abs/2004.14451

3D Shape Dataset. Burgess, C. and Kim, H. (2018).https://github.com/deepmind/3d-shapes/tree/master