Multimodal Relational Reasoning for Visual Question Answering

Progress Report

Name: Benedict Aaron Tjandra

E-mail: bat34@cam.ac.uk

Supervisor: Catalina Cangea

Director of Studies: Prof. Simon Moore

Overseers: Prof. Marcelo Fiore, Prof. Amanda Prorok

The project is 1 week ahead of the project schedule. The core objective of the project is to implement and train the MuRel network such that it outperforms a simple concatenation baseline on the VQAv2 test-std dataset. This core objective has been completed. The MuRel network has been implemented and trained, its loss function minimised, and has beaten the simple concatenation baseline.

This, however, felt largely unfair because the MuRel network is allowed access to different features than the concatenation baseline, is allowed to fine-tune the GRU module in its training, and has the ability to perform a self-attention mechanism over its question features. To make comparison between the MuRel network and the baseline fairer, I implemented and trained 8 other baseline networks and have shown that the MuRel network outperformed them all. These baseline networks each was allowed to fine-tune the GRU module, used different aggregation mechanisms (Mean, Max, Min, Sum) to process the BUTD (Bottom-Up Top-Down) object features, and had the option of using a self-attention mechanism on the question features just like MuRel.

The paper that this project is based on claims that MuRel outperforms an attention network under the same training conditions. In order to verify this, I additionally implemented two attentional networks, each with different fusion mechanisms and discovered that MuRel outperformed them too. In total, I have implemented and trained 11 baseline networks as part of this project, of which MuRel outperformed.

The table below summarises the results of the different baselines and the MuRel network.

Method Name	Yes / No	Number	Other	Overall
Concat: Frozen GRU + Resnet	60.89	0.34	1.14	25.66
Concat: Agg Sum, Unfrozen GRU	61.81	22.8	7.11	31.43
Concat: Agg Sum, Q_att, Unfrozen GRU	61.36	28.62	7.03	31.87
Concat: Agg Max, Unfrozen GRU	63.96	17.94	12.35	34.25
Concat: Agg Max, Q_att, Unfrozen GRU	65.01	19.70	12.81	35.09
Concat: Agg Mean, Unfrozen GRU	65.29	24.68	14.65	36.65
Concat: Agg Mean, Q_att, Unfrozen GRU	65.58	27.63	14.85	37.2
Concat: Agg Min, Unfrozen GRU	68.51	32.31	30.61	46.42
Concat: Agg Min, Q_att, Unfrozen GRU	67.8	32.31	30.61	45.36
Attention, Concat	71.86	37.85	45.31	55.4
Attention, BLOCK	77.5	40.35	52.85	61.58
MuRel (best performing network)	77.53	43.93	52.44	**61.81**

The difficulties of this project was largely dictated by time and resource constraints. The MuRel network is large and takes up a lot of GPU space (~22GB). The Computational Biology GPU machines simply cannot accommodate this, and I was required to employ a

trick (gradient accumulation) which reduces the space the model occupies, but increases training time. Training time, in particular, is also substantial bottleneck as it takes around ~6 days to fully complete MuRel training. Therefore, it is not feasible to explore and fine-tune hyperparameters rigorously as this will take up too much time. The current hyperparameters of MuRel largely follows what was written down in the appendix of the original MuRel paper. As per the project proposal, the runtime of the baseline networks and MuRel will be compared, and the extensions of the MuRel network will be considered in the future.