Title	**Multimodal Learning and Reasoning for Visual Question Answering**
Student name:	**Vasin Srisupavanich**
Supervisor name:	**Dr Srinandan Dasmahapatra**

Aims/research question and Objectives

Visual Question Answering (VQA) is the task of answering question about an image. It is a multimodal task which requires both visual and language understanding. Making progress in this domain could have immense real-world impacts, such as aiding visually impaired individuals and enhancing human machine interaction. In addition, VQA is a great domain to measure the intelligence and reasoning capabilities of a machine, due to the variety of skills required to tackle different types of question.

The aim of this project is to develop a VQA model and evaluate how it performs on the visual reasoning dataset. The model should be capable of responding to relatively complex questions that require multi-steps reasoning and inference over real-world images. Main objectives are summarized as follows:

1. **Background research on existing deep learning models for VQA and visual reasoning task.**
   The first step is to explore the use of various deep learning models, which appeared to perform well on visual reasoning tasks, such as attention mechanism and memory augmented neural networks. In addition, deep learning models that are specifically design for reasoning tasks, such as MAC networks[1], needs to be investigated.

2. **Background research on joint representation of image content and natural language.**
   Convolutional features and BERT[2] (Pre-training of Deep Bidirectional Transformers for Language Understanding) are the current state of the art representation of image and text respectively. However, their representations are separated, which limited the usefulness and accuracy of downstream visual-linguistic tasks. Recent methods, such as VL-BERT[3], which learn joint representations of image content and natural language needs to be explored.

3. **Prepare and setup development workflow.**
   Deep learning models can take a significant amount of time to train, therefore fast GPU servers will be prepared and setup in advanced. To increase productivity, experiment tracking tools, such as *weight and biases* or TensorBoard will be utilized. In addition, version control and automated testing will be setup to ensure the efficiency of the project.

4. **Design the VQA model.**
   The model architecture will be designed to incorporate the joint representation of image and text with reasoning module. Also, hyperparameters, such as learning rate, suitable loss functions and optimization process will be defined in this step.

5. **Implement and train the VQA model**
   The model will be implemented using Python programming language and PyTorch deep learning library. GQA, a dataset for real-world visual reasoning and question answering, will be used to train and evaluate the model.

6. **Optimize and finetune the model**
   To improve the performance of the model, regularization methods, such as dropout and BatchNorm will be investigated and utilized in the training process. Also, different learning rates and optimization algorithms will be tested to improve the training efficiency.

7. **Evaluate the model on GQA dataset**
   The model will be submitted to EvalAI, an open source platform for evaluating and comparing different machine learning algorithms. Qualitative and quantitative analysis will be performed to identify the strengths and weaknesses of the model.

---

[1] Hudson, D.A. and Manning, C.D., 2018. Compositional attention networks for machine reasoning. arXiv preprint arXiv:1803.03067.
[2] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[3] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F. and Dai, J., 2019. Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530.

Summary of proposed research and analysis methodology

1. **Literature Review**
   Papers related to VQA and visual reasoning will be searched and reviewed through information sources, such as Google scholar, arxiv.org, arxiv-sanity.com and paperswithcode.com. In addition, to keep up with recent developments on this domain, papers relevant to this topic from top machine learning conferences, such as NIPS, ICLR and ICML will be reviewed.

2. **Data Collection**
   There are many recent VQA datasets available to use, however, only dataset which puts emphasis on reasoning skills will be considered. Specifically, GQA[4], a dataset for real-world visual reasoning will be used to train and develop the model. This dataset consists of 113K images and 22M questions, measuring different types of reasoning skills, such as object recognition, spatial reasoning, logical inference, and comparison.

3. **Model Development**
   The first step is to investigate and implement the baseline VQA model (LSTM, CNN) to get the idea of the overall process and set the benchmark to improve upon. Second, the best approach that jointly represent both questions and images will be explored and implemented. The next step is to implement the core neural networks model for reasoning process. Finally, the best approaches to integrate each module will be explored and iteratively improved.

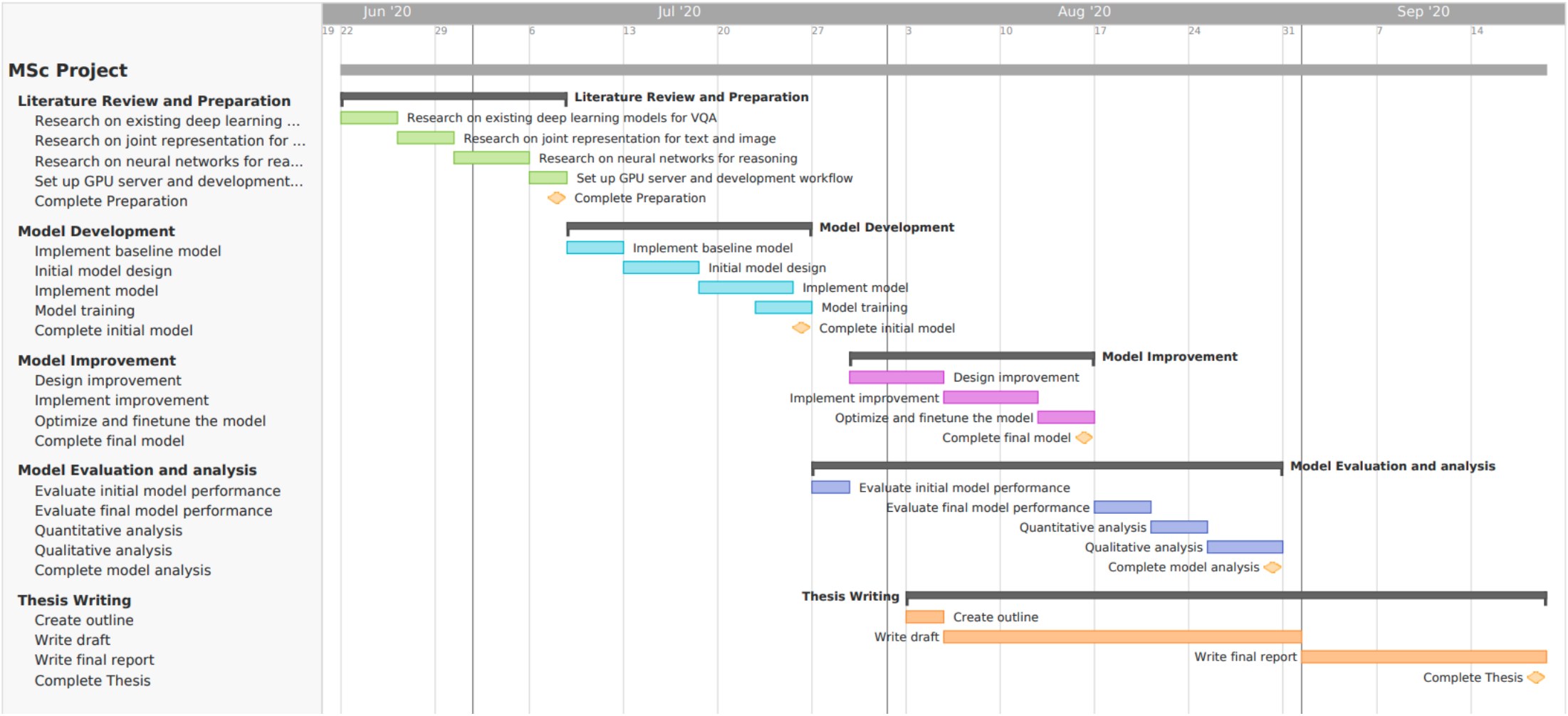
4. **Model Evaluation**
   Apart from the accuracy metric, there are several additional metrics, such as consistency, validity and plausibility metric, which can give insights to the performance of the model's reasoning capability. For instance, the validity metric evaluates whether the model gives valid answer (respond with color for color questions or yes/no for binary questions). For this project, these metrics will be used together to evaluate and compare different models.

5. **Qualitative Analysis**
   If time permits, qualitive analysis of the final model will be investigated. The matter of model's interpretability and explainablity will be discussed and visualized (e.g. visualization of attention mechanism). In addition, it will be useful to study whether the model can generalize to new dataset with different type of questions or images.

6. **Quantitative Analysis**
   The accuracy of each question types will be analyzed and compared with existing models. This will allow the strengths and weaknesses of the model to be identified. Moreover, an analysis on data efficiency should give useful insights on how fast the model learns.

[4] Hudson, D.A. and Manning, C.D., 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6700-6709).

Research plan – Gantt chart or Pert chart



Ethical statement

This project will use the GQA dataset, which contains 113k real-world images. These images are taken from COCO and Flickr, and contain many categories, including images of people and houses. Therefore, I will use this dataset with mindful of privacy and any other ethical concerns, as well as abiding to the term of use.

This project aims to develop an AI system which is capable of understanding both image and text data. Potential real-world applications resulting from this domain includes a system to aid visually impaired individuals, a system to detecting hateful contents, and virtual assistant. There are also applications which requires careful consideration before deploying, such as surveillance system. As with any other technologies, these systems can be used or developed for both good and bad depending on the user and creator. Therefore, I will only contribute to the development of AI systems, which are beneficial to the society.

Legal and commercial aspects

Commercial Aspects:
This project could potentially be developed to be part of many commercial applications e.g. a virtual assistant in digital photo album. However, the potential for commercial applications would still be very limited, since this domain is still in a very early stage of development.

Legal Aspects:
If this project were to be commercialized, then the issues of software licensing must be considered. Frameworks or libraries used in the project would need to be free, open-source, or allowed to be used as part of commercialized products.