MSc Project - Reflective Essay

Project Title:	Unsupervised multimodal machine translation using visual signals as rewards for reinforcement learning
Student Name:	**Gargeya Sharma**
Student Number:	**220278025**
Supervisor Name:	Julia Ive
Programme of Study:	MSc Artificial Intelligence

This essay is about embracing the journey to submitting my dissertation today, which no doubt was filled with challenges, but I must say “totally worth it”.

The journey began during a lecture of my module: artificial intelligence in the first semester, where I encountered Professor Julia Ive for the first time. Her expertise in Transformers and NLP intrigued me, given my background in computer vision and poor experience with her field of research. Eager to merge our interests, I approached her, expressing my desire to venture beyond my comfort zone. Julia's initial skepticism about aligning our research interests was met with my determination to learn, not merely to secure high grades. This proactive approach secured me a position as one of the first students to obtain a dissertation mentor in the EECS department of QMUL.

My passion for artificial intelligence is evident in my self-taught journey during my bachelor's degree in computer science, specializing in cybersecurity and forensics. With Julia's guidance, I delved into NLP and reinforcement learning, thoroughly for the first time, culminating in this dissertation. Recognizing my keen interest in computer vision, Julia ensured its integration into our research topic. Our discussions led to the decision to explore unsupervised machine translation using reinforcement learning, with the unique twist of incorporating multimodal visual signals as rewards.

Deep Dive into the Research Phase

The research phase of this project was akin to navigating uncharted waters. To make this very clear in the beginning, *I am the first person to ever do something like this: Unsupervised multimodal rewards for machine translations through reinforcement learning*. This is a NOVEL approach. My initial excitement and work started with months of reading literature, different novel approaches for unsupervised machine translation, reinforcement learning with machine translation and all the relevant combinations of techniques utilized in this paper. I even did an entire course of reinforcement learning offered by UCL x DeepMind on YouTube and read the relevant sections related to policy gradient reinforcement learning in Manning publications, ‘Grokking Deep reinforcement learning’ book, which I found to be very helpful along with so many public articles circling the area. Platforms like Hugging Face, and GitHub became my anchor, offering a plethora of resources, datasets, and tools that were instrumental in guiding my efforts. Their vast repository and user-friendly interface allowed me to explore and experiment with various models and techniques.

However, the journey was not without its pitfalls. One blunder mistake from my end was the inadvertent use of two different models within the training loop. In practice, the model that was in training, instead of learning from its predictions, was relying on the rewards produced by translations received from another instance of the same model, which is obviously pre-trained but static in nature as it was not the model that was undergoing optimization. This stagnation was a stark reminder of the intricacies of reinforcement

learning and the importance of meticulous implementation. Despite my enthusiasm, the journey was fraught with challenges. From architectural to implementation issues, the learning curve was steep. Such challenges, while initially disheartening, became invaluable learning experiences. They pushed me to delve deeper, question my assumptions, and refine my approach.

Growth in all directions

My prior accomplishments in the realm of artificial intelligence, including published papers in esteemed journals, book chapters, and numerous blogs, had instilled in me a confidence that I could tackle any challenge. This time as always, utilizing my attitude without crossing the area of arrogance helped me delve deeper into the intricacies of reinforcement learning, particularly with unsupervised rewards in large language models and transformers, where I was met with unexpected hurdles.

Libraries that promised seamless integration, like RL4LMs, failed to deliver the expected results. Even when I tried integrating datasets like multi30k, the outcomes were far from satisfactory. These initial setbacks, coupled with the immense pressure I had placed on myself with this particularly advanced topic, began to erode my confidence. I questioned my capabilities and wondered if I had bitten off more than I could chew.

However, Julia's unwavering support was a beacon of hope during these trying times. She constantly reminded me of the groundbreaking nature of my endeavor. Her encouragement to view this project as a learning journey, rather than a mere end goal, rekindled my determination. She emphasized that even if the unsupervised approach didn't yield the desired results, the learnings and insights gained would be invaluable.

The crux of my challenges revolved around the reward function and the optimization of the Proximal Policy Optimization (PPO) trainer. My initial understanding of PPO suggested a straightforward approach, but the anomalies in loss and rewards during training sessions painted a different picture. For more than 10 experiments that I performed in the beginning, none of them finished training because of some or the other errors, the major issue was getting a Value Error stating that the logs that are generated after each step include nan values in them which can’t be projected on the weights and biases monitoring portal. This error didn’t leave me until the very end of this research. Another big reason for the training never finishing was my own decision to pull the plug on the ongoing training because I could monitor the performance at run-time. I could clearly observe that the model was definitely out of track by looking at the explosion of KL divergence, entropy, and a sharp decline in rewards, which at that point started resulting in translations that were nothing short of gibberish or just repeated words. It was a humbling experience, reminding me that even with a solid foundation, there's always more to learn.

Julia's insights were instrumental in reshaping my perspective. She highlighted the distinction between training a model from scratch and fine-tuning a pre-trained one. Realizing that I was inadvertently disrupting the parameters of a model already proficient in German-to-English translations, I adjusted my approach. By reducing the learning rate and increasing the batch size, I aimed to make subtle, effective changes rather than drastic ones. This again sets my research in the right direction at the right time given that I don’t have indefinite time to explore all my options. With these improved configurations, now I had other problems to solve.

Rewards for both me and the model

My journey was marked by numerous trials, so much so that I exhausted the storage of my Weights & Biases free trial. While reflecting on the reward function, I began to view rewards not just as mathematical entities but as intuitive tools, akin to how humans use rewards in real life. The reward function, which was central to the research, proved to be a particularly tough nut to crack. The integration of visual reasoning into the research process introduced a new layer of complexity. My initial exploration led me to various multimodal Transformers, such as ViLBERT, Visual BERT, and others. However, it was Hugging Face's documentation that drew my attention to the ViLT, a multimodal transformer designed for image and text retrieval. This seemed like the perfect fit for the research, given its capabilities and ease of implementation.

Yet, as with any novel approach, the incorporation of ViLT was not without its challenges. The pipeline for image and text retrieval, while promising on paper, presented numerous implementation issues. On the first go, I went with raw scores generated by ViLTforImageAndTextRetrieval as the rewards as they look just fine to be treated as rewards. However, they didn't provide as much reasonability into the pipeline as expected from this incorporation. If I had to control these rewards, I had to normalize them and then try playing around with them. As these raw scores were logit values, their range goes from -infinity to +infinity, so luckily due to various attempts at training, I observed the range of rewards that the model was receiving, they seemed to be between the range of ~-11 to ~+11, so I just forced a min-max clipping on these rewards with the custom range of -10 to +10. Now I have mathematical control over them to increase my ways of experimenting.

Although this shift (standardizing scores through min-max scaling) led to various experiments, each change in the attempt to improve the reward formulation brought its own set of disappointments, from errors to unexpected outputs. Don't forget the error that teased me throughout the time, ValueError: nan values.

In reflection, the journey with ViLT and visual reasoning was a roller-coaster of emotions (mostly bad ones), oscillating between excitement at the potential of the approach and frustration at the myriad challenges encountered. However, each hurdle, and each setback, served as a steppingstone, pushing me to innovate, adapt, and persevere. The experience taught me the importance of continuous learning and how to accept failure more easily than I used to (adaptability), helped me increase my patience to a higher level and instilled better resilience in the face of adversity. During all this, Julia's unwavering support was pivotal.

Setting a firm ground but also an unseen trap

Julia's suggestion to establish a supervised pipeline as a benchmark was a turning point. By demonstrating the efficacy of a supervised model, I could then provide a solid foundation for my comments on the unsupervised approach. This dual approach, while time-consuming, enriched my understanding of machine translation using reinforcement learning. To train the supervised model, I used the industry's best practice for machine translation, i.e., taking cumulative BLEU score for the token vice comparison between prediction translation and target sentence as a whole. The learning rate why high at the first go so the model was damaging itself midway but as an experiment what I did to solve this issue was disturbing and should be conveyed to avoid the same for anyone else. To dissect the counterintuitive directly proportional behaviors between the rewards

and the loss, I just return the negative supervised score to fix the proportional issue. This approach although seemed to work fine with the behavior of the charts, it resulted in something completely socially unethical: pornographic sequences of tokens in the resultant translations.

This served as a stark reminder of the unpredictable nature of AI and the importance of monitoring and refining models. Thankfully, with adjustments to the learning rate and a deeper understanding of the nuances of fine-tuning, I was able to rectify this.

In retrospect to what all I have done to finalize this dissertation, while my results might not be termed "groundbreaking," they represent a significant step in a larger journey. Research is a cumulative effort, built upon the contributions of many. I am proud to have added my insights to this collective knowledge pool. Given more time and resources, I would undoubtedly delve deeper, exploring the potential of multimodal transformers and their applications in 2023 and beyond.

In conclusion, this journey has been a testament to the power of perseverance, collaboration, and continuous learning. As we move towards a future where machines emulate human-like perception, integrating multiple sources of data for a holistic understanding, I am excited about the prospects of unsupervised multimodal machine learning. This dissertation, while a significant milestone, is just the beginning of a long and promising journey ahead.

P.S. I am attaching the link to my Weights and Biases portal showcasing the experiments that I performed (only those are left that were created after deleting previous ones due to low storage)

https://wandb.ai/gargeya/trl/table?workspace=user-grey8magic