Rokos Award Project Report
--Jianhao Yuan 08/10/22

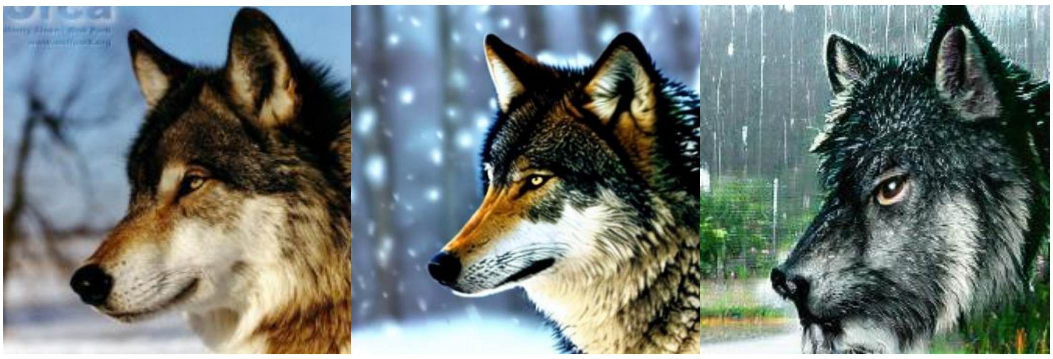
Exploiting The Untapped Potential of Using Text-Vision Generative Model in Domain Generalization

I spent a great summer in the Torr Vision Group, Engineering Department. My project focus on exploring a novel method to improve the generalizability of the current computer vision model. The recent success of deep learning models is largely built on the assumption that the training data and test data have the same distribution, which can sometimes be broken in real-life applications. For example, in autonomous driving, researchers sometimes seek to use a visual model to detect and understand the objects such as humans, cars, traffic lights, etc. for the vehicle to make decisions on driving strategy. To get a good performance, researchers train the model with a bunch of sample photos taken on the street with these objects being labelled. One interesting situation is it happened to be a nice sunny day when researchers went out and took those sample photos. However later, when they test the model on a rainy day, the model suddenly fails to detect most of the things it should have learned. It is the shift in the train and test image's illumination condition that has caused the degradation of model performance. Such annoying behaviours prevalently exist around all types of deep learning models. To tackle this problem, researchers spend a great effort in the area of Domain Generalization, which my project contributes to.

During my internship, my colleagues and I proposed a novel method that uses a large text-vision generative model to solve this problem. The intuition is very simple: if we can have all types of photos taken from sunny, rainy, snowy, foggy, etc. weather, we can train a perfect model, but the only problem is that collecting real-world data is usually very expensive. Instead, a cheaper and more efficient way of doing this is using a model to synthesize or mimic these effects so-called data augmentation. In the past, people used hand-crafted pipelines to manipulate images such as blurring, cropping, jittering, etc. to increase the general robustness of the models. However, these fixed image corruption methods do not guarantee to fully reflect the style shift between training and test data, and we want to introduce a more targeted and explainable way of doing these.

Recently, there have been many text-vision generative models coming out of the community. Those generative models are trained with millions of text-image pairs collected from the Internet and allow people to generate and manipulate images through text-prompting. By specifying the semantics that we want to change through natural language, we can have a precise and controllable way to edit an image. Our method simply uses a human-understandable sentence to instruct the generative model to do the style change for us. We have shown that our methods achieve higher performance than the current state-of-the-art methods through a range of testing datasets, and we have also provided throughout analysis of the limitations and further improvement for better usage of the generative model in domain generalization.

Here are some examples augmented by our methods with different types of generative model:



(a) Original Image (b) Augmented with Stable Diffusion (c) Augmented with VQGAN-CLIP

This research experience has been very enjoyable for me. Although it was not that smooth and we had to make some pivots in the project direction during the process, we finally got on the right track. We get a paper to submit to the conference just by the end of the project. As my first paper output in my life, I am super excited about it. It gives me a more concrete idea of how the research would look like and makes me more determined to have a DPhil to continue to investigate the area.