HEPHAESTUS

Context based outlier detection, interpretation through disentangled latent manifold (DLM) and exploration of DLM through generative models

CONTACT INFORMATION –

Name – Vedansh Sethi

Branch – Computer Science and Engineering

Institute ID – vedansh_s@cs.iitr.ac.in

Enrollment - 25114099

Contact No. – 7976952336

ABOUT ME –

I am a first year student in computer science branch of IITR, I became interested in AI/ML towards the end of my 12th standard when I studied the chapter of biomolecules.

I was really interested in the way proteins work, and are they mutable, that is when I was introduced to AlphaFold, it was the main thing that sparked my interest

Then after JEE advanced, I very briefly studied more about AI, where I got to know about AlphaGo documentary, and it was an eye opener, that in this age, all this is possible.

Now that I am in college, I feel that this is high time for me to get hands-on about this, and make some good projects myself, like in my TMI 102 course, I have worked on a project related to DL, where I used tandem network to use inverse physics problems

My current interest is in interpretability of AI and better use of generative models through knowledge gathered using interpretation methods

INTRODUCTION –

This project aims to tackle the problem of identifying outliers based on context. It ranks input pictures based on how different the picture is from the group using interpretability methods and use generative methods to explore the latent manifolds

- Data – I aim to use the shapes3D dataset to first create a pipeline and check the results of the next two stages, that is interpretability and generative part, then I will extend it to datasets with more real life like features, namely CIFAR 100
- Pre-processing – Reducing the pixel features magnitude from 0-255 to 0-1 to prevent exploding gradient, we use Random Horizontal Shift to make the model insensitive to horizontal orientation
- Rough idea of architecture –
    - I plan to use an architecture like that of **beta-TCVAEs** which are specifically focussed on disentanglement of the latent manifold
    - For quantifying dissimilarity I use two measures, for initial ranking, I use **distance** of the images' encoded representation from the mean of the encoded representations and using **latent dimension-wise distance** to quantify the most differing feature among the mean and most dissimilar
    - For generative models, we can use the default decoder of the beta TCVAE or DDPM models by the advent of Diffusion aided VAEs for sharper images

PRIMARY TARGET –

- Train a model to predict outliers in group of pictures based on the context of the pictures given
- Using interpretability methods as described above, identify different aspects in which the outlier is different from the group.
- Generative model trained to make pictures similar to the group of inputs but different in parameters controlled as input by the virtue of disentanglement

SECONDARY TARGET –

- Use generative models to generate pictures from most similar to the group to most different from the input group

EXAMPLE OF PROJECT'S WORKING –

- We take the example of shapes3D dataset, it has 6 ground truth factors
- If we take 5 pictures matching on 3-4 ground truth factors and in the same group, we have a picture which matches on only 1 ground truth factor on average with each picture, then that image will be an outlier
- We will use interpretability methods to see what features of the image made the model make the decision to rank the group this way, it includes saliency maps and interpretation through the help of disentangled latent space, telling exactly which features made the model confident
- We use the generative models to generate images ranging from most like the group to least similar, and to generate images which are like the group but can be varied in different features in a controlled manner, again because of disentangled latent space

TIMELINE –

- **Week 1** –
    - Research more about model architecture, finding better interpretability methods and diffusion variants
- **Week 2** –
    - Do basic training on shapes3D dataset, applying the interpretability part of the project
    - Start training of the generative model that generates images similar to the group but different in controlled aspects
- **Week 3** –
    - Prepare a report of overall project's performance
    - Do the same on CIFAR 100 database
- **Week 4** –
    - Train the second generative model that generates the gradient of images from most similar to most different
    - Apply the same to CIFAR 100 model too

REFERENCES –

- **VAEs** –
  - Introduction to VAE
  - original VAE paper
  - more detailed discussion on VAEs
- **Variants of VAEs focussing on disentanglement** –
  - introduction to disentanglement with beta-VAEs
  - original beta-VAE paper
  - understanding disentangling in beta-VAE
  - isolating sources of disentanglement in beta-VAEs
- **Interpretability and generative methods to explore latent space** –
  - introduction to XAI
  - introduction to Grad-CAM
  - examining interpretable disentangled representations
  - How diffusion models work - Deepbean
  - DDPM and DDIM by 3B1B and Welch labs
  - Original diffusion model paper
  - original DDPM paper
  - paper on diffusion VAEs