COM322: Computer Vision
Preliminary Research Project Topic and Scope

Derin Gezgin

Image classification is a widely-discovered area in computer vision. In this project, I plan to create an image classification pipeline for the classification of food images. While there are many options for my classification pipeline, such as a (Deep) Convolutional Neural Network (CNN/DCNN), Support Vector Machine (SVM), Decision tree, and K-Neerest Neighbor (KNN), I am planning to use a vision transformer model in this project. The vision transformers is a new image classification method recently introduced in 2020 [1].

There are several datasets (which I will explain in more detail in my Project Description paper) for food images that are publicly available for me to use. Similarly, considering the time and computing power constraints and lack of a significantly large dataset, I am planning to fine-tune the weights of an existing vision transformer model to see how it will perform in a different classification task. At the same time, I plan to do the same process with a DCNN/CNN to see how it performs in fine-tuning compared to a vision transformer. Similarly, I plan to experiment (depending on the time) with how a model trained in a specific cuisine will classify a different cuisine. Lastly, even though it is doubtful that I will have time for this, I plan to experiment with different vision transformer models like the Swin-Transformer [2], Deep Vision Transformer (DeepViT) [3], etc.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020.

[2] Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," *CoRR*, vol. abs/2103.14030, 2021.

[3] D. Zhou, B. Kang, X. Jin, *et al.*, "Deepvit: Towards deeper vision transformer," *CoRR*, vol. abs/2103.11886, 2021.