COM322: Computer Vision
Project Description

Derin Gezgin

Problem Definition

My project mainly focuses on *food image classification*, which is a widely-explored area [1]. As I am specifically interested in vision transformers and their variations, my primary focus in this project would be to create a vision transformer model to classify food images. At the same time, I will explore some variations of this problem, which I explain more in the *Methodology* and *Goals* section.

Methodology

As with most computer vision projects, I have two main libraries I can use: PyTorch [2] and TensorFlow [3]. I will follow this process during implementation of my project:

1. **Reading the dataset from the directory.**

   Considering our datasets are already publicly available and properly formatted, we can use the built-in functions from TensorFlow (`image_dataset_from_directory`) and PyTorch (`DataLoader`). These functions will automatically assign the folder numbers as labels to the images. As we have a label to food-name file for most of the datasets, we can crosscheck them later if we need to know the exact names of the food items.

2. **Pre-Processing and Data Augmentation**

   As an optional step, I can apply basic data augmentation to images. At the same time, depending on the model I use, I might have to pre-process the images to make them suitable for the input to my network.

3. **Fine-Tuning Process**

   For the fine-tuning section of my project, there are options in both PyTorch (`torch.hub.load`) and TensorFlow (`tensorflow_hub.KerasLayer`) have built-in model weights that I can directly load. Both libraries have options that recreate the vision transformer model architectures from cutting-edge papers [4][5][6]. Moreover, I can easily experiment with other models like ResNet [7], GoogLeNet [8], etc. In the fine-tuning step, I will keep the general model architecture while re-training the last (or last 3) fully connected layers. This will reduce the need for computing resources and training time.

4. **Testing and Evaluating the Trained Model**

   After the fine-tuning (training) of the network, we can test it as we test a regular model and check the evaluation criteria that I explained in the *Evaluation* section. Both TensorFlow (`TopKCategoricalAccuracy`) and PyTorch (there is no direct built-in function for this) have options for this type of accuracy measure.

Datasets on Food Images

There are many publicly available datasets for me to use in my research project. These are some of the well-known examples for the food-image datasets:

Dataset Name	Class Count	Total Image	Cuisine
Food2k [9]	2000	1,036,564	Mixed
ISIA Food-500 [10]	500	399,726	Mixed
Food524DB [11]	524	247,636	Mixed
CNFOOD-241 [12]	241	191,786	Chinese
ChineseFoodNet [13]	207	185,628	Chinese
FoodX-251 [14]	251	158,846	Mixed
ETHZ Food-101 [15]	101	101,000	Mixed
FFoCat [16]	156	58,962	Mixed
UEC Food-100 [17]	256	31,397	Mixed
Food-11	11	16,643	Mixed
UEC Food-256 [18]	100	9,060	Mixed
KenyanFood13 [19]	13	8,174	Kenyan
FoodSeg103	104	7,118	Mixed
Indian Food Image Dataset	N/A	5,000	Indian
Food-5k	2	5,000	Mixed

Table 1: Some of the available food image datasets

In my research project, depending on the exact computational requirements of my transfer-learning process, I plan to use a medium-sized dataset like CNFOOD-241 or FoodX-251. During the building part of my project, I have to consider having enough data to train a vision transformer (It is known that vision transformers are data hungry [20][21][4]) and also do not run out of computing resources.

Goals

In my final project, I have several sub-goals for my general goal of *food image classification*. I will start from goal 1. in my implementation and try my best to complete all my goals. I am a little over-shooting in this part to ensure I have something to do in case I finish everything early.

1. Food image classification of a specific dataset by fine-tuning the weights of a vision transformer.
2. Food image classification of a specific dataset by fine-tuning the weights of a (Deep) Convolutional Neural Network, or any network structure other than a vision transformer variation. This is to experiment with how significantly different architectures perform on the same task.
3. After I train/fine-tune a Vision Transformer model or a (D)CNN, I plan to test the trained classification model with a different dataset, preferably a dataset on a different cuisine. The main issue is when I train a model with a specific dataset, the model structure is specific for a certain number of output classes. I can fine-tune the model trained specifically on a specific cuisine to adapt it to another cuisine. I do not think that using the same model directly will work.
4. Finally, I am planning to experiment with different types of vision transformer architectures such as, CvT (Convolution for Vision Transformers [22]), DeepViT [23], Swin Transformer [5], and DeiT [24]. This final part of my goals depends on how much time I have left until the end of the semester and how well I can implement these architectures, as some of them are newly introduced and do not have proper resources for implementation.

Evaluation

For the evaluation part of my project, my initial goal is to have a better result than a pure random guess. In this case, it is Test Accuracy $> \frac{1}{\text{Class Count}}$. While this is a low bar for an evaluation, I think it is a nice starting point. After passing this bar, I aim to have as high accuracy as possible.

I am planning to have three different measures as my accuracy measure to have a better understanding of how my model performs:

- *True/False Accuracy:* In this case, it is the basic accuracy measure of checking whether our top guess is correct or wrong.
- *Top-3 Accuracy:* In this case, checking if the correct answer is among one of the top-3 guesses.
- *Top-5 Accuracy:* Datasets I have has a class count ranging from 2 to 2000. In datasets with many classes (More than 100), I plan to also look for the top-5 accuracy score.
- *Top-X Accuracy:* As I said in the previous bullet, I found and might use some datasets with many classes. I might increase the threshold (over 5) for accuracy to evaluate my work better. In this case, my limit is from %5 to %10 of the class count as the X value.

References

[1] Y. Zhang, L. Deng, H. Zhu, *et al.*, "Deep learning in food category recognition," *Information Fusion*, vol. 98, p. 101 859, 2023.

[2] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 8024–8035.

[3] Martín Abadi, Ashish Agarwal, Paul Barham, *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. [Online]. Available: `https://www.tensorflow.org/`.

[4] 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.

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

[6] Z. Tu, H. Talebi, H. Zhang, *et al.*, *Maxvit: Multi-axis vision transformer*, 2022.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[8] C. Szegedy, W. Liu, Y. Jia, *et al.*, *Going deeper with convolutions*, 2014.

[9] W. Min, Z. Wang, Y. Liu, *et al.*, "Large scale visual food recognition," *CoRR*, vol. abs/2103.16107, 2021.

[10] W. Min, L. Liu, Z. Wang, *et al.*, "ISIA food-500: A dataset for large-scale food recognition via stacked global-local attention network," *CoRR*, vol. abs/2008.05655, 2020.

[11] G. Ciocca, P. Napoletano, and R. Schettini, "Learning cnn-based features for retrieval of food images," in *New Trends in Image Analysis and Processing – ICIAP 2017*, S. Battiato, G. M. Farinella, M. Leo, and G. Gallo, Eds., 2017, pp. 426–434.

[12] C.-S. Chen, G.-Y. Chen, D. Zhou, D. Jiang, and D.-S. Chen, "Res-vmamba: Fine-grained food category visual classification using selective state space models with deep residual learning," *arXiv preprint arXiv:2402.15761*, 2024.
[13] X. Chen, H. Zhou, and L. Diao, "Chinesefoodnet: A large-scale image dataset for chinese food recognition," *CoRR*, vol. abs/1705.02743, 2017.
[14] P. Kaur, K. Sikka, W. Wang, S. J. Belongie, and A. Divakaran, "Foodx-251: A dataset for fine-grained food classification," *CoRR*, vol. abs/1907.06167, 2019.
[15] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – mining discriminative components with random forests," in *European Conference on Computer Vision*, 2014.
[16] I. Donadello and M. Dragoni, "Ontology-driven food category classification in images," in *ICIAP (2)*, ser. Lecture Notes in Computer Science, vol. 11752, Springer, 2019, pp. 607–617.
[17] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," in *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, 2012.
[18] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in *Proc. of ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, 2014.
[19] K. Wang, M. Jalal, S. Jefferson, Y. Zheng, E. O. Nsoesie, and M. Betke, "Scraping social media photos posted in kenya and elsewhere to detect and analyze food types," *CoRR*, vol. abs/1909.00134, 2019.
[20] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *CoRR*, vol. abs/2012.12877, 2020.
[21] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," *CoRR*, vol. abs/2102.10662, 2021.
[22] H. Wu, B. Xiao, N. Codella, *et al.*, "Cvt: Introducing convolutions to vision transformers," *CoRR*, vol. abs/2103.15808, 2021.
[23] D. Zhou, B. Kang, X. Jin, *et al.*, "Deepvit: Towards deeper vision transformer," *CoRR*, vol. abs/2103.11886, 2021.
[24] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *CoRR*, vol. abs/2012.12877, 2020.