CSE 519 -- Data Science (Fall 2024)
Prof. Steven Skiena
Homework 3: Data Science in the Era of Foundation Models
Due: Thursday, November 7, 2024

This homework will investigate several aspects of data science in the new era of LLMs and other foundation models that perform vision and coding tasks. Modern tools like copilot promise to improve the productivity of software developers—although I think I see the beginnings of a generation too quick to turn off their brains and outsource too much to the machines. They also promise to deskill NLP and computer vision tasks, by reducing them to simple calls to pretrained foundation models. We will explore these situations in this homework assignment.

The Google colab notebook (Use your @cs.stonybrook.edu email to access the file. If you do not have one, use the form that comes up to request access) contains boilerplate code to download the data to your google drive and a dictionary containing the features along with its data type. **Make a copy of the notebook before you start your HW.**

Tasks (100 pts)

1. Download the CIFAR-10 dataset, which contains 60,000 32x32 color images across 10 different classes, including airplanes, cars, birds, cats, and others. You can download it using the torchvision.datasets.CIFAR_10 module. Load the dataset and split it into training (50,000 images) and testing (10,000 images) sets. (5 points)
2. To perform analysis on these images, use ResNet-18 (torchvision.models.resnet18), a well-known convolutional neural network (CNN) designed for image classification. Remove the final classification layer of ResNet18 (since it's trained for ImageNet) and use the rest of the network to extract embeddings (features) from the test set. (5 points)
3. After extracting the embeddings, apply the t-SNE algorithm (sklearn.manifold.TSNE), which reduces the high-dimensional data (embeddings) into 2D for visualization. The resulting plot will show how well the model clusters similar images. How well do the embeddings reflect the structure of the classes? (5 points)
4. Nearest neighbor classification: Construct the centroid of each object class in embedding space from the training data. Now perform object classification by the label of the nearest neighbor among the 10 centroids.
    a. Experiment with both cosine similarity and Euclidean distance. For each object construct the confusion matrix and evaluate how well the nearest neighbor classification works. (5 points)
    b. For each of the ten object types, identify which image is furthest from its centroid, ie. is an outlier to its class? Are these difficult cases? Are they mislabeled by mean closest to another centroid? (5 points)

5. Image Classification from a model: Build a classification model that uses the top level embedding as features to classify the objects. This could be logistic regression, random forests, or other methods.
    a. As above, build the confusion matrix and evaluate how well it works. (5 points)
    b. Compare the performance of nearest-neighbor and model based classification: do they make similar mistakes or not? Which approach does best? (5 points)
6. Experiment with dimension reduction on these embeddings, using PCA or SVD. Cut them down to 10 dimensions and 50 dimensions, and compare the performance on classification and TSNE to the full embeddings. Do we lose much performance? Or might it even get better? Explain your findings. (10 points)
7. Download the AG News dataset using the Hugging Face datasets library, which contains news articles across four categories: World, Sports, Business, and Science/Technology. Use the pre-trained DistilBERT model from Hugging Face to extract embeddings from the dataset by removing the final classification layer. **Repeat steps 1-6 for this NLP problem using the AG News dataset and DistilBERT embeddings**. Train and test splits are already provided in this dataset. You may sample the dataset according to your available computation resources. (45 points split as above)
8. Use your favorite code-generation system (ChatGPT or CoPilot or Google Gemini) and ask it to solve this homework assignment. Check carefully what it is doing, and report on your experiences:
    a. Evaluate the usefulness of your code-generation system on this assignment. What does it get right and what does it fail on?
    b. Did you discover any subtle mistakes when you read the resulting code? Do you get trapped into spending more time debugging than you thought you would?
    c. Do you think using AI tools for simple tasks frees up your brainpower for more advanced problem-solving, or does it reduce your overall understanding?

There is no wrong answer here: we are looking for depth of self-reflection, not fidelity to the party line. (10 points)

Rules of the Game

1. This assignment must be done **individually by each student**. It is not a group activity.
2. Get started on this as early as possible! As we have seen, it helps to start simple and then iterate to better solutions (KISS).
3. This project is designed to show how easy NLP and vision tasks can be when starting from foundation models. It is even easier if you have had an NLP or vision course before. If not, it is even more important that you start early here.
4. Some of these experiments may take substantial computation time if you are in a compute poor environment. If so, just downsample the data and proceed. Perhaps start by downsampling to a small reasonable amount and then upsample until you get tired of waiting.

5. I will soon distribute semester project descriptions, and the period of this assignment will overlap that of the time you are preparing your project proposals. Be prepared to work on both simultaneously.
6. All of your written responses should be put in the appropriate place in your notebook template. Get the template notebook form from here. *Use your @cs.stonybrook.edu email to access the file. If you do not have one, use the form that comes up to request access.*
   You are allowed to add more cells, but definitely fill out the cells we give.
7. I intend to have covered all relevant material before the assignment is due. But I look forward to discussing things in class and on Piazza, so please ask. And feel free to read ahead in the book and/or lecture slides.

Submission

Submit everything through Google classroom. As mentioned above, you will need to upload:

1. The Jupyter notebook all your work is in (.ipynb file), derived from the provided template
2. PDF (export the notebook as a pdf file)

These files should be named with the following format, where the italicized parts should be replaced with the corresponding values:

1. cse519_hw3_*lastname_firstname_sbuid*.ipynb
2. cse519_hw3_*lastname_firstname_sbuid*.pdf