UNIVERSITY OF SOUTHERN DENMARK
FACULTY OF SCIENCE

DS807: Applied Machine Learning

Winter 23/24 (ordinary exam)

1. Assigner: Christian Møller Dahl.
2. Hand-out: December 18th, 2023, 12:00 (noon).
3. Hand-in: January 31st, 2024, 12:00 (noon).
4. All pages, incl. the front page, should contain the following: Full name and SDU username (**not** CPR-number).
5. All pages must be numbered.

Form of examination for the certificate:
Take-home assignment.

Supplementary information for the form of the exam:
The exam may be solved in groups of up to 5 students or individually. Working in groups is encouraged. You should make the definition of your group in System DE-Digital Exam before the start of the exam. Follow this guideline.

Further:

1. In your report, be sure to state explicitly who is responsible for which parts to facilitate individual assessment.
2. Location: Home assignment.
3. Internet access: Necessary.
4. Hand-out: System DE-Digital Exam.
5. Hand-in: System DE-Digital Exam.
6. Extent: No longer than 30 pages, excluding references, appendices, and code.
7. Exam aids: All exam aids are allowed.
8. File format: The report must be submitted as a **.pdf** file. The code may be submitted as one of: **.ipynb**, **.html**, or **.py**.

Grading according to the Danish 7-point scale. Grading based on the performance of the individual student compared to the learning goals.

Exam questions

For all questions in the exam, be sure to state how and why you prepare the data, including considerations for how to further split the data, scale the data, reshape the data etc.[1] In particular, it might be necessary to limit data and models due to the hardware limitations on your system. This is perfectly fine, but please be sure to discuss the potential implications and consequences. In general, if you use the same method for multiple questions, it is sufficient to describe the procedure once and refer to it in subsequent questions (however, it must still be motivated).

Problem 1:

During the semester you have become very excited about working on the PatchCamelyon (PCam) data. Like Veeling et al (2018), you are primarily interested in developing machine learning models that, based on patches of whole-slide images of lymph node sections, can assist pathologists in tumor detection.

The primary objective of this exam is to perform image classification on the PCam dataset, with a focus on using autoencoders for image compression.

As a reminder, the PCAM dataset consists of 327,680 color images (96x96pxs) extracted from histopathologic scans of lymph node sections. Each image is annotated with a binary label indicating presence of metastatic tissue. Specifically, the dataset is divided into 262,144 training images, 32,768 validation images, and 32,768 test images. You must use the training/validation images to train/validate models that perform well at classifying the test images.

Importantly, you are not required to use the full dataset. Use the amount of data that is feasible for you and your hardware configuration. The PCam dataset is available from many online sources but some of them are very slow. Therefore, I recommend accessing the dataset from my SDU repository: Link to PCAM. Further instructions and hints on how to load the data efficiently and flexibly by using the module tfds|Tensorflow Dataset are available on the course site's itslearning platform.

Questions

You wish to perform image classification. However, you are a little concerned about the size of the PCAM images. Their size might be too demanding for your hardware configuration, potentially hindering proper hyperparameter tuning and downstream model evaluation. Hence, you wish as a pre-processing step to compress the images by using autoencoders (both AEs and VAEs).

1. Discuss how you can use autoencoders to compress images. Motivate why you think compressing the images by using AE/VAE will work here and will be better than just simple resizing for example by using the function:

[1] Even if you do not perform one or more of these steps, motivate why you choose not to.

```python
def resize_and_normalize_image(image):
    image = tf.image.resize(image,[32,32])
    image = tf.image.rgb_to_grayscale(image)
    return image / 255.0
```

Present your full AE/VAE based compression pipeline.

2. Build and train autoencoders that efficiently compress the 96x96pxs color images. As a minimum train one AE model and one VAE model.
3. Based on the compressed images from part 2. train CNN models for image classification. Train at least one CNN model using AE compressed images and at least one CNN model using VAE compressed images. Compare the performance to a CNN trained on the resized images (using a function like resize_and_normalize_image())

Importantly, for parts 2. and 3. above, you must explicitly:

a. Consider and discuss alternative AE/VAE/CNN-model architectures. Motivate your baseline model architectures.
b. Discuss different optimization methods and motivate your final choice. Use visualizations to show the relative performance of the optimizers.
c. Visualize how regularization (such as dropout, weight regularization, or early stopping) impacts the training of your model. Here, be sure to visualize plots of train and validation losses and accuracies both with and without the use of regularization. Discuss regularization and its relation to overfitting.
d. Discuss and apply transfer learning. Motivate what type of transfer learning you use and how you apply it, including considerations for how to prepare the data for this. Here, be sure to visualize plots of train and validation losses and accuracies.

4. Having considered all the "experiments" above (a.-d.), select your preferred models (motivate why they are your preferred models) i.e., present the results for the preferred trained AE/VAE/CNN models. Calculate and report their performance on the test data and present the result in a table for easy comparison. Based on your results, also discuss the effectiveness of using AE/VAE for image compression as a pre-processing step compared to simple resizing methods, such as using a function like resize_and_normalize_image().

Problem 2:

In the rapidly evolving field of Natural Language Processing (NLP), the development of robust models for text classification has become increasingly important. These models have a wide range of applications, from sentiment analysis to automated content generation.

You aim to advance your expertise in this domain by working with the comprehensive and diverse Amazon Review Data (2018), accessible here.

This dataset, an updated version of the previously released Amazon review datasets, offers a rich compilation of reviews, product metadata, and transaction metadata. The sheer volume and variety of the data make it an ideal candidate for developing and testing advanced NLP models.

Your focus is on the "Luxury Beauty" product category, aiming to develop and train models for text classification, i.e., models that predict customer review ratings from 1 to 5, based on textual feedback (e.g., reviews and review summaries). However, to enhance the robustness and generalizability of your models, you have opted to exclude this category from your training and validation datasets. This decision simulates a real-world scenario where a model encounters data not seen during its training and validation phases. This approach will allow you to rigorously test the model's ability to generalize and perform on unseen data.

Therefore, the "Luxury Beauty" product category will solely be used for testing purposes. Data from all other product categories within the Amazon Review Data (2018) can “freely” be utilized for the training and validation of your models. You believe this strategy will enable you to explore the models' effectiveness in text classification tasks across diverse domains, while also providing a focused evaluation on the "Luxury Beauty" category.

Questions and suggested tasks:

1. Training and Validation Data Selection and Exploratory Data Analysis:
   Question: Describe the product categories you have chosen for your training and validation dataset. How many observations from each product category have you included?
   Suggested Tasks: Conduct an exploratory data analysis (EDA) to justify the selection of your training and validation datasets. Your EDA should feature visualizations illustrating, for example, the distribution of review ratings across different product categories, the length of textual reviews, along with their most frequent words and/or characters, the length of review summaries, including their prevalent words and/or characters etc. Provide a rationale for your choices of training and validation data and discuss any patterns or insights you observe from the EDA.

2. Review Rating Categories and Rationale:
   Question: How many categories, up to five, are you planning to include in your model for classifying review ratings?
   Suggested Tasks: Explain your decision regarding the number of review rating categories (outcome classes). Discuss the factors influencing your choice and why you believe this is the optimal number of review rating categories.

3. Baseline Model for Review Rating Classification:
   Question: As a baseline for your review rating classification task, you are required to build and train a shallow learning model. Which shallow learning model have you chosen for this task, and how well does it perform?
   Suggested Tasks: Discuss your chosen shallow learner and the preprocessing steps you applied to the textual data before training the model. Describe the process and reasoning behind your preprocessing choices. After training the model, calculate and report its performance on both the training, validation, and test data. Present key metrics such as

accuracy, precision, recall etc. Discuss any insights or observations you can draw from the model's performance.

4. RNN Model for Review Rating Classification:
   Question: For the review rating classification task, you are now required to develop and train models using Recurrent Neural Networks (RNNs). Why are RNNs suitable for this task, and how do they perform?
   Suggested Tasks:
   a. Model Development: Describe the architecture of your RNN model(s). Include details such as the type of RNN cells used (e.g., LSTM, GRU), the number of layers, and any other relevant architectural choices.
   b. Data Preparation: Explain how you prepared the textual data specifically for the RNN model(s). If any, discuss additional preprocessing steps or transformations that were necessary to make the data suitable for RNN training. Include details on vocabulary size, sequence length and embedding dimensions.
   c. Training Process: Outline the process you followed to train the RNN model(s). Include details on the training parameters such as batch size, number of epochs, choice of optimizer, the presence of dropout and other regularization techniques. As in **Problem 1** visualize how all these factors affect training and validation losses and accuracies.
   d. Performance Evaluation: Evaluate the performance of your RNN model(s) on both the training, validation, and test data. Present key metrics such as accuracy, precision, recall.
   e. Insights and Analysis: Provide an analysis of the RNN model's performance. Discuss any challenges you faced while training the RNN and how they were addressed. Reflect on the suitability of RNNs for the text classification task based on your results.

5. Comparison of Shallow Learner and RNN Models:
   Question: How does the performance of Recurrent Neural Network (RNN) model(s) compare to that of a shallow learning model in review rating classification tasks? Additionally, would implementing an ensemble approach combining the models enhance overall performance?
   Suggested Tasks:
   a. Performance Comparison: Analyze and compare the performance of the RNN model(s) with the shallow learner model. Also consider an ensemble approach that combines these models. Focus again on the key metrics such as accuracy, precision, and recall.
   b. Strengths and Weaknesses: Discuss the strengths and weaknesses of each model. Consider aspects such as model complexity, training time, generalization to unseen data, and their ability to handle the nuances of language in the dataset.
   c. Model Suitability: Reflect on the suitability of each model for the review rating classification task. Consider the characteristics of the dataset and the specific requirements of the task when making this assessment.
   d. Insights and Conclusions: Provide insights gained from this comparative analysis. Discuss any surprising findings or confirmations of expected behavior. Conclude with your recommendations on which model is more appropriate for the review rating classification task in this context and why.

General hints for the exam

This section provides a list of 'best practices' for answering exams, applicable not only to this specific exam but to exams in general.

1. Be sure to explicitly answer everything that is asked. This sounds obvious, but you may have missed something! Be very critical here – read through the exam carefully and ensure you have answered every question, considered every task, and discussed all necessary topics. Also carefully study the front page and its list of requirements!
2. Make your answers as short and precise as possible.
3. Stay on topic! You are welcome to discuss topics beyond what is explicitly asked, *provided it is relevant*. Do not start discussing unrelated topics! If a specific part of the curriculum is not asked for in an exam, it does not improve your exam if you start discussing it.
4. The objective of the exam is *not* to achieve the highest test performance, but to demonstrate your understanding of the various concepts you are asked to discuss and apply. I expect *reasonable* values (both with respect to parameters and performance), not optimal values!
    a. To expand, this means that while it is often a very good idea to search to some extent for the best parameters of your models, I am not interested in seeing a test of thousands of values.

Loading data (important!)

As mentioned above, I will provide hints and instructions on how to download and prepare the data. I have posted these on the course site's itslearning platform under the plan named: "**Take home exam: Instructions and hints**". It is important to review the notebooks provided.