UNIVERSITY OF SOUTHERN DENMARK
FACULTY OF SCIENCE

DS807: Applied Machine Learning

Winter 2022

1. Assigner: Christian Møller Dahl.
2. Hand-out: January 23, 2022, 12:00 (noon).
3. Hand-in: January 31, 2022, 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: https://mitsdu.dk/-/media/sdunet/filer/vaerktoejer/brugeradgang/digitaleksamen/digitaleksamen-stud/uk+vejledninger/uk_s02_guide_stud+gruppeaflevering.pdf.

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

During the semester you have developed a great interest in handwritten character recognition (HCR), and you have enjoyed working with problems related to the classification of the MNIST dataset. You have an appetite for more HCR and you have noticed that Kaggle recently posted the 'DIDA' dataset (https://www.kaggle.com/ayavariabdi/didadataset). Part of DIDA contains images with handwritten year-strings in the format CCDY, e.g., 1826, 1810, etc. You are excited about this discovery, and you decide to use the data available in DIDA to design and train models for the transcription of handwritten year-strings which you know potentially will be extremely useful to researchers working with historical data.

Your primary objective (of this exam) is to perform image classification of images with handwritten year-strings based on the DIDA dataset. For this purpose, and to simplify, you decide to work with three separate classification models, as illustrated in Figure 1, denoted the CC-D-Y modelling strategy:

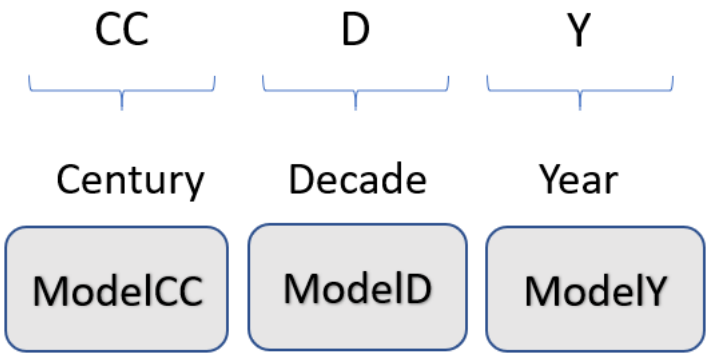

Figure 1: The CC-D-Y modelling strategy.

Specifically, instead of using one single model for predicting the entire year-string, you build and train one model denoted ModelCC to classify the century, one model denoted ModelD to classify the decade and one model denoted ModelY to classify the year. By combining the predictions/classifications from each of these three models you hope to be able to get at robust prediction/classification of images with handwritten year-strings.

For training/validation you start by using the part of the DIDA dataset denoted DIDA_12000_String_Digit_images with associated labels DIDA_12000_String_Digit_Labels.csv.[1] This part of the DIDA dataset consists of 12000 color images.[2] For the CC labels, to be used in ModelCC, you are defining 2 classes for the outcomes (18, not 18)=:(0,1). For the D and Y labels, to be used in ModelD and ModelY respectively, you are potentially interested in defining 11 classes, i.e., (0, 1, 2, 3, 4, 5, 6, 7, 9, 10) where 10 is the residual class. Please check that all 11 classes are represented in your training data and adjust if needed. Note that some images in the

[1] Downloading and organizing data for training/testing is part of the exam.
[2] Note that the raw images do not have identical/fixed dimensions. Imposing a fixed dimension can be done easily, for example, by a tf.image.resize function call in the image preprocessing/transformation steps.

dataset are not depicting year-strings and that some images only contain parts of a complete CC-D-Y year-string. These images might need some special attention when preparing your data and this is where the residual class 10 becomes handy.[3] All such irregularities are assigned to this residual class. Alternatively, you can delete images with irregularities from the training data but first consider if this is really necessary. Hints on how to prepare your data for the CC-D-Y modelling strategy are provided below in the "Hints on how to prepare the labels for the CC-D-Y modelling strategy" section below.

You can choose to split the pre-labeled data in DIDA_12000_String_Digit_images into training/validation and test, or you can choose to use all the pre-labelled images for training/validation and then evaluate the model on unseen/unlabeled images from the DIDA_30K_String_Digit_images part of the DIDA dataset. The latter option would require you to label a random subset of images in DIDA_30K_String_Digit_images. Importantly, you must use the training images of your choice to train models that perform well at predicting the test images of your choice.

When evaluating and comparing the accuracy of the different CC-D-Y models please consider using also the following metrics:

- "Sequence" accuracy: A prediction is recorded as correct (earning 1 point) only if all three sub models in the CC-D-Y model are making correct predictions.
- "Character" accuracy: A prediction is recorded as 1/3 correct (earning 1/3 point) if only one of the three sub models in the CC-D-Y model is making a correct prediction. Similarly, the prediction is recorded as 2/3 correct (earning 2/3 point) if two of the sub models in the CC-D-Y model are making correct predictions. If all three sub models are correct the prediction of the CC-D-Y model is recorded as correct (earning 1 point).

For all questions in the exam, be sure to state how and why you prepare the data, including considerations for how to split the data, scale the data, and reshape the data.[4] 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).

Question 1

Use non-deep learning to perform image classification according to the CC-D-Y modelling strategy. Specifically, you must:

1. Discuss how the problem can be solved using support vector machines, random forests, and boosting (discuss each method separately).
2. Use one of the methods above to solve the problem. A combination of two or all three of the methods may also be used, if you believe this is better (regardless of whether you use one or multiple methods, this must be motivated). Calculate and report the method's performance on the training, validation, and test data. Does the performance differ between the different sets? If yes, does this surprise you (explain why or why not)?

[3] The residual class can also be referred to as the "wildcard" class.

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

Question 2

Use deep learning to perform image classification according to the CC-D-Y modelling strategy. Specifically, you must:

1. Discuss why convolutional neural networks (CNNs) could be an appropriate type of model architecture to use for this task. [5]
2. Train a CNN to solve the problem. Here, you must explicitly:
    a. Discuss different optimization methods and motivate your choice.
    b. 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.
    c. Visualize how data augmentation 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 data augmentation. Discuss data augmentation 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.
3. Having run the experiments above, select your preferred model (motivate why it is your preferred model). Calculate and report its performance on the test data.

Question 3

To solve this problem, restrict attention to your preferred CNN from Question 2. The objective of this question is to get a better understanding of your model. Specifically, you must:

1. Discuss visualizations of activations, how to generate images that excite certain layers, and heatmaps of activations. Perform at least one of these using your preferred model CC-D-Y model. If your preferred model includes a spatial transformer network, consider visualizing spatial transformations.
2. Investigate if your model's performance is particularly good or bad at correctly classifying certain classes (i.e., it might be very good at correctly classifying centuries but not years, or it might be good at correctly classifying some decades but not certain other decades). Does it mix up certain classes? If yes, does this surprise you (explain why or why not)?

General hints for the exam

This section provides a list of "best practices" for answering exams – not specific to this one, but 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 carefully through the exam and be absolutely sure you have answered every question, discussed what needs discussing, and so forth. Also carefully study the front page and its list of requirements!

[5] If you disagree, you need to motivate why but still use CNNs to solve the subsequent questions.

2. Make your answers as short and precise as possible.
3. Stay on topic! You are welcome to discuss topics further than what is explicitly asked for, *if 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 get the highest test performance, but rather to show you have understood the different concepts you are asked to discuss and use. 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. If you at any point feel that you are unable to answer the exam due to limitations of computational resources, you are doing something wrong!

Hints on how to prepare the labels for the CC-D-Y modelling strategy

Below, a suggestion on how to create labels for each of the three models in the CC-D-Y strategy from the raw labels (**Label** column below) in DIDA_12000_String_Digit_Labels.csv.

Index	Label	CC	D	Y
1	1836	0	3	6
2	1836	0	3	6
3	1840	0	4	0
4	1840	0	4	0
5	1823	0	2	3
6	1823	0	2	3
7	59	1	5	9
8	61	1	6	1
9	62	1	6	2
10	63	1	6	3
11	65	1	6	5
12	1830	0	3	0
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
10661	1807	0	0	7
10662	1807	0	0	7
10663	192010	1	10	10
10664	111214	1	10	10

Note, however, that you do not have prepare the data this way, but in answering how you prepare the data, it is *not* sufficient to refer to “this is how it was done in the slides” – you need to show you understand *why* we prepare the data in a certain way.