DD2437 Online Exam Questions, 22VT (2022/03/12)

Question 1 (4p)

What are key differences and similarities (shared features) between transfer learning and multi-task learning approaches in deep learning? In what situations are they used and for what purpose?

Question 2 (9p)

You are requested to design a system for naming different species of birds shown in images. Those images were collected by botanists around the world who photographed selected birds in laboratory environment. As a result, a large number of bird photos accompanied with Latin names of bird species have been collected. Your task is to design a neural network based mobile app that can produce a name of bird species for photos taken by bird enthusiast out in the nature with their mobile phone cameras (of inferior quality). The botanists assume that they managed to account for more than 95% of all known bird species on the Earth. To simplify the task they created 48 categories of bird species and assigned them a new Latin name.

A) Assuming that the given 48 categories of bird species cover all known bird in the world (i.e. all bird can be assigned one of 48 Latin names), please design a neural network based app, dedicated for a mobile phone, that maps bird images to the corresponding names, and can generate the name for birds photographed by bird enthusiasts. Describe the proposed architecture (including key dimensionalities), name learning algorithms, identify relevant hyperparameters and provide some key dimensionalities. What type of problem is it? Please describe also critical challenges and constraints you recognise and suggest how you could deal with them or minimise their negative impact.

B) How does the problem change if we drop the assumption about the availability of 48 bird species categories and instead recognise that either bird could be assigned to more than a single category or there could be bird that does not fit the existing categorisation of available images (this could be a problem for those bird species that have never been photographed)? Would you extend the neural network architecture proposed to tackle the problem A so that you could also address the problem defined here (B)? Alternatively, would you propose a new neural network architecture for that task. How would you perform new more flexible groupings and deal with samples without relevant label (a sample without the corresponding representative with a name in the botanists' dataset). Describe your approach to B with analogous details as in A.

Question 3 (8p)

The interest to eat fish varies a lot between individuals, geographical regions and over the year (particular seasons or holidays associated with fish products, for instance). It also increases a lot following a TV-program about cooking whenever a nice fish dish is presented and then the interest can dramatically decline after media coverage of "negative factors" (over-fishing, water pollution etc). A wholesales retailer of fish products wants to get a better grip on the demand for fish over the next 3 days (one estimate for each of the days tomorrow, the day after tomorrow and the day after that). They will use this information to set prices when local shops and restaurants contact them for products. They have been in contact with your consultant company regarding the problem and your company has accumulated a lot of information for the last 6 months in 3692 locations. Daily sales of fish products (18 categories) have been obtained together with information about region, number of TV-programs about cooking the past 24 hours, number of media articles about "negative factors" the past 24 h, a total of 23 factors. After an initial analysis, your team concludes it is enough to look at 7 days of data (one week) since correlations between time-points seem to decay really quickly so looking at more days would essentially only add more "noise".

What type of problem does the company want to address? Accordingly, propose a neural network solution in your design assignment. In particular, motivate the choice of your network type, briefly characterise its topology clearly indicating the inputs and outputs, and describe how the network should be trained – how the data should be used and what learning algorithm you recommend. Please, explain also how you would optimise and estimate the generalization capacity. Finally, identify key challenges and potential difficulties/risks concerning the problem and your approach to effectively solving it.

Question 4 (9p)

As a data scientist you support a group of biologists that study samples of biological tissues infected by a new type of virus. The biologists take snapshots of their high-resolution microscopy images of a sample once per quarter for 24 hours in total. There are about 25 different types of tissue each represented by roughly 80-100 samples. In each image (snapshot) one can recognise characteristic quantitative features such the proportion of infected cells, changes in their dimensions etc. Following careful analysis the biologists managed to identify 5 key quantitative descriptors per image for about 90% of available samples. They wonder if there are patterns of changes in these descriptors for different tissue types that could develop over cycles of 3 hours, which they consider as a characteristic time constant. You are asked to support their heroic efforts in the following way:

A) automate the extraction of quantitative descriptors from images (there will be new volumes of images and labelling them by hand could prove infeasible),

B) identify, predict cyclic patterns (sequences) of changes in those descriptors at the given time scale of 3 hours,

C) determine if these sequences are typical of different tissues or alternatively if there are categories of tissue types that can be collectively characterised by the observed sequences of quantitative descriptors of virus inflicted changes,

D) as an extension of C) you are asked if it is feasible to actually predict those sequences potentially evolving on longer time scales than 3 hours.

E) build a prototype of a diagnostic tool that recognises a type of tissue (or a group of tissue types) attacked by the given virus based on the results of your work in C).

Please describe for each task (A-E) what neural network approach you would follow, how you would train the proposed network (how and what data you use, which learning algorithm(s) are involved) and apply it to provide meaningful results for your analysis. Please describe how you would reason about your results to provide answers that the biologists pose in different tasks.

Question 5 (7p)

Your company manufactures equipment for analysis of air and, in particular, for recognising the presence of solid particles (dust) in the air. Customers are government agencies and companies with the role of monitoring and reporting environmental factors, including cleanliness of air. The air processed by the equipment is fed through a camera-based microscope which takes images at high resolution. Using a particle detection and image-classification software, each particle detected is measured along a number of features (area, roundness, aspect ratio, solidity, etc, a total of 14 different features). The company has been running this system over the world in a multitude of locations (a total of 1375834 measurements have been stored). At each location, the average value of each of the 14 features and also the total particle count over 24 hours have been measured. For each location, the company has also stored a number of factors (human population density, percentage of the population employed by heavy industry, percentage of the population employed by light industry, etc a total of 28 socio-economic and geo-political factors, so-called SEGP-factors).

The company now has an idea for a spin-off using the data. The idea is to produce a system that can estimate the air particle quality given data on the 28 SEGP-factors. They will sell the system to agencies responsible for environmental factors so that they could produce their reports for

locations that have not been measured by the air analysis equipment but where the SEGP-factors are known. This would save a lot of money for the agencies to avoid collecting air measurement samples.

What type of problem does the company want to address? Accordingly, propose a neural network solution in your design assignment. In particular, motivate the choice of your network type, briefly characterise its topology clearly indicating the inputs and outputs, and describe how the network should be trained – how the data should be used and what learning algorithm you recommend. Please, explain also how you would optimise and estimate the generalization capacity. Finally, identify key challenges and potential difficulties/risks concerning the problem and your approach to effectively solving it.

Question 6 (9p)

You collaborate with a company manufacturing shock absorbers for trucks. The company has discovered that the lifetime of shock absorbers varies across trucks largely depending on the truck exploitation. In fact, the company has been asked to develop a diagnostic tool that will assess the condition of a shock absorber so that its repair or exchange can be well timed and, consequently, financial costs saved. On a closer inspection it turns out that there are 5 types of conditions ranging from healthy condition to failure. To make the assessment the company has developed a test based on induced vibrations where the frequency/spectral response of the absorber indicates its state. You are asked to automate the process of interpreting these spectral responses (a matrix of spectral components in 180 frequency bands and 100 time bins) and producing the diagnostic output – a condition of the sample shock absorber. However, it turns out there are some extra challenges ahead so you need to approach the problem in the following steps:

A) The available data is unbalanced. You have a few thousand samples without any annotations and then 5000 samples representing healthy condition, good condition – 1500, moderate condition – 3000, poor condition – 500, failure – 300. How would you approach this problem assuming that a diagnostic system should be trained on balanced classes, especially to identify poor condition and failure cases? How could we with the help of neural networks level out the number of class representatives?

B) If you manage to balance your classes you can start designing your diagnostic system. It turns out however that there is a sizeable proportion of noisy samples, particularly for the originally well represented cases. How would you deal with these noisy data?

C) Finally, having removed most of these disturbing obstacles you can develop and evaluate your model. Besides describing the underlying neural network architecture, please elaborate on how you validate it, select hyperparameter values, and perform final evaluation / measure performance.

Consider that you need to choose between two architectures of the same class of neural networks, how would you compare them?

D) Beyond the automated system for interpreting the shock absorber's spectral response and diagnosing the absorber's condition the company dreams of a neural network based predictor of the remaining lifetime for shoch absorbers. You are asked to advice the company on what type of data have to be collected to come closer to fulfilling their dream. At the same time, you are requested to identify key risks, problems that you envisage in this endeavour. Given that you can be offered the new data according to your requirements, how would you conceptually approach this prediction problem (no need for details, just name the family of neural network architectures that you would find useful, please motivate your answer)?

For the given steps and analyses (A-C) please propose and motivate a neural network approach – name and/or illustrate its general architecture, provide key dimensionalities (outputs, inputs, layers), propose a learning algorithm (without describing its operational details) and how you use the available data for training, model selection and generation of desirable results. If you recognise any *risks*, *constraints* that should be considered in your design, please clearly indicate them and, where applicable, suggest how they could be handled.