THE UNIVERSITY OF AUCKLAND

SEMESTER ONE 2023
Campus: City

COMPUTER SCIENCE

Advanced Topics in Machine Learning

(Time Allowed: TWO hours)

NOTE:

This exam is out of **100** marks. Attempt **ALL** questions.
Write your answers in the separate booklet provided to you.
Write as clearly as possible, and clearly indicate the question you are answering.
A list of reference papers discussed in class is provided on the last page.

Answer ALL questions.

Part 1: Deep Neural Networks [33 marks]

Question 1: Deep Neural Networks and generalisation [7 marks]

a) For the following situations, describe a simple solution and the intuition behind it:

   i. Your model has a high bias after training. [2 marks]

   ii. Your model has a low bias and a high variance after training. [2 marks]

b) Zhang et al.'s (2021) paper discussed in class, questioned traditional approaches for reasoning about generalisation. Through randomisation experiments, the authors show that large successful DNNs can easily fit random labels. What is the implication of this observation on DNNs effective capacity? [3 marks]

Question 2: Attention mechanism and Transformers [6 marks]

a) Explain in your own words what is multi-head attention. [3 marks]

b) Explain why multi-head attention is useful for Transformer models. [3 marks]

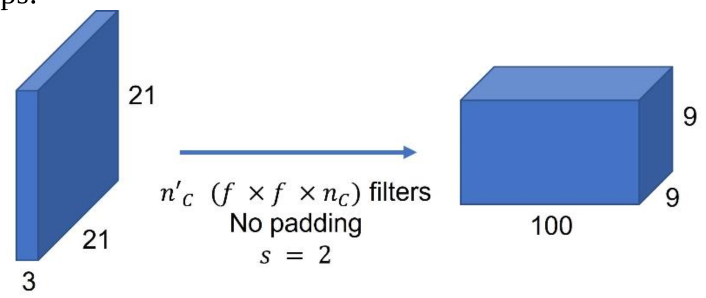
Question 3: Large Language Models [12 marks]

a) Most recent Large Language Models (LLMs) follow a parametric learning approach, i.e., the knowledge learned about the world is stored implicitly in the parameters of the model. Discuss **two** possible limitations of this setting. [6 marks]

b) Izacard et al.'s (2022) paper discussed in class, introduced the Atlas model which is a retrieval augmented language model. Retrieval augmented models employ a non-parametric memory, e.g., a neural retriever over a large, external, potentially non-static knowledge source to enhance a parametric language model. Discuss **two** aspects of how this approach can help alleviate some limitations of parametric LLMs. [6 marks]

Question 4: Convolutional Neural Networks [8 marks]

a) Given the following representation of a convolutional layer with input and output feature maps:



What are the values of the following hyperparameters (explain your workings):

   i. The number of filters n'_C [1 mark]

   ii. The filter size f [2 marks]

   iii. The filter width n_C [1 mark]

b) Consider a Convolutional Neural Network trained to classify objects in an image. After training, if you were to visualise the neurons' functions at different depths in the network, which neurons are more likely to behave like edge detectors? Explain your reasoning. [4 marks]

Part 2: Data Stream Mining and Transfer Learning [34 marks]

Question 5: Data Stream Mining [14 marks]

a) The majority of data stream research assumes that accurate class labels are readily available for each processed example or batch of data points. However, this assumption is impractical as it would result in exorbitant labelling costs. Describe **one** approach for data stream mining that enables the construction of a model when there are only a small number of labelled data points available. [8 marks]

b) Li et al.'s (2022) paper discussed in class, introduced the DDG-DA method as a means of addressing predictable concept drift through future data distribution forecasting. DDG-DA is designed to predict the sequential data distribution for the next time-step, enabling downstream learning models to be trained on data samples from the predicted distribution, rather than relying solely on catching up with the latest concept drift. As a dynamic data generator, DDG-DA creates sample data from previously observed data by following the predicted future data distribution. Specifically, it generates the resampling probability for each historical data sample to construct an estimated future data distribution. However, in practice, training this data generator to maximise the similarity between the predicted data distribution (represented by weighted resampling on historical data) and the ground truth future data distribution (represented by future data) can be challenging. For the challenge presented, can you discuss **one** potential solution? This can either be the solution proposed by the paper, or you can propose your own solution. Explain your answer. [6 marks]

Question 6: Transfer Learning [20 marks]

a) Discuss the differences between the following transfer learning methods:
   i. feature transfer, and
   ii. parameter transfer. [6 marks]

b) Explain the concept of fine-tuning using a pre-trained model. [4 marks]

c) In a study, two tasks (A and B) were created by randomly splitting 1000 classes into two groups creating two datasets. Two eight-layer convolutional networks, called *baseA* and *baseB*, were trained on Tasks A and B, respectively. Consider the following scenarios:

- The first four layers were copied from *baseA* and frozen, and the remaining four layers (Layers 5 to 8) were initialised randomly and trained on the target dataset, Task B. This transfer network is called A4B.
- The first four layers were copied from *baseA* and frozen, and the remaining four layers were initialised randomly and trained on the target dataset Task B, and fine-tuning was allowed on all layers. This transfer network is called A4B*.
- The first four layers are copied from *baseB* and frozen. The four remaining layers are initialised randomly and trained on Task B. This network is called B4B.

- The first four layers are copied from *baseB* and frozen. The four remaining layers are initialised randomly and trained on Task B, and fine-tuning was allowed on all layers. This network is called B4B*.

The average accuracy over the validation set for the trained networks described above is shown in this table:

Network	Accuracy (%)
A4B	59
A4B*	69
B4B	62
B4B*	64

Does the transfer across different tasks boost the performance? Explain your answer and discuss any assumptions you make. [5 marks]

d) Given the following scenario, assume you have a large amount of source data and a small amount of target data. Labelled data is available in both source and target datasets and the target tasks are different from the source tasks. Describe **one** transfer learning method suitable for this scenario. Explain your answer. [5 marks]

Part 3: Continual Learning, Domain Generalisation and Self-Supervised Learning [33 marks]

Question 7: Machine Learning Paradigms [6 marks]

Choose which of the following machine learning paradigms (transfer learning, multi-task learning, continual learning, domain adaptation, domain generalisation) is the most appropriate to model each of these problems, and justify your choice:

a) We have an existing machine learning model which predicts whether a patient has a respiratory disease, from patient data. We want to update the model, so that it can also predict whether the patient has diabetes. We no longer have access to the respiratory disease data that we originally used to train the model. [2 marks]

b) We are creating a language model which will summarise text. We only need the model to summarise text in three languages: English, French and Arabic. We will train this model using three datasets of summarised text, one from each of these three languages. [2 marks]

c) We are creating a model which will use X-ray scans to diagnose broken bones. We want this model to work for various hospitals which all have different X-ray machines that produce slightly different scans. [2 marks]

Question 8: Continual Learning [9 marks]

a) In Lopez-Paz et al.'s (2017) paper discussed in class, they introduce Gradient Episodic Memory, which stores a buffer of previous data instances to prevent forgetting. Explain why they constrain the performance of these previous task instances rather than re-training on them directly. [4 marks]

b) Two types of methods for continual learning are regularisation-based methods and architecture-based methods. Describe these two types of techniques and give one advantage of each type. [5 marks]

Question 9: Domain Generalisation **[6 marks]**

a) Define a "domain" in machine learning. [2 marks]

b) Explain how adversarial learning can be used to make models learn domain-invariant representations, and explain why learning domain-invariant representations is useful in domain generalisation. [4 marks]

Question 10: Self-Supervised Learning **[12 marks]**

a) Suppose that you want to create a model which generates missing sound for videos that were recorded without sound. You will train a machine learning model that takes a video as input and generates audio. You have a large set of YouTube videos to use as training data.

   i. Describe how you could manipulate this data to train your model to complete this task, using a self-supervised learning training scheme. [3 marks]

   ii. After you have finished training your model, it works well in generating missing sound on videos. Your friend suggests that you may be able to fine-tune your model and train it to classify the genre of videos instead of generating audio. Do you think that your model could be fine-tuned to perform well for this purpose? Explain your reasoning. [4 marks]

b) Describe the process of contrastive learning techniques for self-supervised learning. Explain the importance of negative sampling in this process. [5 marks]

Reference List

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. Communications of the ACM, 64(3), 107-115.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. Advances in Neural Information Processing Systems, 31.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155.

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., ... & Grave, E. (2022). Few-shot learning with retrieval augmented language models. arXiv preprint arXiv:2208.03299.

Ke, Z., Liu, B., & Huang, X. (2020). Continual learning of a mixed sequence of similar and dissimilar tasks. Advances in Neural Information Processing Systems, 33, 18493-18504

Li, W., Yang, X., Liu, W., Xia, Y., & Bian, J. (2022). DDG-DA: Data Distribution Generation for Predictable Concept Drift Adaptation. Proceedings of the AAAI Conference on Artificial Intelligence, 36(4), 4092-4100.

Sun, J., Wei, D., Ma, K., Wang, L., & Zheng, Y. (2022). Boost Supervised Pretraining for Visual Transfer Learning: Implications of Self-Supervised Contrastive Representation Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 36(2), 2307-2315.

Zhu, Y. N., & Li, Y. F. (2020, April). Semi-supervised streaming learning with emerging new labels. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 04, pp. 7015-7022).

Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C. Y., Ren, X., Su, G., Perot, V., Dy, J. & Pfister, T. (2022). Dualprompt: Complementary prompting for rehearsal-free continual learning. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI (pp. 631-648). Cham: Springer Nature Switzerland.

Garg, S., Balakrishnan, S., Lipton, Z. C., Neyshabur, B., & Sedghi, H. (2022). Leveraging unlabeled data to predict out-of-distribution performance. In International Conference on Learning Representations.

Lopez-Paz, D., & Ranzato, M. A. (2017, December). Gradient episodic memory for continual learning. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 6470-6479).