UiT

THE ARCTIC
UNIVERSITY
OF NORWAY

Home Exam

Home Exam in	**FYS-8604 Multi-Modal Learning**
Hand-out:	**Monday June 16, 2025, 09:00**
Hand-in:	**Monday August 11, 2025, 12:00**

The Home Exam contains 4 pages including this cover page

Contact person:	**Kristoffer Wickstrøm**
Email:	**kristoffer.k.wickstrom@uit.no**

UNIVERSITY OF TROMSØ

Before You Start

Portfolio instructions

Your code should be submitted together with your report (see instructions below). **Further, please include a discussion of the results obtained and a discussion of the implementation in your report, which show that you understand what you are doing.**

The code should be commented in such a way that any person with programming knowledge should be able to understand how the program works. Like your report, the code must be your own individual work.

You are permitted to use deep learning frameworks such as Pytorch and Tensorflow. As there is a lot of code available online, please make sure that your report and code clearly show that you understand what you are doing.

Hand-in format

Please submit your report (in pdf format) to WISEflow and attach *one* single `.zip` file that contains two folders, one called `doc` that contains your report, and another one called `src` containing the code. The file name of the `.zip` file should follow the format `homeexam_candidateXX.zip` (replace *XX* with your candidate number obtained from WISEflow) for anonymity. *Your report should be maximum 8 pages*, not including references.

Please include your candidate number and the course name on the frontpage of your report.

Follow the hand-in instruction in Wiseflow. Upload the pdf as the main file and then attach your zip as an attachment. Note, the reports will be processed by a plagiarism checker and the **pdf file size must not exceed 15MB**.

Problem 1

You are working as a machine learning engineer in an international post sorting company, and your boss has tasked you with solving a challenging problem. Customers often provide handwritten digits that must be automatically classified as image data for efficiency, but the camera you have available frequently fails and introduces a great deal of noise into the data. You come up with an idea to incorporate a fail-safe into your system. In addition to writing the digits down, the customers will now also repeat the digits into a microphone, thus creating both an audio and an image version of their desired digits. Unfortunately, your available microphone is not of the best quality, and also frequently introduces noise in the audio recordings. Nevertheless, you believe that processing both modalities at the same time will lead to satisfactory performance.

To test your idea, you collect 17560 pairs of images and audio of digits with a corresponding label for training. This data is available on Canvas under "Files/Exam". In addition, you also collect an independent set of 5859 pairs of images and audio of digits with a corresponding label to test your system.

Note: The image and audio samples are paired, but stored in different data files. You must therefore take care when you process the data, since e.g. random shuffling can cause issues if not done consistently.

(1a) Describe and design a unimodal deep learning classification model for classifying the image data on its own. You can use the architecture provided during the first practical exercises of the summer school (available in Canvas under "Files/Practical/"). Train your model on the training data and evaluate on the test data. Provide plots of the training loss and test accuracy during training. Discuss the performance of your model. Plot 5 randomly sampled images from the training set and 5 randomly sampled images from the test set. What do you observe?

(1b) Repeat the problem 1a), but now for the audio data. Provide plots of the training loss and test accuracy during training. Discuss the performance of your model. Plot 5 randomly sampled audio recordings from the training set and 5 randomly sampled audio recordings from the test set. What do you observe?

(1c) Compare and discuss the results from problem 1a) and 1b).

(1d) Explain how multi-modal learning can be used to process and combine information from multiple sources. Explain how information from modalities are typically fused together in multi-modal deep learning-based models. Describe at least two ways how this fusion can be performed and explain how they work.

(1e) Describe and design a multi-modal deep learning classification model that can process both the image and audio data to perform digit classification. Train your model on the training data and evaluate on the test data. Provide plots of the training loss and test accuracy during training. Discuss the performance of your model and compare with the results from problem 1a) and 1b). What do you observe?

(1f) Repeat problem 1e but with a different fusion strategy. Compare the results of this fusion strategy with the results from problem 1e.

Problem 2

In this problem you will analyse the normalized temperature-scaled cross entropy (NT-Xent) loss, which is one of the most common loss functions in self-supervised learning, and which is widely used in multi-modal learning. You will compare the NT-Xent loss to a logistic loss where we consider creating pairs of positive and negative sample for training.

(2a) The NT-Xent loss for a positive pair of samples is deifned as:

$$l_{ij} = -\log\left(\frac{\exp\left(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp\left(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau\right)}\right), \tag{1}$$

where $\mathbb{1}_{[k\neq i]}$ is an indicator function that evaluates to 1 when $k \neq i$, τ is a temperature parameter, $\mathbf{z}_i$ and $\mathbf{z}_j$ is a positive pair of input vectors, $\mathbf{z}_k$ is a negative input sample, and $\mathrm{sim}(\cdot,\cdot)$ is the cosine similarity between two vectors. Show that for input vectors $\mathbf{u}$, $\mathbf{v}^+$, and $\mathbf{v}^-$, where $\mathbf{v}^+$ and $\mathbf{v}^-$ are positive and negative samples with respect to $\mathbf{u}$, Equation 1 can be decomposed into the following form (assuming a negative loss function):

$$\text{NT-Xent} = \frac{\mathrm{sim}(\mathbf{u}, \mathbf{v}^+)}{\tau} - \log\left(\sum_{\mathbf{v}\in\{\mathbf{v}^+,\mathbf{v}^-\}} \exp\left(\mathrm{sim}(\mathbf{u}, \mathbf{v})/\tau\right)\right) \tag{2}$$

(2b) Show that the gradient of Equation 2 with respect to $\mathbf{u}$ can be calculated as:

$$\frac{\partial \text{NT-Xent}}{\partial \mathbf{u}} = \left(1 - \frac{\exp\left(\mathrm{sim}(\mathbf{u}, \mathbf{v}^+)/\tau\right)}{Z(\mathbf{u})}\right)\frac{\mathbf{v}^+}{\tau} - \sum_{\mathbf{v}^-} \frac{\exp\left(\mathrm{sim}(\mathbf{u}, \mathbf{v}^-)/\tau\right)}{Z(\mathbf{u})}\frac{\mathbf{v}^-}{\tau}, \tag{3}$$

where

$$Z(\mathbf{u}) = \sum_{\mathbf{v}\in\{\mathbf{v}^+,\mathbf{v}^-\}} \exp\left(\mathrm{sim}(\mathbf{u}, \mathbf{v})/\tau\right) \tag{4}$$

and $\mathbf{u}$, $\mathbf{v}^+$, and $\mathbf{v}^-$ are normalized such that $\mathrm{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T\mathbf{v}$.

(2c) An alternative loss function for self-supervised learning is to use a logistic loss between pairs of positive and negative samples. This logistic loss (assuming a negative loss function) can be described as:

$$\text{NT-Logistic} = \log\left(\sigma\left(\frac{\mathrm{sim}(\mathbf{u}, \mathbf{v}^+)}{\tau}\right)\right) + \log\left(\sigma\left(-\frac{\mathrm{sim}(\mathbf{u}, \mathbf{v}^-)}{\tau}\right)\right), \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid activation function. Show that the gradient of Equation 5 can be calculated as

$$\frac{\partial \text{NT-Logistic}}{\partial \mathbf{u}} = \sigma\left(-\frac{\mathrm{sim}(\mathbf{u}, \mathbf{v}^+)}{\tau}\right)\frac{\mathbf{v}^+}{\tau} - \sigma\left(\frac{\mathrm{sim}(\mathbf{u}, \mathbf{v}^-)}{\tau}\right)\frac{\mathbf{v}^-}{\tau}. \tag{6}$$

Discuss the differences in using the NT-Xent loss compared to the NT-Logistic loss, with a particular focus on how negative samples are weighted.

(2d) Discuss the role of the temperature parameter τ in both loss functions. What happens when τ becomes very small or very large?