**Final Exam Winter 2023**

**COMP 499/691**

| **Student Name** |  |
| --- | --- |
| **Student ID** |  |

(please write your name and ID clearly)

**Instructions:**

* The final exam is closed book with no cheat sheet.
* Discussing with the other students is not allowed. Ask the instructor if a question is not clear to you.
* A scientific calculator is required. Laptops/tablets cannot be used.
* Bags, coats, and cell phones must be placed at the edge of the room. Cell phones must be placed on silent, before being stowed away in a bag/coat.
* You have **2 hours** (120 minutes) to reply to the 15 questions. Each question can give you up to 2 points.
* If you need more space than that allocated in the boxes, please use the white pages attached at the end of this document. Add a note both on the box and on the white paper to make clear the connection (e.g. write “Continue at the end” in the box of P.1.1 and “Continuation of P.1.1” in the white paper).
* In multiple-choice questions, one or more options can be true.

**Marks:**

| **P1** | Self-Supervised Learning | / 6 |
| --- | --- | --- |
| **P2** | Speech Recognition | / 10 |
| **P3** | (Large) Language Models | / 10 |
| **P4** | Other Conversational AI Tasks | / 4 |
| **Total** |  | / 30 |

| **P1** | **Self-Supervised Learning** |
| --- | --- |

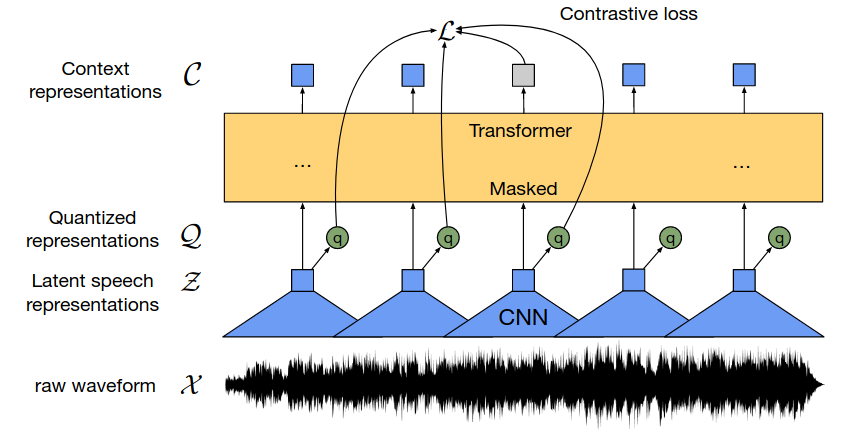
**P1.1 [2 points]** Briefly explain the difference between supervised and self-supervised learning. Also, mention the limitations of supervised learning that self-supervised learning attempts to overcome.

| . | |
| --- | --- |

**P1.2 [2 points]** Which of the following statements on autoencoders are true?

1. The output of the decoder is a latent representation that summarizes the content of the input signal.
2. The loss is the categorical cross-entropy.
3. The bottleneck prevents the model from learning an identity mapping (i.e., copying the inputs into the outputs).
4. A speech autoencoder can be trained with raw waveforms as input, but cannot be trained with spectrograms due to their bi-dimensional representation.
5. In a denoising autoencoder, the input is the clean signal and the output is the noisy one.

**P1.3 [2 points]** The figure below shows the wav2vec2 model. Describe how the quantized representations **q** are derived.



| . | |
| --- | --- |

| **P2** | **Speech Recognition** |
| --- | --- |

**P2.1 [2 points]** The Conformer is a model often used in state-of-the-art speech recognizers. Please, explain the difference between the Pointwise and Depthwise Convolution used in each convolution module.

|  | |
| --- | --- |

**P2.2 [2 points]** Which of the following statements on the Encoder-Decoder ASR are true?

1. A cross-attention mechanism connects the encoder and decoder states.
2. The number of output time steps can be larger than the number of input time steps.
3. The encoder is autoregressive.
4. Mean-Squared-Error (MSE) is typically used as a training objective.
5. A self-supervised model such as wav2vec2 or WavLM can be used as an encoder.

**P2.3 [2 points]** Which of the following statements on the Whisper are true?

1. It is a CTC-based model.
2. It is trained with a large multilingual dataset.
3. It can be used for speech recognition only.
4. The encoder is pretrained using unlabelled data.
5. The decoder jointly predicts the output tokens and the original input features.

**P2.4 [2 points]** Describe the architecture of Transducers. Explain the role of the Encoder, Predictor, and Joiner modules.

|  | |
| --- | --- |

**P2.5 [2 points]** Write an equation for the Word Error Rate (WER) in terms of insertions I, deletions D, and substitution S errors. Provide an example for each type of error.

|  | |
| --- | --- |

| **P3** | **(Large) Language Models** |
| --- | --- |

**P3.1 [2 points]** Write an equation for factorizing the joint probability P(w1,w2,w3, …, wN) over N words. Explain why it is challenging to estimate the terms that compose it with a simple counting approach.

| . | |
| --- | --- |

**P3.2 [2 points]** Which of the following statements on n-gram language modelsare true?

1. In bigram language models, we consider a history of two words.
2. We can significantly speed up the training of an n-gram language model using a GPU.
3. Interpolation techniques consider a lower-order n-gram if we have zero probability for a higher-order n-gram.
4. The number of possible n-grams increases exponentially with n.
5. N-gram language models suffer from data sparsity.

**P3.3 [2 points]** Can we generate text with BERT? If so, explain how.

| . | |
| --- | --- |

**P3.4 [2 points]** Which of the following statements on the Generative Pre-trained Transformer

(GPT-1) are true?

1. Similarly to BERT, GPT-1 is non-autoregressive.
2. A mask is used at training time to avoid attending the future inputs.
3. GPT-1 can learn good embeddings, but cannot be used to generate new text.
4. GPT-1 is trained on the WebText corpus (40 GB).
5. The model is trained to minimize the Negative Log-Likelihood (NLL) loss.

**P3.5 [2 points]** Which of the following statements on the Reinforcement Learning from Human Feedback (RLHF) used in ChatGPT are true?

1. RLHF is used to fine-tune a large language model like GPT3.5 that is trained in a self-supervised way.
2. In ChatGPT, RLHF is only used to provide supervised labels for training a reward model.
3. Proximal Policy Optimization is used to fine-tune the large language model using a pretrained reward model.
4. The Plackett-Luce model is a neural network used to convert ranks into single scores.
5. According to the authors, RLHF is helpful because it injects some “safeguards” into the model preventing it from responding negatively/harmfully or in a biased way too often.

| **P4** | **Other Conversational AI Tasks** |
| --- | --- |

**P4.1 [2 points]** You would like to design a neural network for frame-level emotion identification. You have a pretty large dataset composed of variable-length speech recordings corresponding to four different emotions (neutral, sad, happy, angry) annotated for each time step of the speech recording. The system should provide an output for every time step in real time while recording a signal (streamability). Explain how you would design such a system. In particular, mention:

1. *Which features would you use?*
2. *Which neural network?*
3. *Which loss?*
4. *How can you make the model streamable?*

|  | |
| --- | --- |

**P4.2 [2 points]** Which of the following statements on Speech Enhancement are true?

1. It is a regression task.
2. Autoregressive speech enhancement models are often used to remove reverberation.
3. One common practice for estimating the enhanced waveform is to apply the inverse short-time Fourier transform (ISTFT) using the enhanced spectrogram and the phase information from the original noisy signal.
4. In masked-based approaches, the mask can be applied to a latent representation.
5. Mask-based speech enhancement systems are trained with binary cross-entropy as a training objective.