CS5710 Machine Learning

Fall 2025

Home Assignment

Student name: Gangadhar Nersu

Part A — Short-Answer

1. **Positional Encoding Concepts**

a) Why do we need positional encodings in transformer models?
Ans - Transformers do not process tokens in a fixed order; the self-attention mechanism treats the entire sequence as a set.
Because of that, the model has no built-in sense of which token comes first or last. Positional encodings supply numerical patterns that indicate each token's location, allowing the model to understand order-dependent meaning such as word order in sentences or temporal structure in time-series data.

b) Describe two key requirements for a good positional encoding scheme.
Ans – 1. Each position should have a distinct representation.
The model must be able to differentiate index 1 from index 10, so every position in the sequence should map to a unique encoding vector.
2. It should generalize well to unseen lengths or shifts.
A robust scheme allows the model to handle sequences longer than what it saw during training or interpret relative distances consistently, even if absolute positions change.

c) What does it mean for the positional encoding matrix $M$to be *unitary* and *norm-preserving*?
Ans – 1.Unitary / Orthogonal:
Saying $M$is unitary means multiplying by it does not distort inner products; essentially, $M^T M = I$.
This maintains the geometric relationships between embeddings.
2.Norm-Preserving:
A norm-preserving matrix ensures that transforming a vector with does not change its length.
This prevents the encoding from unintentionally amplifying or shrinking embeddings, which keeps training stable and attention computations consistent.

2. **Attention Mechanism**

a) Define “attention score” and explain how it determines the weight of each token.
Ans - An attention score reflects how strongly a query matches a key. A higher score means the model considers that token more relevant. After scores are normalized, tokens with higher compatibility receive greater influence in the final output.

b) What mathematical operation is applied to convert alignment scores into attention weights?
Ans - The **softmax** function is applied to the scores so they become a probability distribution, ensuring all weights are positive and sum to one.

c) How is the *context vector* computed from these weights and values?
Ans - The context vector is obtained by taking a weighted sum of the value vectors, where the weights come from the softmax-normalized attention scores.

3. **Multi-Head Attention**

a) What is the main advantage of using multiple attention heads?
Ans - Multiple heads allow the model to analyze different types of patterns (e.g., long-distance links, syntax, semantics) at the same time, making the representation richer.

b) How does splitting Q, K, and V across different subspaces improve model representation?
Ans - By projecting Q, K, and V into smaller subspaces, each head can specialize in learning different relationships, giving more expressive attention behavior.

c) After multi-head attention, why is concatenation followed by another linear projection necessary?
Ans - Concatenation combines all heads' insights, and the final linear layer merges them into a single, unified representation suitable for the next layer.

4. Ethical Foundations

a) Explain why *ethics* is not the same as *laws* or *feelings*.
Ans - Laws are formal rules enforced by institutions; ethics evaluates whether actions are morally right beyond legal boundaries. Feelings are personal emotions, whereas ethical reasoning relies on principles and logical justification.

b) Briefly describe two classical ethical theories (e.g., utilitarianism and deontology) and how they would handle an AI decision scenario.
Ans – 1. Utilitarianism: Chooses the action producing the greatest overall benefit. In an AI scenario, it favors the decision with the best collective outcome.

2. Deontology: Judges actions based on duties or moral rules. In AI, it would avoid using any decision process that violates moral constraints, even if outcomes seem beneficial.

c) Why do philosophers argue that no single ethical theory clearly “wins” in all contexts?

Ans - Different theories prioritize different moral values. Real-world situations involve conflicting concerns - rights, outcomes, fairness - so no one theory resolves every scenario effectively.

5. Types of AI Harms

a) Define allocational harm and representational harm in AI systems.

Ans – 1. Allocational harm: When an AI system disproportionately gives or withholds opportunities or resources.

2. Representational harm: When AI outputs reinforce stereotypes or misrepresent certain groups.

b) Provide an example of each from real-world applications (e.g., translation, hiring, or facial recognition).

Ans – 1. Allocational: A loan approval model consistently declines applicants from a certain demographic.

2. Representational: A translation model defaults certain job roles to one gender, reinforcing bias.

c) Why is representational harm often harder to measure than allocational harm?

Ans - It deals with cultural meaning, identity, and perception—factors that are difficult to express with clear numerical metrics.

6. Sources of Dataset Bias

a) List three reasons why bias arises during data collection or annotation in AI datasets.

Ans – 1.**Sampling bias:** Collected data doesn't represent the full population.

2. **Annotator bias:** Human labelers inject personal assumptions.

2. **Measurement bias:** Sensors, scraping methods, or collection tools introduce structured distortion.

b) What kinds of data or groups tend to be under-represented in large language datasets?

Ans - Low-resource languages, marginalized communities, and non-Western cultural content are typically underrepresented in modern datasets.

c) How can bias amplification occur even after initial data preprocessing?

Ans - Learning algorithms often emphasize frequent training patterns, causing models to rely even more heavily on majority-group data than the dataset initially did.

7. Safety, Security, and Privacy

a) Define data poisoning and describe how it can manipulate a model's predictions.

Ans - Data poisoning means injecting malicious training examples so the model learns incorrect behaviors, such as misclassifying a specific input on purpose.

b) What are the ethical implications of model memorization (e.g., GPT-2 reproducing private or copyrighted text)?

Ans - If a model memorizes training text, it can leak private or copyrighted passages, violating privacy and intellectual property protections.

c) How does model stealing threaten privacy and intellectual property in AI research?

Ans - By repeatedly querying a model, an attacker can recreate a copy of it. This undermines the creator’s intellectual property and allows unauthorized systems to use or misuse the cloned model.