

Background Knowledge Diagnostic

CIS 5590: Vision-Language Models

Name: _____

Date: _____

Purpose: These questions assess whether you have the mathematical and machine-learning foundations needed for the course. Concise answers are sufficient; no knowledge of the assigned papers is expected.

- 1. Two-layer neural network.** Write the formula for a two-layer feed-forward neural network that maps an input vector $\mathbf{x} \in \mathbb{R}^d$ to an output vector $\mathbf{y}$. Identify the weight matrices, bias vectors, and nonlinear activation function, and state their dimensions.
- 2. Backpropagation and optimization.** What is a gradient, and how is it used by gradient descent to update a neural-network parameter θ ? Write the basic update equation and explain the role of the learning rate.
- 3. Softmax.** Define the softmax function for a vector of logits $\mathbf{z}$. What properties make its output interpretable as a probability distribution?
- 4. Cross-entropy loss.** Write the cross-entropy loss for a multiclass classification problem with a one-hot target. How is it related to maximizing the likelihood of the correct class?
- 5. Next-token prediction.** Describe next-token prediction in an autoregressive language model. Given tokens $x_1, \dots, x_T$, write the training loss and explain what probability distribution the model learns at every position.
- 6. Embeddings and similarity.** What is an embedding? Define cosine similarity between two embedding vectors and explain why cosine similarity is commonly used to compare image and text representations.
- 7. Self-attention.** Given an input matrix X , write the formulas for queries, keys, and values,

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V,$$

and then write the scaled dot-product self-attention operation. Explain why the factor $1/\sqrt{d_k}$ is used and what the attention weights represent.

- 8. Images as transformer tokens.** An RGB image of size 224×224 is divided into non-overlapping 16×16 patches. How many visual tokens are produced? Describe how each patch can be converted into a transformer input embedding.
- 9. Evaluation metrics.** Define precision, recall, and intersection over union (IoU). Give an example in which accuracy could be high even though a model performs poorly on the class of interest.
- 10. Generalization and experimental design.** Explain the difference among training, validation, and test sets. What is overfitting, and why must examples from the same image, video, patient, or data source sometimes be kept within only one split?

This diagnostic is intended to identify topics that may benefit from review before or early in the course.