Prep Questions for Midterm Examination: Computer Vision

1. **Question:** In a standard Vision Transformer (ViT-Base), what component is used to make the final classification prediction?

   A) The average of all patch tokens
   B) The first convolutional feature map
   C) The CLS token passed through an MLP head
   D) The last attention head only

   Answer: C) The CLS token passed through an MLP head.

2. **Question:** DETR (DEtection TRansformer) reframes object detection as:

   A) Pixel-wise segmentation
   B) Anchor regression
   C) Set prediction
   D) Sliding window classification

   Answer: C) Set prediction.

3. **Question:** What fundamentally distinguishes the two dominant paradigms of Vision-Language Models (VLMs), such as Flamingo vs. LLaVA?

   A) They use completely different loss functions.
   B) Flamingo uses cross-attention to keep vision and language separate, while LLaVA uses a unified token space where visual information is converted to tokens and processed in the same self-attention layers as text.
   C) LLaVA relies exclusively on Convolutional Neural Networks for text generation.
   D) Flamingo does not use a vision encoder.

   Answer: B) Flamingo uses cross-attention to keep vision and language separate, while LLaVA uses a unified token space where visual information is converted to tokens and processed in the same self-attention layers as text.

4. **Question:** DETR requires Non-Maximum Suppression (NMS) during inference to remove duplicate bounding box predictions. **Answer:** False. DETR uses bipartite matching (Hungarian matching) to ensure a one-to-one assignment between predictions and ground truth, eliminating the need for NMS.

5. **Question:** Median filtering is a linear filtering operation. **Answer:** False. Median filtering is a non-linear operation, meaning $\text{filter}(I_1 + I_2) \neq \text{filter}(I_1) + \text{filter}(I_2)$.

6. **Question:** In-context learning allows a Large Language Model to adapt to new tasks without updating its model parameters. **Answer:** True.

7. **Question:** Explain the concept of "separability" in the context of Gaussian filters and why it is computationally useful. **Answer:** A 2D Gaussian kernel is separable, meaning it can be factored into a product of two 1D Gaussians. This is highly useful because it allows a 2D convolution to be reduced to two separate 1D convolutions (one along the rows and one along the columns), which drastically reduces the computational complexity from $O(n^2m^2)$ to $O(n^2m)$.

8. **Question:** What is the main role of a backbone in a modern vision architecture? **Answer:** The main role of the backbone is to extract features from the input image. It encodes the visual world into a feature map, usually decreasing spatial resolution while increasing semantic depth, which is then passed to a prediction head.

9. **Question:** Briefly explain the difference between Semantic Segmentation, Instance Segmentation, and Panoptic Segmentation. **Answer:**

   - **Semantic Segmentation:** Classifies every pixel into a category, but treats multiple objects of the same class as a single mass.
   - **Instance Segmentation:** Detects and segments each individual object instance, distinguishing between separate objects of the same class.
   - **Panoptic Segmentation:** A unified framework that collectively segments both individual objects ("things") and background regions ("stuff").

10. **Question:** Write the universal mathematical formula for Scaled Dot-Product Self-Attention. **Answer:** $\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$. (Where Q represents Queries, K represents Keys, V represents Values, and d_k is the dimensionality of the key vectors used for scaling).

11. **Question:** Assuming an input feature map has a spatial resolution of $H \times W$ with K input channels, how many operations are required to compute the output feature volume for a convolutional layer utilizing L filters of size $F \times F \times K$? **Answer:** F^2KLHW operations.

12. **Question:** Given a Vision Transformer (ViT) processing an image of size $H \times W$, write the formula to calculate the number of patch tokens N generated if the patch size is $P \times P$. If the patch size P is halved, what happens to the sequence length and the computational cost of the self-attention mechanism? **Answer:** The number of patch tokens is calculated using the formula $N = (H/P) \times (W/P)$. If the patch size P is halved, the total number of tokens N quadruples (increases by a factor of 4). Because the computational cost of the self-attention mechanism scales quadratically with the sequence length (an $O(N^2)$ operation), quadrupling the tokens increases the attention computational cost by a factor of 16.

13. **Question:** Compare the number of weights (parameters) required for a single 7×7 convolutional layer versus three stacked 3×3 convolutional layers, assuming both configurations have K input and output channels. Why was the latter architectural approach favored in models like VGGNet? **Answer:** A single 7×7 convolutional layer with K input and output channels requires $49K^2$ weights. In contrast, three stacked 3×3 convolutional layers require $3 \times (3 \times 3 \times K \times K) = 27K^2$ weights. The stacked 3×3 approach was favored in VGGNet because it achieves the same effective receptive field as a 7×7 convolution but significantly reduces the number of parameters and allows for the insertion of non-linear ReLU activations between each layer.

14. **Question:** Explain the mathematical concept of autoregressive factorization in Large Language Models (LLMs). How is the joint probability of a token sequence $x_{1:T}$ decomposed? **Answer:** Autoregressive factorization decomposes the joint probability of an entire token sequence into a chain of conditional next-token predictions. Mathematically, the probability of the sequence $p_\theta(x_{1:T})$ is represented as the product of the conditional probabilities of each individual token given all preceding tokens: $\prod_{t=1}^{T} p_\theta(x_t|x_{<t})$. This architectural design dictates that at each step during inference, the model predicts the next token based sequentially on the context of all previously generated tokens.

15. **Question:** Why is padding necessary when applying a linear filter to an image?
    Answer: Padding is necessary to **control the size of the output**. Without padding, applying a filter would cause the output image to be smaller than the input image, so padding allows the output to either remain the same size or become larger depending on the implementation.

16. **Question:** Why does a convolutional architecture tile units over the input image and share their weights instead of using a standard multi-layer perceptron (MLP) with fully connected layers?
    Answer: Fully connected layers are inefficient for images because they do not account for spatial locality. Convolutional architectures resolve this by **limiting the receptive fields of units and sharing their weights** across the image, which acts as learned local templates (feature maps).

17. **Question:** In dense prediction architectures, what are the two necessary steps required to upsample a feature map by a factor of 2?
    Answer: The first step is to **increase the resolution of the feature grid** (often by dilating the input by inserting rows and columns of zeros), and the second step is to **interpolate to get the missing values**.

18. **Question:** How does a Vision Transformer (ViT) acquire spatial awareness for its patch tokens since transformers process patches as an unordered set?
    Answer: ViTs achieve spatial awareness by **adding a positional encoding** (either a learnable vector or a fixed sinusoidal function) to each patch embedding. This allows the model to understand where each patch originated in the image.

19. **Question:** How does the Segment Anything Model (SAM) shift the fundamental approach to segmentation compared to traditional semantic or instance segmentation models?
    Answer: SAM shifts segmentation from predicting fixed object categories to **promptable segmentation**. This means it functions as a foundation model that predicts a mask dynamically **conditioned on a user prompt** (like text, points, or bounding boxes) rather than relying on a fixed class vocabulary.

20. **Question:** In contrastive learning models, what is the concept of "collapse," and how do negative pairs prevent it?
    Answer: "Collapse" occurs when a model trivializes the task by **mapping all embeddings to an identical vector** to minimize the distance between positive pairs. Negative pairs prevent this by **introducing a repulsion force**, which enforces separation and structure in the joint embedding space.

21. **Question:** Explain the problem of "exposure bias" that occurs during the autoregressive generation loop of Large Language Models.
    Answer: During training, models use "teacher forcing," meaning they are conditioned on real, ground-truth prefixes from the dataset. However, during inference, models condition sequentially on their own generated predictions. This mismatch can cause **small mistakes early on to snowball into larger deviations**, making long-horizon generation particularly vulnerable to drift and instability.

22. **Question:** In the context of Vision-Language Models, what is visual "hallucination" and what is its primary root cause?

Answer: Visual hallucination occurs when a model generates plausible-sounding content that **contradicts what is actually visible** in the input image. Its primary root cause is that **the language prior overrides the visual evidence**, meaning the model relies on statistical assumptions (e.g., assuming a microwave is present just because the scene is a kitchen) rather than what is actually depicted.