Компьютерное зрение **ФИО:** _____
Весна 2025 / Университет Иннополис
Промежуточный экзамен **Группа:** _____
11 марта 2025 г.
Продолжительность: 45 минут

Промежуточный экзамен содержит 5 страниц (включая титульный лист) и 10 вопросов с множественным выбором и свободными ответами по различным темам курса. Вопросы различаются по сложности (4, 6 или 9 баллов за вопрос). Проверьте, нет ли пропущенных страниц. Подпишите каждый лист.

- Напишите свое имя и название группы четко и разборчиво на листе ответов.
- Внимательно прочитайте каждый вопрос и выберите один или несколько вариантов ответа, либо дайте подробный ответ, если это необходимо.
- Не разговаривайте и не общайтесь с другими студентами во время экзамена, а также не используйте электронные устройства, книги или подготовленные ответы.

За неправильные ответы не предусмотрено штрафа, поэтому рекомендуется не оставлять ни одного вопроса без ответа. После завершения работы оставьте свой экзаменационный лист в ящике, предназначенном для вашей группы. Обратите внимание, что вы можете покинуть аудиторию только после сдачи работы. Удачи!

Total Score	**Grade**

Question	**A**	**B**	**C**	**D**	**Max. Score**	**Score**
1					0.2	
2					0.2	
3					0.2	
4	-	-	-	-	0.5	
5	-	-	-	-	0.5	
6	-	-	-	-	1.0	
7	-	-	-	-	1.5	
8	-	-	-	-	1.7	
9	-	-	-	-	1.7	
10	-	-	-	-	2.5	

1. Какова основная функция механизма self-attention в Vision Transformers?
   A. Он позволяет учитывать глобальные зависимости между патчами изображения.
   B. Он заменяет необходимость использования позиционных эмбеддингов.
   C. Он уменьшает размерность входных данных без потери информации.
   D. Он обеспечивает локальную фильтрацию, подобно свёрточным ядрам.

2. Какую функцию потерь чаще всего применяют для борьбы с дисбалансом классов в задачах сегментации?
   A. Кросс-энтропию.
   B. Dice loss.
   C. MSE (среднеквадратичная ошибка).
   D. Hinge loss.

3. Функция потерь Triplet Loss содержит оператор max в выражении

$$\mathcal{L} = \max\left(||f(a) - f(p)||^2 - ||f(a) - f(n)||^2 + \alpha,\, 0\right).$$

   Почему эта функция потерь пригодна для оптимизации методами градиентного спуска, несмотря на то, что функция $\max(x, 0)$ не является дифференцируемой в точке $x = 0$?
   A. В точке $x = 0$ используется понятие субградиента, позволяющее вычислять направление градиентного спуска даже при недифференцируемости.
   B. Недифференцируемость в одной точке не оказывает влияния на оптимизацию, так как эта точка никогда не встречается на практике.
   C. Современные оптимизаторы основаны на численных методах, которые не требуют аналитической дифференцируемости.
   D. Функция $\max(x, 0)$ считается дифференцируемой, так как градиент в точке $x = 0$ задаётся равным 1.

4. Дано изображение $x \in \mathbb{R}^{224\times224\times3}$, которое разбивается на патчи размером 16×16. Рассчитайте:
   a) Количество патчей N.
   b) Размерность вектора каждого патча после выравнивания.

   Объяснение:
   a) Число патчей: $(224/16)^2 = 14^2 = 196$.
   b) Каждый патч имеет размер 16×16 с 3 каналами, значит размерность = 16*16*3 = 768.
   a) 196; b) 768.

5. При фьюжене данных для оценки глубины используют взвешенное объединение данных с LiDAR и RGB-камеры. Если LiDAR-оценка глубины $D_{\text{LiDAR}} = 3.5$ м, а оценка по RGB $D_{\text{RGB}} = 4.0$ м, и веса фьюжена равны $w_1 = 0.6$ и $w_2 = 0.4$ соответственно, вычислите итоговую глубину D_{final} по формуле:

$$D_{\text{final}} = w_1 \cdot D_{\text{LiDAR}} + w_2 \cdot D_{\text{RGB}}.$$

3.7 м.

6. Даны Ground Truth маска (100 пикселей объекта, 900 фона), предсказание 80 TP (True Positives), 20 FN (False Negatives), 50 FP (False Positives). Рассчитайте IoU (отношение пересечения к объединению), Dice Coefficient (отношение удвоенного пересечения к общей площади). Учтите, что метрики используются для оценки качества сегментации объекта, а не фона.

   1. **IoU (Intersection over Union)**:
   $$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$
   2. **Dice Coefficient**:
   $$\text{Dice} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}$$

   Подсказки:

   - **IoU**: Отношение пересечения (TP) к объединению (TP + FP + FN).
   - **Dice**: Похож на IoU, но удваивает TP в числителе для учета дисбаланса классов.

   Ответ:
   $$\text{IoU} = \frac{80}{80 + 50 + 20} = \frac{80}{150} \approx 0.53$$
   $$\text{Dice} = \frac{2 \times 80}{2 \times 80 + 50 + 20} = \frac{160}{230} \approx 0.69$$

7. Опишите разницу между использованием одного большого свёрточного ядра (например, 7×7) и последовательным применением нескольких малых ядер (например, трёх последовательных свёрток 3×3) в архитектуре CNN. Охарактеризуйте влияние на: эффективное поле зрения, вычислительную стоимость, возможность введения нелинейностей.

   Использование одного большого ядра (7×7) даёт широкое поле зрения, но за счёт большого числа параметров и вычислительной сложности, а также с меньшим числом нелинейных слоёв. Три последовательные свёртки 3×3 обеспечивают эквивалентное эффективное поле зрения, требуют меньше параметров, снижают вычислительные затраты и позволяют применять больше нелинейных активаций, что улучшает качество модели.

8. Рассмотрим сверточный слой, получающий входной тензор размером $32 \times 32 \times 3$. Слой применяет 64 фильтра размером 5×5 с шагом (stride) $s = 2$ и заполнением (padding) $p = 2$. 1. Вычислите пространственные размеры (высота и ширина) выходного тензора, используя формулу:
   $$r = \left\lfloor \frac{n + 2p - k}{s} \right\rfloor + 1.$$

   2. Определите общее число обучаемых параметров (веса и смещения) в этом слое.

   Объяснение:

   а) Подставляем: n = 32, p = 2, k = 5, s = 2 $\rightarrow r = \lfloor(32+4-5)/2\rfloor+1 = \lfloor 31/2 \rfloor+1 = 15+1 = 16$.

   Выходной тензор имеет размер $16 \times 16 \times 64$.

b) Для каждого фильтра: число весов $= 5 \times 5 \times 3 = 75$; плюс 1 смещение, итого 76 параметров. Всего $64 \times 76 = 4864$ параметров.

a) $16 \times 16 \times 64$; b) 4864 параметров.

9. Остаточные связи (residual connections, или skip connections) играют ключевую роль в обучении очень глубоких нейронных сетей. Объясните, как остаточные связи помогают смягчить проблему затухания градиентов. Приведите краткое математическое объяснение базового остаточного блока и обсудите, как тождественное отображение (identity mapping) способствует более стабильному обучению и повышению производительности глубоких CNN.

   Объяснение:

   Остаточные связи позволяют градиенту протекать напрямую через блоки, что снижает проблему затухания градиентов. Базовый остаточный блок описывается выражением $y = F(x) + x$, где $F(x)$ – свёрточные преобразования с нелинейностями. Тождественное отображение x обеспечивает прямой путь для обратного распространения градиента. Остаточные связи помогают решать проблему затухания градиентов, предоставляя прямой путь для распространения градиента через блоки сети. Базовый остаточный блок описывается формулой $y = F(x) + x$, где $F(x)$ представляет собой серию свёрточных слоёв с нелинейными активациями. Тождественное отображение x гарантирует сохранение исходной информации, что стабилизирует процесс обучения и повышает эффективность глубоких CNN.

10. Даны два патча (токена), преобразованные в эмбеддинги, включающие по две части:

$$\mathbf{z}_1 = \mathbf{x}_1 + \mathbf{p}_1, \quad \mathbf{z}_2 = \mathbf{x}_2 + \mathbf{p}_2,$$

   где: $\mathbf{x}_1 = [1, 0]$, $\mathbf{x}_2 = [0, 1]$ - содержательные эмбеддинги, $\mathbf{p}_1 = [\alpha, \beta]$, $\mathbf{p}_2 = [\gamma, \delta]$ - позиционные эмбеддинги.

   Весовые матрицы и матрицы Q, K, V для входной последовательности $X \in \mathbb{R}^{N \times d}$ (где N количество патчей, d размерность эмбеддинга):

$$W_Q = W_K = W_V = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, Q = XW_Q, \quad K = XW_K, \quad V = XW_V.$$

   Упрощенная формула внимания (scaled dot-product):

$$\text{Attention} = \frac{QK^T}{\sqrt{d}}.$$

   1. **Без позиционных эмбеддингов** ($\alpha = \beta = \gamma = \delta = 0$). Покажите, что механизм Attention делает модель нечувствительной к порядку токенов.

   2. **С позиционными эмбеддингами** ($\alpha \neq \gamma$, $\beta \neq \delta$). Покажите обратное, что механизм Attention чувствителен к порядку токенов.

   3. **Анализ**: а) Какие задачи (NLP/CV) сильнее страдают от отсутствия позиционных эмбеддингов? б) Предложите альтернативу синусоидальным эмбеддингам для кодирования позиций.

Решение

1. Без позиционных эмбеддингов:

$$\mathbf{z}_1 = \mathbf{x}_1 = [1, 0], \quad \mathbf{z}_2 = \mathbf{x}_2 = [0, 1].$$

Вычисляем внимание:

$$\text{Attention}(\mathbf{z}_1, \mathbf{z}_2) = \frac{[1, 0] \cdot [0, 1]^T}{\sqrt{2}} = \frac{0}{\sqrt{2}} = 0.$$

$$\text{Attention}(\mathbf{z}_2, \mathbf{z}_1) = \frac{[0, 1] \cdot [1, 0]^T}{\sqrt{2}} = 0.$$

Вывод: Результаты идентичны → модель не различает порядок.

2. С позиционными эмбеддингами:

$$\mathbf{z}_1 = [1 + \alpha, \beta], \quad \mathbf{z}_2 = [\gamma, 1 + \delta].$$

Вычисляем внимание:

$$\text{Attention}(\mathbf{z}_1, \mathbf{z}_2) = \frac{(1 + \alpha)\gamma + \beta(1 + \delta)}{\sqrt{2}}.$$

$$\text{Attention}(\mathbf{z}_2, \mathbf{z}_1) = \frac{\gamma(1 + \alpha) + (1 + \delta)\beta}{\sqrt{2}}.$$

Вывод: Если $\alpha \neq \gamma$ или $\beta \neq \delta$, результаты различны → порядок влияет на внимание.