Instruções: Responda às questões de múltipla escolha assinalando apenas uma alternativa. Nas questões dissertativas, apresente o raciocínio passo a passo, incluindo fórmulas e cálculos intermediários quando solicitado. Mantenha a notação S para a matriz de similaridades, τ para a temperatura e $\mathcal{L}$ para a perda, conforme utilizadas em aula.

Questões de Múltipla Escolha.

I) Considere o paradigma dominante em visão computacional antes do CLIP, baseado em pré-treino supervisionado em `ImageNet` seguido de *linear probe* ou *fine-tuning*. Qual das alternativas abaixo melhor caracteriza a principal limitação dessa abordagem para tarefas de vocabulário aberto?

(a) O conjunto de classes é fixado em tempo de treino e a rotulação manual é cara, o que dificulta cobrir conceitos novos.

(b) A função de ativação utilizada na última camada impede que a rede generalize para novas distribuições de imagens.

(c) As camadas convolucionais não conseguem extrair características semânticas úteis sem regularização adicional.

(d) O número de parâmetros do *backbone* é insuficiente para representar mais de 1000 classes simultaneamente.

II) O CLIP é treinado em pares (imagem, texto) coletados em escala da web. Qual é, em essência, o sinal de supervisão utilizado pelo modelo durante o pré-treino?

(a) Rótulos categóricos atribuídos por anotadores humanos a cada imagem do conjunto.

(b) Pseudo-rótulos produzidos por uma CNN previamente treinada em `ImageNet`.

(c) A associação natural entre cada imagem e a legenda que a acompanha na página de origem.

(d) Máscaras de segmentação derivadas automaticamente de modelos generativos auxiliares.

III) Em um *batch* de N pares (imagem, texto) durante o treino do CLIP, quantos pares são tratados como positivos e quantos como negativos na matriz $S \in \mathbb{R}^{N \times N}$?

(a) N positivos e N negativos, um par positivo e um negativo por linha.

(b) N positivos (na diagonal) e $N(N-1)$ negativos (fora da diagonal).

(c) $N(N-1)/2$ positivos e $N(N-1)/2$ negativos, distribuídos simetricamente.

(d) 1 positivo e $N^2 - 1$ negativos, escolhidos por amostragem aleatória.

IV) Antes de compor a matriz S usada na *InfoNCE*, o CLIP projeta as saídas dos *encoders* em um espaço comum e normaliza os vetores. Qual operação de similaridade entre um vetor de imagem $\mathbf{u}$ e um vetor de texto $\mathbf{v}$ é efetivamente computada após essa normalização?

(a) Distância euclidiana $\|\mathbf{u} - \mathbf{v}\|_2$ entre as projeções.

(b) Produto interno cru $\mathbf{u}^\top \mathbf{v}$ sobre os vetores não normalizados.

(c) Similaridade do cosseno, equivalente a $\tilde{\mathbf{u}}^\top \tilde{\mathbf{v}}$ após normalização L2.

(d) Divergência de Kullback-Leibler entre as distribuições $\mathbf{u}$ e $\mathbf{v}$.

V) Em uma linha da matriz S obtém-se os logits $(0.96,\ 0.64)$ associados ao par positivo (coluna 1) e a um negativo (coluna 2). Aplicando softmax a S/τ, qual coluna recebe maior probabilidade quando $\tau = 0.1$ em comparação com $\tau = 1.0$?

(a) Em ambos os valores de τ a distribuição se mantém praticamente uniforme entre as duas colunas.

(b) Com $\tau = 1.0$ o positivo recebe cerca de 0.58 e com $\tau = 0.1$ esse valor sobe para aproximadamente 0.96.

(c) Com $\tau = 1.0$ o positivo recebe cerca de 0.96 e com $\tau = 0.1$ a distribuição fica uniforme.

(d) Diminuir τ inverte a ordem das probabilidades, fazendo o negativo passar a dominar.

VI) ☠️ Suponha um *batch* com $N = 2$ e a matriz de similaridades já dividida por τ:

$$S/\tau = \begin{pmatrix} 9.6 & 6.4 \\ 3.6 & 9.6 \end{pmatrix}.$$

Considerando o termo $\mathcal{L}_{i\to t} = -\frac{1}{N}\sum_i \log \frac{\exp(S_{ii}/\tau)}{\sum_j \exp(S_{ij}/\tau)}$, qual é o valor aproximado dessa perda?

(a) Aproximadamente 0.49.

(b) Aproximadamente 0.25.

(c) Aproximadamente 0.02.

(d) Aproximadamente 0.69.

VII) A perda total do CLIP é definida como $\mathcal{L}_{\text{CLIP}} = \frac{1}{2}(\mathcal{L}_{i\to t} + \mathcal{L}_{t\to i})$. Qual é a justificativa principal para essa simetrização?

(a) Garantir que a matriz S seja simétrica em todos os passos de treino.

(b) Permitir que cada modalidade atue como consulta sobre a outra, evitando viés em uma única direção de recuperação.

(c) Reduzir pela metade o custo computacional do passo de *backpropagation*.

(d) Forçar os dois *encoders* a compartilhar pesos durante o treino.

VIII) Sobre a arquitetura do CLIP, qual afirmação descreve corretamente a relação entre os dois *encoders*?

(a) Os *encoders* de imagem e de texto compartilham pesos e formam um único *transformer* multimodal.

(b) O *encoder* de texto recebe como entrada o vetor de saída do *encoder* de imagem para então gerar a representação textual.

(c) Os *encoders* são independentes, treinados conjuntamente, e se conectam apenas no espaço de projeção onde a similaridade é calculada.

(d) Apenas o *encoder* de imagem é treinado; o *encoder* de texto é congelado a partir de um BERT pré-treinado.

IX) Para realizar classificação *zero-shot* com CLIP em um problema com K classes, qual pipeline descreve corretamente o procedimento de inferência?

(a) Treinar uma camada linear sobre o *encoder* de imagem usando exemplos rotulados das K classes antes da inferência.

(b) Codificar a imagem de teste e cada uma das K classes como texto, calcular a similaridade entre os vetores e tomar o arg max das similaridades.

(c) Submeter a imagem a uma busca por vizinhos mais próximos no `ImageNet` e propagar o rótulo majoritário.

(d) Gerar legendas candidatas com um *decoder* autoregressivo e selecionar aquela com maior probabilidade.

X) Estudos com CLIP mostram que utilizar *prompts* como "*a photo of a* {*classe*}" produz acurácia maior do que usar apenas o nome da classe ("{*classe*}"). Qual é a explicação mais consistente com o modo como o modelo foi treinado?

(a) O *prompt* aumenta o número de *tokens* de entrada, o que melhora a capacidade do *transformer* de texto.

(b) As legendas vistas em treino são frases completas, então frases bem formadas ficam mais próximas dessa distribuição que palavras isoladas.

(c) O *prompt* introduz *tokens* de pontuação que estabilizam a normalização L2 no espaço de projeção.

(d) A média sobre múltiplos *tokens* reduz a variância da estimativa de temperatura τ.

XI) Considere quatro estratégias para adaptar o CLIP a uma tarefa específica: (1) *full fine-tuning*, (2) *linear probe*, (3) *adapter modules*, (4) LoRA. Qual delas mantém o *backbone* congelado e treina apenas matrizes de baixo posto inseridas em camadas selecionadas?

(a) *Full fine-tuning*, que ajusta todos os pesos da rede.

(b) *Linear probe*, que aprende uma única camada linear sobre as *features*.

(c) *Adapter modules*, que adicionam pequenos blocos com não linearidade entre as camadas.

(d) LoRA, que adiciona produtos $AB^{\top}$ de baixo posto às matrizes de pesos congeladas.

XII) Sobre as limitações conhecidas do CLIP, qual alternativa identifica corretamente um cenário em que o modelo apresenta desempenho notavelmente fraco?

(a) Recuperação semântica em grandes coleções de imagens a partir de consultas textuais curtas.

(b) Contagem precisa do número de objetos presentes em uma cena complexa.

(c) Classificação *zero-shot* de classes amplas presentes em `ImageNet`.

(d) Discriminação entre imagens fotográficas e desenhos esquemáticos da mesma categoria.

Questões Dissertativas.

I) ☠️ (**Perda *InfoNCE***) Considere um *batch* com $N = 2$ pares e a matriz de similaridades

$$S = \begin{pmatrix} 0.96 & 0.64 \\ 0.36 & 0.96 \end{pmatrix}.$$

(a) Escreva a expressão de $\mathcal{L}_{i\to t}$ em função de S e τ e explique, em uma frase, qual é o papel da temperatura.

(b) Calcule $\mathcal{L}_{i\to t}$ para $\tau = 1.0$ e para $\tau = 0.1$. Apresente os valores das duas linhas de $\mathrm{softmax}(S/\tau)$ e a perda final em cada caso.

(c) Comente o que acontece com o gradiente associado ao positivo da primeira linha à medida que τ diminui, e por que isso justifica o *clipping* de $1/\tau$ em $\log 100$ feito pelos autores.

II) (**Pipeline *zero-shot***) Você precisa classificar imagens em três espécies de plantas: *samambaia*, *cacto* e *orquídea*, sem dispor de dados rotulados. Descreva, passo a passo, como construir o classificador usando o CLIP pré-treinado. Sua resposta deve cobrir:

(a) A escolha de um *template* de *prompt* e o motivo pelo qual ele tende a superar o uso do nome da classe isolado.

(b) Como obter os vetores de cada classe a partir do *encoder* de texto e o que fazer com a normalização.

(c) Como decidir a classe para uma imagem nova e como tratar empates ou similaridades muito próximas (por exemplo, indicar abstenção ou usar um *ensemble* de *prompts*).

III) (**Recuperação semântica**) Descreva como utilizar embeddings do CLIP para construir um motor de busca sobre uma coleção de 1 000 000 de imagens, considerando os seguintes pontos:

(a) Etapa *offline*: o que pré-computar e armazenar para cada imagem, e qual estrutura de índice é razoável para consultas rápidas.

(b) Etapa *online*: como a consulta em texto é transformada em vetor, qual métrica é usada e como obter os k resultados mais próximos.

(c) Discuta o comportamento esperado quando a consulta é uma frase descritiva longa em comparação com uma única palavra, considerando o limite de 76 *tokens* do *encoder* de texto e a distribuição das legendas vistas em treino.

IV) (**Vieses e mitigação**) O CLIP foi treinado em pares (imagem, texto) coletados sem curadoria fina da web. Liste dois modos concretos de falha que podem surgir desse processo (por exemplo, associações demográficas indesejadas, reconhecimento privilegiado de marcas, estereótipos profissionais) e proponha um experimento de auditoria. O experimento deve indicar:

(a) Quais *prompts* usar para sondar a falha (incluindo *prompt ensembling*, se aplicável).

(b) Qual métrica reportar, e como interpretar a magnitude observada.

(c) Uma mitigação possível (curadoria de subconjuntos, calibração de *prompts*, restrição do vocabulário de classes, etc.) e os limites dessa mitigação.

Gabarito

Múltipla Escolha.

I) **(a)** O conjunto de classes do `ImageNet` é fixo e a rotulação manual é cara, então cobrir conceitos fora dessas 1000 classes exige novos rótulos. As outras alternativas descrevem limitações que não correspondem à realidade do paradigma supervisionado.

II) **(c)** A supervisão vem da co-ocorrência natural entre imagem e legenda em páginas web, sem rótulos humanos explícitos. Esse é o mecanismo que viabiliza o conjunto WIT-400M.

III) **(b)** A diagonal contém os N pares verdadeiros, enquanto as $N^2 - N = N(N-1)$ entradas fora da diagonal funcionam como negativos "de graça" obtidos do próprio *batch*.

IV) **(c)** Após a normalização L2 das projeções, o produto interno coincide com a similaridade do cosseno. Sem essa normalização a temperatura τ teria que absorver a escala dos vetores.

V) **(b)** Com $\tau = 1.0$, $\text{softmax}(0.96, 0.64) \approx (0.579, 0.421)$. Com $\tau = 0.1$, $\text{softmax}(9.6, 6.4) \approx (0.961, 0.039)$. Diminuir τ "afia" a distribuição em direção ao maior *logit*.

VI) **(c)** Linha 1: $\text{softmax}(9.6, 6.4) \approx (0.961, 0.039)$, $-\log 0.961 \approx 0.040$. Linha 2: $\text{softmax}(3.6, 9.6) \approx (0.0025, 0.9975)$, $-\log 0.9975 \approx 0.0025$. Média: ≈ 0.021, ou seja, aproximadamente 0.02.

VII) **(b)** Cada modalidade pode ser usada como consulta, então é desejável que recuperar texto a partir de imagem e recuperar imagem a partir de texto produzam gradientes simétricos. Sem a simetrização, o modelo poderia ficar enviesado a uma direção.

VIII) **(c)** Os *encoders* são independentes (uma CNN ou ViT para imagens, um *transformer* para texto), treinados conjuntamente, e só interagem no espaço de projeção comum onde se calcula a similaridade. Não há compartilhamento de pesos.

IX) **(b)** O procedimento padrão é codificar a imagem, codificar as K descrições de classe como texto, calcular as similaridades e tomar $\arg\max$. Esse é o pipeline mostrado no exemplo trabalhado III da aula.

X) **(b)** As legendas vistas em treino são frases completas, então frases construídas com *templates* ficam mais próximas dessa distribuição que palavras isoladas. As demais alternativas invocam mecanismos que não correspondem ao treino do modelo.

XI) **(d)** LoRA insere produtos de baixo posto $AB^\top$ em camadas de atenção ou *feed-forward*, treinando apenas A e B e mantendo o resto congelado. O *linear probe* treina apenas a camada final, *adapters* adicionam blocos não lineares completos, e *full fine-tuning* ajusta todos os pesos.

XII) **(b)** Contagem precisa de objetos é uma falha conhecida do CLIP, junto com leitura de texto em imagens, tarefas de granularidade fina e raciocínio espacial. Recuperação semântica e classificação *zero-shot* em classes amplas estão entre seus pontos fortes.

Dissertativas.

I) (a) $\mathcal{L}_{i\to t} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^{N}\exp(S_{ij}/\tau)}$. A temperatura τ controla a "nitidez" da distribuição: τ pequeno aproxima a softmax do $\arg\max$, τ grande aproxima-a do uniforme.

(b) Para $\tau = 1.0$: linha 1 $\text{softmax}(0.96, 0.64) \approx (0.579, 0.421)$, linha 2 $\text{softmax}(0.36, 0.96) \approx (0.354, 0.646)$, então $\mathcal{L}_{i\to t} \approx \frac{1}{2}(-\log 0.579 - \log 0.646) \approx 0.492$. Para $\tau = 0.1$: linha 1 $\approx (0.961, 0.039)$, linha 2 $\approx (0.0025, 0.9975)$, então $\mathcal{L}_{i\to t} \approx \frac{1}{2}(0.040 + 0.0025) \approx 0.021$.

(c) Quando $\tau \to 0$, a softmax concentra toda a massa no maior *logit*, então um negativo enganoso (com S_{ij} próximo ao positivo) gera gradiente extremamente grande na direção de afastar esse par. Para evitar instabilidade numérica e gradientes explosivos, os autores aprendem $\log(1/\tau)$ e o recortam em $\log 100$, ou seja, $\tau \geq 0.01$.

II) (a) Usar um *template* como "*a photo of a* {*classe*}" aproxima a entrada da distribuição de legendas vistas em treino, em que frases completas predominam sobre palavras isoladas. Isso reduz o "gap" entre o treino e a inferência *zero-shot*.

(b) Para cada classe $k \in \{\text{samambaia}, \text{cacto}, \text{orquídea}\}$, gere o *prompt* correspondente, passe-o pelo *encoder* de texto e aplique normalização L2 ao vetor de saída para obter $\tilde{\mathbf{v}}_k$. Esses três vetores podem ser pré-computados uma única vez.

(c) Dada uma imagem nova, codifica-se a imagem (com normalização L2 também), calcula-se a similaridade do cosseno com cada $\tilde{\mathbf{v}}_k$, escala-se por $1/\tau$ e aplica-se softmax. A classe escolhida é o arg max. Para empates ou similaridades muito próximas, pode-se usar *prompt ensembling* (média dos vetores de vários *templates*, como "*a photo of a* {*classe*}", "*a close-up of a* {*classe*}", "*an image of a* {*classe*}") e definir um limiar de confiança mínimo, abaixo do qual o sistema se abstém.

III) (a) Etapa *offline*: para cada imagem da coleção, passar pelo *encoder* de imagem, normalizar L2 e armazenar o vetor (em geral 512 ou 768 dimensões) junto com o identificador da imagem. Para 1M de itens, um índice aproximado como FAISS (HNSW ou IVF-PQ) é razoável, pois busca exata por força bruta ainda é viável mas mais lenta.

(b) Etapa *online*: a consulta textual é passada pelo *encoder* de texto, normalizada L2 e usada como vetor de busca. A métrica é a similaridade do cosseno, equivalente a produto interno sobre vetores normalizados. O índice retorna os k vizinhos com maior similaridade.

(c) Frases descritivas longas tendem a funcionar bem porque se aproximam da distribuição de legendas usadas em treino, desde que respeitem o limite de 76 *tokens*. Palavras isoladas ficam mais distantes dessa distribuição e podem produzir vetores menos discriminativos; o impacto é semelhante ao que motiva o uso de *templates* no *zero-shot*. Consultas que excedem o limite são truncadas, então detalhes ao final da frase são descartados pelo *encoder*.

IV) Exemplo de resposta:

(a) Falhas plausíveis incluem associações demográficas (por exemplo, profissões correlacionadas a gênero ou etnia em *prompts* como "*a photo of a CEO*") e privilégio a marcas conhecidas (logos de grandes empresas reconhecidos com alta confiança, em detrimento de pequenas marcas). Para sondar, monta-se um conjunto de *prompts* pareados que diferem apenas no atributo sensível ("*a photo of a male nurse*" vs. "*a photo of a female nurse*") e usa-se *prompt ensembling* sobre cada lado para reduzir variância de formulação.

(b) Uma métrica útil é a diferença de probabilidade média entre os lados pareados em um conjunto fixo de imagens neutras, ou a taxa de classificação correta condicionada ao atributo sensível. Magnitudes pequenas e consistentemente próximas de zero indicam pouca dependência; diferenças sistemáticas grandes indicam viés que merece intervenção.

(c) Mitigações incluem curadoria do conjunto de classes (remover ou unir rótulos com forte associação demográfica), calibração de *prompts* (subtrair a contribuição de *prompts* neutros) e restrição do vocabulário em produção. Os limites: essas mitigações atuam na inferência e não corrigem o viés latente dos pesos, então casos novos de associação podem aparecer fora do conjunto auditado.