

Lista de Exercícios 3

Fundamentos da Inteligência Artificial Generativa

Conteúdos: Métricas de avaliação, falhas de GANs, tokenização, aprendizado com poucos exemplos, espaço latente e modelos multimodais — aprofundamento

Nome: _____ Data: ____/____/____

1. A perplexidade é uma métrica usada para avaliar modelos de linguagem, relacionada à probabilidade que o modelo atribui às palavras corretas de um texto de teste; de forma simplificada, pode ser aproximada por $PPL \approx 1/p$, sendo p a probabilidade média atribuída pelo modelo à palavra correta em cada posição do texto. Um modelo de linguagem atribui, em média, probabilidade de 0,20 à palavra correta em um texto de teste. Calcule a perplexidade aproximada desse modelo, e explique se um valor menor de perplexidade indica um modelo melhor ou pior.

Fórmula(s): $PPL \approx 1 / p$ (perplexidade aproximada)

2. Explique o que é o problema de 'mode collapse' (colapso de modos) em uma GAN, e por que ele compromete a qualidade e a diversidade das imagens (ou outros dados) geradas pelo modelo.

Fórmula(s): Questão teórica — mode collapse em GANs

3. Explique o que é a tokenização de texto em um modelo de linguagem, e por que os modelos processam o texto dividido em tokens (que podem ser palavras inteiras, partes de palavras ou caracteres) em vez de processá-lo caractere por caractere. Um texto tem, em média, 4 caracteres por token. Calcule aproximadamente quantos tokens são necessários para representar um texto de 2.000 caracteres.

Fórmula(s): $tokens \approx caracteres / 4$

4. Explique a diferença entre 'zero-shot learning' e 'few-shot learning' no contexto de modelos de IA Generativa, dando um exemplo de como um prompt seria formulado em cada uma dessas duas abordagens para pedir ao modelo que classifique o sentimento (positivo ou negativo) de uma frase.

Fórmula(s): Questão teórica — zero-shot × few-shot learning

5. Explique o que é o 'espaço latente' (latent space) em modelos generativos como GANs, autoencoders e modelos de difusão, e qual é o seu papel no processo de geração de novos dados, como imagens.

Fórmula(s): Questão teórica — espaço latente em modelos generativos

6. Uma rede neural simples tem um neurônio na camada oculta que recebe uma única entrada $x = 1,0$, com peso $w_1 = 0,6$ e bias $b_1 = 0,1$, produzindo uma saída h (soma ponderada, sem função de ativação não linear nesta questão). Em seguida, h é usada como entrada de um neurônio de saída, com peso $w_2 = 0,8$ e bias $b_2 = -0,05$. Calcule o valor da saída final desse neurônio de saída.

Fórmula(s): $h = x \cdot w_1 + b_1$; saída = $h \cdot w_2 + b_2$

7. Um modelo de difusão gera uma imagem realizando 50 etapas sucessivas de remoção de ruído (denoising), sendo que cada etapa leva, em média, 0,2 segundos para ser processada. Calcule o tempo total necessário para gerar uma única imagem com esse modelo, e explique, de forma geral, por que aumentar o número de etapas de denoising tende a melhorar a qualidade da imagem gerada, mas também aumenta o tempo total de geração.

Fórmula(s): $\text{Tempo total} = \text{número de etapas} \times \text{tempo por etapa}$

8. Uma camada de uma rede neural tem 10 neurônios de entrada totalmente conectados a uma camada seguinte com 5 neurônios. Calcule o número total de pesos (parâmetros ajustáveis) existentes entre essas duas camadas, sem contar os bias, e explique por que redes neurais maiores (com mais neurônios e mais camadas) geralmente exigem mais dados e mais poder computacional para serem treinadas.

Fórmula(s): $\text{Número de pesos} = \text{neurônios de entrada} \times \text{neurônios de saída}$

9. Explique o que é a técnica RAG (Retrieval-Augmented Generation, ou 'geração aumentada por recuperação'), descrevendo as duas etapas principais desse processo, e por que essa técnica ajuda a reduzir o problema das alucinações em modelos de IA Generativa de texto.

Fórmula(s): *Questão teórica — RAG (recuperação + geração)*

10. Explique o que são modelos de IA Generativa multimodais, capazes de processar e/ou gerar mais de um tipo de dado (como texto, imagem e áudio) simultaneamente, e dê um exemplo de uma aplicação prática que se beneficia dessa capacidade multimodal.

Fórmula(s): *Questão teórica — modelos de IA Generativa multimodais*
