PTIC — 1º Ano

3º Bimestre — 2026

Gabarito 2 com Resolução

Fundamentos da Inteligência Artificial Generativa

Conteúdos: Arquiteturas generativas, redes neurais e mecanismos de atenção — aprofundamento técnico e cálculos

---

1. O modelo autoregressivo gera texto de forma sequencial, prevendo a próxima palavra (token) com base em todas as palavras já geradas anteriormente, um passo de cada vez. O modelo de difusão, usado em geração de imagens, começa com uma imagem de ruído aleatório e aplica um processo iterativo de 'remoção de ruído' guiado por um prompt, refinando progressivamente a imagem até que ela corresponda à descrição desejada — o processo é global (ajusta a imagem inteira a cada passo) em vez de sequencial elemento por elemento.

2. $z = (0{,}3\times0{,}5) + (0{,}7\times0{,}2) + (0{,}2\times0{,}9) + (-0{,}1) = 0{,}15 + 0{,}14 + 0{,}18 - 0{,}1 = 0{,}37$

3. $f(0{,}37) = 1/(1+e^{-0{,}37}) \approx 1/(1+0{,}691) \approx 1/1{,}691 \approx 0{,}591$

4. Acurácia = 520/1000 = 52%. Uma acurácia próxima de 50% significa que o discriminador está tendo desempenho próximo ao de uma escolha aleatória (chute), ou seja, está tendo grande dificuldade em distinguir imagens reais das geradas pelo gerador — isso geralmente indica que o gerador da GAN está produzindo imagens sintéticas de alta qualidade, muito próximas das reais.

5. Um embedding é uma representação numérica (um vetor de números) de uma palavra, frase ou outro dado, em um espaço multidimensional, aprendida de forma que capture relações semânticas entre os elementos. Palavras com significados semelhantes (como 'rei' e 'monarca') tendem a aparecer em contextos parecidos durante o treinamento, e o modelo aprende a posicionar seus vetores de embedding próximos nesse espaço, permitindo que operações matemáticas (como distância entre vetores) reflitam similaridade de significado.

6. A arquitetura Transformer processa todos os tokens de uma sequência simultaneamente, usando mecanismos de atenção para relacionar cada token a todos os outros diretamente, o que permite alto grau de paralelização durante o treinamento e melhor captura de dependências de longo alcance no texto. As RNNs processam a sequência token a token, de forma estritamente sequencial, o que dificulta a paralelização (cada passo depende do anterior) e torna mais difícil capturar relações entre palavras muito distantes entre si no texto.

7. Palavras ≈ 8000 × 0,75 = 6000 palavras

8. A temperatura é um parâmetro que controla o grau de aleatoriedade na escolha da próxima palavra gerada pelo modelo. Uma temperatura baixa (próxima de 0) faz o modelo escolher quase sempre a palavra mais provável, gerando respostas mais previsíveis, repetitivas e conservadoras. Uma temperatura mais alta aumenta a chance de escolher palavras menos prováveis, tornando as respostas mais variadas, criativas e, em excesso, potencialmente incoerentes.

9. Fine-tuning é o processo de continuar o treinamento de um modelo já pré-treinado (que aprendeu padrões gerais de linguagem ou imagem a partir de um grande volume de dados) usando um conjunto de dados menor e específico de uma tarefa ou domínio particular, ajustando seus parâmetros para essa aplicação. Uma vantagem é que esse processo exige muito menos dados, tempo e poder computacional do que treinar um modelo do zero, já que o modelo parte de um conhecimento geral já consolidado.

10. Alucinações são respostas geradas por um modelo de IA Generativa que soam coerentes e convincentes, mas que são factualmente incorretas, inventadas ou não fundamentadas nos dados reais (por exemplo, o modelo citar uma fonte ou um evento que não existe). Uma estratégia para reduzir esse problema é fornecer ao modelo informações de contexto confiáveis diretamente no prompt (ou usar técnicas como RAG — Retrieval-Augmented Generation, que busca informações em fontes verificadas antes de gerar a resposta), além de sempre verificar/checar as informações geradas antes de utilizá-las.