FAQ — Núcleo do Transformer e Variações Modernas

BERT, Atenção, RoPE, Scaling, Temperatura e Mixture of Experts

Introdução

Este documento consolida dúvidas frequentes surgidas durante o estudo prático do BERT e das variações modernas do Transformer. As respostas são fundamentadas em implementação real, geometria vetorial e análise matemática do mecanismo de atenção, servindo como material de apoio técnico para cursos, livros e repositórios públicos.

1 Input Embeddings no BERT

1) Para gerar o input embedding no BERT, após gerar os tensores de token, posição e segmento, o modelo soma todos?

Sim.

No BERT, o vetor de entrada associado a cada token é obtido pela soma elemento a elemento de:

- embedding lexical (token);
- embedding posicional;
- embedding de segmento (token type).

Formalmente, para cada posição i:

$$\mathbf{x}_i = \mathbf{e}_i^{(\text{tok})} + \mathbf{e}_i^{(\text{pos})} + \mathbf{e}_i^{(\text{seg})}, \quad \mathbf{x}_i \in \mathbb{R}^{768}.$$

Essa soma ocorre antes da primeira camada de self-attention. Após isso, aplica-se *Layer Normalization* e *dropout*, mas ainda não existe interação entre tokens.

2) Se os embeddings já incluem posição e segmento, isso não torna os tokens contextuais?

Não.

Apesar da soma com posição e segmento, cada token ainda é tratado de forma independente. Tokens idênticos em posições diferentes terão vetores de entrada distintos, mas não trocam informação.

A contextualização surge exclusivamente quando a self-attention mistura informações entre diferentes posições da sequência.

2 Tokens Especiais e Geometria

3) Por que o token `[CLS]` aparece distante na PCA enquanto `[SEP]` fica próximo de outros tokens?

Esse comportamento é esperado e decorre de três fatores:

1. **Tokens especiais não são outliers geométricos por definição.**
   Ambos possuem embeddings aprendidos como qualquer outro token.

2. **Funções diferentes implicam distribuições diferentes.**
   O token `[CLS]` é treinado como agregador global da sequência, enquanto `[SEP]` atua como marcador estrutural.

3. **PCA é uma projeção de alta para baixa dimensão.**
   Distâncias não são preservadas fielmente ao projetar 768 dimensões em 2.

Tokens são especiais por função no treinamento, não por obrigação geométrica.

3 Rotação, Posição e Seleção

4) Um token foi rotacionado porque foi definido `top-k = 5`?

Não.

São conceitos distintos:

- RoPE (Rotary Position Embeddings) aplica rotações baseadas na posição do token.
- Top-k é usado apenas para seleção (pesos, logits ou especialistas).

Top-k não altera embeddings. RoPE não depende de top-k.

4 Embeddings Posicionais Modernos

5) Qual a diferença entre embeddings posicionais aditivos e RoPE?

No BERT clássico, a posição é incorporada por soma:

$$\mathbf{X} = \mathbf{E}_{\text{token}} + \mathbf{E}_{\text{pos}}.$$

No RoPE, a posição atua como operador geométrico, rotacionando consultas e chaves:

$$\mathbf{Q} \to \mathbf{Q}_r, \quad \mathbf{K} \to \mathbf{K}_r.$$

A posição emerge no produto escalar, dependendo da diferença relativa entre posições.

6) A rotação do RoPE altera a magnitude dos vetores?

Não.

Rotações preservam norma:

$$\|R(\theta)\mathbf{x}\| = \|\mathbf{x}\|.$$

O que muda é a orientação relativa entre vetores, modulando compatibilidades sem alterar escala.

5 Scaling e Temperatura

7) O que realmente muda nos Transformers modernos?

O núcleo permanece:

$$\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}.$$

O que muda são refinamentos:

- geometria (RoPE);
- escala (attention scaling e temperatura);
- capacidade (Mixture of Experts).

8) Por que existe a divisão por $\sqrt{d_k}$ na atenção?

A variância dos produtos escalares cresce com a dimensão:

$$\mathbb{V}[\mathbf{q}_i^\top \mathbf{k}_j] \propto d_k.$$

Dividir por $\sqrt{d_k}$ normaliza a escala dos logits, evitando saturação do softmax e garantindo estabilidade numérica.

9) A temperatura controla a "velocidade" da atenção? Existe equação diferencial?

Não.

A temperatura atua diretamente no softmax:

$$\alpha_{ij} = \text{softmax}\left(\frac{Q_i K_j^\top}{\tau}\right).$$

Ela regula a entropia da distribuição, não uma dinâmica temporal. Não há equação diferencial envolvida.

10) A temperatura e o fator $1/\sqrt{d_k}$ fazem a mesma coisa?

Não exatamente.

- $1/\sqrt{d_k}$: correção fixa e necessária para estabilidade.
- Temperatura τ: controle contínuo do regime do softmax.

6 Mixture of Experts

11) O que é Mixture of Experts (MoE) e qual o papel do top-k?

MoE aumenta capacidade sem custo linear por token.

Cada token ativa apenas k especialistas:

$$\text{FFN}_{\text{MoE}}(\mathbf{x}) = \sum_{e \in \mathcal{E}(\mathbf{x})} g_e(\mathbf{x}) \cdot \text{Expert}_e(\mathbf{x}).$$

Aqui, top-k é roteamento seletivo, não transformação geométrica.

7 Embeddings de Sentença

12) O token `[CLS]` é sempre um bom embedding de sentença?

Depende da tarefa.

No BERT original, `[CLS]` é treinado para classificação, mas não como embedding semântico universal. Para busca semântica, modelos especializados costumam funcionar melhor.

13) O SBERT será abordado neste curso?

Não diretamente.

O foco é o núcleo do Transformer. O SBERT é uma adaptação voltada a embeddings de sentença comparáveis, sendo uma extensão natural para tarefas de recuperação semântica.

Resumo Final

A atenção permanece a mesma;
o que muda nos modelos modernos é a geometria (RoPE),
a escala (scaling e temperatura)
e a densidade de parâmetros por token (MoE).