Attention Models: da motivação às variantes modernas

Intuição geométrica, formulação matemática e aplicações

Seu Nome

August 12, 2025

Table of Contents

Roteiro da Apresentação

1 Roteiro

- Motivação e histórico

Roteiro da Apresentação

1 Roteiro

- Motivação e histórico
- Modelos de sequência (RNN/LSTM) + Attention

Roteiro da Apresentação

1 Roteiro

- Motivação e histórico
- Modelos de sequência (RNN/LSTM) + Attention
- Embeddings, interpretação geométrica e matemática

Roteiro da Apresentação

1 Roteiro

- Motivação e histórico
- Modelos de sequência (RNN/LSTM) + Attention
- Embeddings, interpretação geométrica e matemática
- Transformers e posicionais

Roteiro da Apresentação

1 Roteiro

- Motivação e histórico
- Modelos de sequência (RNN/LSTM) + Attention
- Embeddings, interpretação geométrica e matemática
- Transformers e posicionais
- Aplicações: Séries Temporais, Language Models, ViT, GAT

Roteiro da Apresentação

1 Roteiro

- Motivação e histórico
- Modelos de sequência (RNN/LSTM) + Attention
- Embeddings, interpretação geométrica e matemática
- Transformers e posicionais
- Aplicações: Séries Temporais, Language Models, ViT, GAT
- Treinamento, variantes e hiperparâmetros

Roteiro da Apresentação

1 Roteiro

- Motivação e histórico
- Modelos de sequência (RNN/LSTM) + Attention
- Embeddings, interpretação geométrica e matemática
- Transformers e posicionais
- Aplicações: Séries Temporais, Language Models, ViT, GAT
- Treinamento, variantes e hiperparâmetros
- Tamanho ideal (scaling), Mixture of Experts

Roteiro da Apresentação

1 Roteiro

- Motivação e histórico
- Modelos de sequência (RNN/LSTM) + Attention
- Embeddings, interpretação geométrica e matemática
- Transformers e posicionais
- Aplicações: Séries Temporais, Language Models, ViT, GAT
- Treinamento, variantes e hiperparâmetros
- Tamanho ideal (scaling), Mixture of Experts
- Pruning e Distillation

Roteiro da Apresentação

1 Roteiro

- Motivação e histórico
- Modelos de sequência (RNN/LSTM) + Attention
- Embeddings, interpretação geométrica e matemática
- Transformers e posicionais
- Aplicações: Séries Temporais, Language Models, ViT, GAT
- Treinamento, variantes e hiperparâmetros
- Tamanho ideal (scaling), Mixture of Experts
- Pruning e Distillation
- Limitações, diagnóstico e conclusões

Motivação & Histórico

1 Roteiro

Table of Contents

Por que Atenção?

2 Motivação

- Capturar dependências longas e filtrar informação relevante

Por que Atenção?

2 Motivação

- Capturar dependências longas e filtrar informação relevante
- RNNs sofrem com gradientes e paralelização limitada

Por que Atenção?

2 Motivação

- Capturar dependências longas e filtrar informação relevante
- RNNs sofrem com gradientes e paralelização limitada
- Pesos de atenção ajudam na interpretabilidade

Por que Atenção?

2 Motivação

- Capturar dependências longas e filtrar informação relevante
- RNNs sofrem com gradientes e paralelização limitada
- Pesos de atenção ajudam na interpretabilidade
- Melhor relação capacidade/compute em diversas tarefas

Histórico em 3 passos

- **Additive Attention** (Bahdanau) para NMT

Histórico em 3 passos

- **Additive Attention** (Bahdanau) para NMT
- **Dot-Product Attention** (Luong) — caminho para escalabilidade

Histórico em 3 passos

2 Motivação

- **Additive Attention** (Bahdanau) para NMT
- **Dot-Product Attention** (Luong) — caminho para escalabilidade
- **Transformers** (Vaswani et al., 2017): “Attention is All You Need”

Fundamentos

2 Motivação

Table of Contents

RNN/GRU/LSTM — limites práticos

- Cálculo sequencial $\Rightarrow$ baixa paralelização

RNN/GRU/LSTM — limites práticos

- Cálculo sequencial $\Rightarrow$ baixa paralelização
- Dependências longas: vanish/explode (mitigado, não resolvido)

RNN/GRU/LSTM — limites práticos

- Cálculo sequencial $\Rightarrow$ baixa paralelização
- Dependências longas: vanish/explode (mitigado, não resolvido)
- Memória finita e custo de treino elevado para contextos longos

LSTM + Attention (Encoder–Decoder)

Atenção aditiva (Bahdanau)

Equações

$$e_{ij} = v^{\top} \tanh\big(W_1 h_i + W_2 s_j\big), \quad \alpha_{ij} = \mathrm{softmax}_i(e_{ij}), \quad c_j = \sum_i \alpha_{ij} h_i$$

- **Alinhamento** dinâmico entre estados do encoder e passos do decoder

LSTM + Attention (Encoder–Decoder)

Atenção aditiva (Bahdanau)

Equações

$$e_{ij} = v^{\top} \tanh\big(W_1 h_i + W_2 s_j\big), \quad \alpha_{ij} = \mathrm{softmax}_i(e_{ij}), \quad c_j = \sum_i \alpha_{ij} h_i$$

- **Alinhamento** dinâmico entre estados do encoder e passos do decoder
- Reduz gargalos de um único vetor de contexto

Table of Contents

4 Embeddings & Interpretações

Word/Subword Embeddings

4 Embeddings & Interpretações

- Estáticos (Word2Vec, GloVe) vs. Contextuais (ELMo, BERT)

Word/Subword Embeddings

4 Embeddings & Interpretações

- Estáticos (Word2Vec, GloVe) vs. Contextuais (ELMo, BERT)
- Subword (BPE/Unigram) para robustez morfológica

Word/Subword Embeddings

4 Embeddings & Interpretações

- Estáticos (Word2Vec, GloVe) vs. Contextuais (ELMo, BERT)
- Subword (BPE/Unigram) para robustez morfológica
- Análogos em outras modalidades: patches (ViT), time2vec (TS), node2vec (grafos)

Interpretação Geométrica da Atenção

- Q, K, V são projeções lineares do latente

Interpretação Geométrica da Atenção

4 Embeddings & Interpretações

- Q, K, V são projeções lineares do latente
- Similaridade (coseno/dot) guia uma *combinação convexa* de V

Interpretação Geométrica da Atenção

- Q, K, V são projeções lineares do latente
- Similaridade (coseno/dot) guia uma *combinação convexa* de V
- Multi-head $\Rightarrow$ múltiplas métricas/projeções simultâneas

Interpretação Matemática: Self-Attention

4 Embeddings & Interpretações

Fórmula central (scaled dot-product)

$$\mathrm{Att}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

- Máscara causal/atencional conforme a tarefa

Interpretação Matemática: Self-Attention

4 Embeddings & Interpretações

Fórmula central (scaled dot-product)

$$\mathrm{Att}(Q,K,V) \;=\; \mathrm{softmax}\!\left(\frac{QK^{\top}}{\sqrt{d_k}}\right) V$$

- Máscara causal/atencional conforme a tarefa
- Complexidade $\mathcal{O}(n^2)$ em tempo/memória

Interpretação Matemática: Self-Attention

4 Embeddings & Interpretações

Fórmula central (scaled dot-product)

$$\mathrm{Att}(Q,K,V) \;=\; \mathrm{softmax}\!\left(\frac{QK^{\top}}{\sqrt{d_k}}\right) V$$

- Máscara causal/atencional conforme a tarefa
- Complexidade $\mathcal{O}(n^2)$ em tempo/memória
- Gradientes e saturação da softmax

Positional Encodings

4 Embeddings & Interpretações

- Absolutos senoidais vs. aprendidos

Positional Encodings

4 Embeddings & Interpretações

- Absolutos senoidais vs. aprendidos
- Relativos e RoPE (rotary): melhor generalização/composição

Positional Encodings

4 Embeddings & Interpretações

- Absolutos senoidais vs. aprendidos
- Relativos e RoPE (rotary): melhor generalização/composição
- Impacto em extrapolação e contextos longos

Transformers

4 Embeddings & Interpretações

Table of Contents

Arquiteturas Transformer

5 Arquiteturas & Variantes

- **Encoder–Decoder**, **Encoder-only**, **Decoder-only**

Arquiteturas Transformer

5 Arquiteturas & Variantes

- **Encoder–Decoder**, **Encoder-only**, **Decoder-only**
- Bloco: MHSA + FFN, *residual* + *layer norm*

Arquiteturas Transformer

5 Arquiteturas & Variantes

- **Encoder–Decoder**, **Encoder-only**, **Decoder-only**
- Bloco: MHSA + FFN, *residual* + *layer norm*
- Pré-norm vs. pós-norm (estabilidade de treino)

Atenções Eficientes

5 Arquiteturas & Variantes

- **Esparsidade**: Longformer/BigBird (padrões locais+globais)

Trade-off

Complexidade $\mathcal{O}(n^2) \rightarrow \tilde{\mathcal{O}}(n)$ com perda controlada de exatidão e/ou restrições de padrão.

Atenções Eficientes

5 Arquiteturas & Variantes

- **Esparsidade**: Longformer/BigBird (padrões locais+globais)
- **Aproximação**: Linformer, Nyströmformer

Trade-off

Complexidade $\mathcal{O}(n^2) \rightarrow \tilde{\mathcal{O}}(n)$ com perda controlada de exatidão e/ou restrições de padrão.

Atenções Eficientes

5 Arquiteturas & Variantes

- **Esparsidade**: Longformer/BigBird (padrões locais+globais)
- **Aproximação**: Linformer, Nyströmformer
- **Kernels**: Performer (favor+), FlashAttention (IO-aware)

Trade-off

Complexidade $\mathcal{O}(n^2) \rightarrow \tilde{\mathcal{O}}(n)$ com perda controlada de exatidão e/ou restrições de padrão.

Atenções Eficientes

5 Arquiteturas & Variantes

- **Esparsidade**: Longformer/BigBird (padrões locais+globais)
- **Aproximação**: Linformer, Nyströmformer
- **Kernels**: Performer (favor+), FlashAttention (IO-aware)
- **Outros**: Reformer (LSH)

Trade-off

Complexidade $\mathcal{O}(n^2) \rightarrow \tilde{\mathcal{O}}(n)$ com perda controlada de exatidão e/ou restrições de padrão.

Table of Contents

Práticas de Treino

6 Treino & Hiperparâmetros

- AdamW, warmup + decaimento; label smoothing quando aplicável

Práticas de Treino

6 Treino & Hiperparâmetros

- AdamW, warmup + decaimento; label smoothing quando aplicável
- Regularização: dropout, stochastic depth, weight decay

Práticas de Treino

6 Treino & Hiperparâmetros

- AdamW, warmup + decaimento; label smoothing quando aplicável
- Regularização: dropout, stochastic depth, weight decay
- AMP/mixed precision, grad clipping, checkpointing

Práticas de Treino

6 Treino & Hiperparâmetros

- AdamW, warmup + decaimento; label smoothing quando aplicável
- Regularização: dropout, stochastic depth, weight decay
- AMP/mixed precision, grad clipping, checkpointing
- Dados: curriculum, masking, augmentation (TS/ViT/GAT)

Hiperparâmetros Essenciais

- Profundidade, d_{model}, #heads, d_{ff}, dropout

Hiperparâmetros Essenciais

- Profundidade, d_{model}, #heads, d_{ff}, dropout
- Comprimento de contexto, batch size, *LR schedule*

Hiperparâmetros Essenciais

6 Treino & Hiperparâmetros

- Profundidade, d_{model}, #heads, d_{ff}, dropout
- Comprimento de contexto, batch size, *LR schedule*
- Específicos: tokenização (LM), *patch size* (ViT), janela/patch (TS)

Qual tamanho ideal? (Scaling)

- Leis de escala vs. limite de dados/compute

Qual tamanho ideal? (Scaling)

- Leis de escala vs. limite de dados/compute
- Tokens vs. parâmetros; saturação com contexto

Qual tamanho ideal? (Scaling)

- Leis de escala vs. limite de dados/compute
- Tokens vs. parâmetros; saturação com contexto
- Regra prática: dimensione para o dataset e o *budget* de inferência

Aplicações

6 Treino & Hiperparâmetros

Table of Contents

Atenção em Séries Temporais

- Codificação temporal: absolutos/relativos, time2vec, calendários/sazonalidade

Atenção em Séries Temporais

7 Séries Temporais

- Codificação temporal: absolutos/relativos, time2vec, calendários/sazonalidade
- Exógenas e *cross-attention*; *patching* para contextos longos

Atenção em Séries Temporais

7 Séries Temporais

- Codificação temporal: absolutos/relativos, time2vec, calendários/sazonalidade
- Exógenas e *cross-attention*; *patching* para contextos longos
- Tarefas: previsão, imputação, detecção de anomalias

Table of Contents

Language Models (LM)

- Atenção **causal**, *next-token* e *masking*

Language Models (LM)

8 Language Models

- Atenção **causal**, *next-token* e *masking*
- Pré-treino vs. *fine-tuning*; Instrução/LoRA/Adapters

Language Models (LM)

8 Language Models

- Atenção **causal**, *next-token* e *masking*
- Pré-treino vs. *fine-tuning*; Instrução/LoRA/Adapters
- Métricas: perplexidade e tarefas *downstream*

Table of Contents

Vision Transformer (ViT)

- Imagem $\rightarrow$ *patches* + `[CLS]` token

Vision Transformer (ViT)

- Imagem $\rightarrow$ *patches* + `[CLS]` token
- Posicionais 2D; augmentations (RandAug, Mixup/CutMix)

Vision Transformer (ViT)

- Imagem $\rightarrow$ *patches* + `[CLS]` token
- Posicionais 2D; augmentations (RandAug, Mixup/CutMix)
- Transfer: *linear probe* vs. *fine-tune*

Table of Contents

Graph Attention Networks (GAT)

Coeficientes de Atenção (um cabeçalho)

$$\alpha_{ij} = \mathrm{softmax}_j\big(\mathrm{LeakyReLU}\big(a^\top [Wh_i \,||\, Wh_j]\big)\big)$$

- Multi-head; sobre-*smoothing* e escalabilidade

Graph Attention Networks (GAT)

Coeficientes de Atenção (um cabeçalho)

$$\alpha_{ij} = \mathrm{softmax}_j\big(\mathrm{LeakyReLU}\big(a^\top [Wh_i \,||\, Wh_j]\big)\big)$$

- Multi-head; sobre-*smoothing* e escalabilidade
- Heterógrafos e atenção relacional

Escala & Compressão

10 Graph Attention

COPPE 60 anos UFRJ

Table of Contents

Mixture of Experts (MoE)

- *Gating* top-1/top-2; balanceamento de carga

Mixture of Experts (MoE)

11 MoE, Pruning, Distillation

- *Gating* top-1/top-2; balanceamento de carga
- Roteamento esparso: capacidade vs. ociosidade

Mixture of Experts (MoE)

11 MoE, Pruning, Distillation

- *Gating* top-1/top-2; balanceamento de carga
- Roteamento esparso: capacidade vs. ociosidade
- Custos de comunicação e estabilidade no treino

Pruning

- Não estruturado (magnitude) vs. estruturado (canal/bloco, n:m)

Pruning

- Não estruturado (magnitude) vs. estruturado (canal/bloco, n:m)
- Iterativo vs. *one-shot*; impacto em latência real

Pruning

11 MoE, Pruning, Distillation

- Não estruturado (magnitude) vs. estruturado (canal/bloco, n:m)
- Iterativo vs. *one-shot*; impacto em latência real
- Interação com quantização e *sparsity-aware kernels*

Distillation

Perda típica (Hinton)

$$\mathcal{L} = (1-\lambda)\,\mathrm{CE}(y, s) + \lambda\, T^2\, \mathrm{KL}\big(p_T \,\|\, q_T\big)$$

- *Teacher* $\rightarrow$ *student*; temperatura T

Distillation

Perda típica (Hinton)

$$\mathcal{L} = (1-\lambda)\,\mathrm{CE}(y, s) + \lambda\, T^2\, \mathrm{KL}\big(p_T \,\|\, q_T\big)$$

- *Teacher* $\rightarrow$ *student*; temperatura T
- *Intermediate layer hints*; *task-specific* vs. generalista

Distillation

Perda típica (Hinton)

$$\mathcal{L} = (1-\lambda)\,\mathrm{CE}(y, s) + \lambda\, T^2\, \mathrm{KL}\big(p_T \,\|\, q_T\big)$$

- *Teacher* $\rightarrow$ *student*; temperatura T
- *Intermediate layer hints*; *task-specific* vs. generalista
- Benefícios: latência/energia e *edge deployment*

Interpretação & Limitações

11 MoE, Pruning, Distillation

Table of Contents

Interpretação & Diagnóstico

12 Diagnóstico & Limites

- *Attention rollout/flow*; atenção $\neq$ causalidade

Interpretação & Diagnóstico

12 Diagnóstico & Limites

- *Attention rollout/flow*; atenção $\neq$ causalidade
- *Probing* de camadas; ablação de cabeças

Interpretação & Diagnóstico

12 Diagnóstico & Limites

- *Attention rollout/flow*; atenção $\neq$ causalidade
- *Probing* de camadas; ablação de cabeças
- Ferramentas de *explainability* por domínio

Limitações & Trade-offs

12 Diagnóstico & Limites

- Custo $\mathcal{O}(n^2)$, viés de dados, *OOD* robustness

Limitações & Trade-offs

12 Diagnóstico & Limites

- Custo $\mathcal{O}(n^2)$, viés de dados, *OOD* robustness
- Contaminação de treino e privacidade

Limitações & Trade-offs

12 Diagnóstico & Limites

- Custo $\mathcal{O}(n^2)$, viés de dados, *OOD* robustness
- Contaminação de treino e privacidade
- Energia/carbono e restrições de hardware

Table of Contents

Conclusões

13 Conclusões

- Atenção como princípio unificador em sequência, visão e grafos

Conclusões

13 Conclusões

- Atenção como princípio unificador em sequência, visão e grafos
- Escolha guiada por dados, métricas e orçamento (treino/inferência)

Conclusões

13 Conclusões

- Atenção como princípio unificador em sequência, visão e grafos
- Escolha guiada por dados, métricas e orçamento (treino/inferência)
- Escala e compressão (MoE/pruning/distillation) para produção

Attention Models: da motivação às variantes modernas

Obrigado pela Atenção!
Alguma Pergunta?

Natanael Moura Junior

natmourajr@poli.ufrj.br, natmourajr@lps.ufrj.br

Perguntas?

13 Conclusões

Obrigado!

Table of Contents

Equações úteis (resumo)

14 Backup (Opcional)

- Scaled dot-product, aditiva (Bahdanau)
- Atenção relativa e RoPE
- GAT detalhado; máscaras causais

Hiperparâmetros por tarefa (resumo)

14 Backup (Opcional)

- LM: contexto, BPE, heads/profundidade típicos
- ViT: patch size, MLP ratio, augmentations
- TS: janela, patching, covariáveis, perdas (MAE/MSE/Quantile)

Table of Contents

Referências Bibliográficas

15 Referências Bibliográficas