Modelos Generativos Profundos

Clase 7: Arquitectura Transformer y LLMs (parte I)

Fernando Fêtis Riquelme

Otoño, 2025

Facultad de Ciencias Físicas y Matemáticas
Universidad de Chile

Clase de hoy

Variaciones en la arquitectura Transformer

Entrenamiento y generación

Variaciones en la arquitectura Transformer

Variaciones en la arquitectura Transformer

- **Tokenización:** tokens espaciales, problema de tokenización, BPE, SentencePiece.
- **Positional encoding:** absoluto, relativo, sinusoidal, RoPE, inducido por masking.
- **Mecanismos de normalización:** LayerNorm, RMSNorm.
- **Mecanismos de atención:** atención cruzada, atención de Badhanau, scaled dot-product, Sliding window attention, atención diferencial.
- **Bloque feed forward:** función de activación, MoE.
- Conexiones residuales y dropout.
- **Arquitectura original:** seq2seq.

Variaciones en la arquitectura Transformer

- **Tokenización:** tokens espaciales, problema de tokenización, BPE, SentencePiece.
- **Positional encoding:** absoluto, relativo, sinusoidal, RoPE, inducido por masking.
- **Mecanismos de normalización:** LayerNorm, RMSNorm.
- **Mecanismos de atención:** atención cruzada, atención de Badhanau, scaled dot-product, Sliding window attention, atención diferencial.
- **Bloque feed forward:** función de activación, MoE.
- Conexiones residuales y dropout.
- **Arquitectura original:** seq2seq.

Variaciones en la arquitectura Transformer

- **Tokenización:** tokens espaciales, problema de tokenización, BPE, SentencePiece.
- **Positional encoding:** absoluto, relativo, sinusoidal, RoPE, inducido por masking.
- **Mecanismos de normalización:** LayerNorm, RMSNorm.
- **Mecanismos de atención:** atención cruzada, atención de Badhanau, scaled dot-product, Sliding window attention, atención diferencial.
- **Bloque feed forward:** función de activación, MoE.
- Conexiones residuales y dropout.
- **Arquitectura original:** seq2seq.

Variaciones en la arquitectura Transformer

- **Tokenización:** tokens espaciales, problema de tokenización, BPE, SentencePiece.
- **Positional encoding:** absoluto, relativo, sinusoidal, RoPE, inducido por masking.
- **Mecanismos de normalización:** LayerNorm, RMSNorm.
- **Mecanismos de atención:** atención cruzada, atención de Badhanau, scaled dot-product, Sliding window attention, atención diferencial.
- **Bloque feed forward:** función de activación, MoE.
- Conexiones residuales y dropout.
- **Arquitectura original:** seq2seq.

Variaciones en la arquitectura Transformer

- **Tokenización:** tokens espaciales, problema de tokenización, BPE, SentencePiece.
- **Positional encoding:** absoluto, relativo, sinusoidal, RoPE, inducido por masking.
- **Mecanismos de normalización:** LayerNorm, RMSNorm.
- **Mecanismos de atención:** atención cruzada, atención de Badhanau, scaled dot-product, Sliding window attention, atención diferencial.
- **Bloque feed forward:** función de activación, MoE.
- Conexiones residuales y dropout.
- **Arquitectura original:** seq2seq.

Variaciones en la arquitectura Transformer

- **Tokenización:** tokens espaciales, problema de tokenización, BPE, SentencePiece.
- **Positional encoding:** absoluto, relativo, sinusoidal, RoPE, inducido por masking.
- **Mecanismos de normalización:** LayerNorm, RMSNorm.
- **Mecanismos de atención:** atención cruzada, atención de Badhanau, scaled dot-product, Sliding window attention, atención diferencial.
- **Bloque feed forward:** función de activación, MoE.
- Conexiones residuales y dropout.
- **Arquitectura original:** seq2seq.

Variaciones en la arquitectura Transformer

- **Tokenización:** tokens espaciales, problema de tokenización, BPE, SentencePiece.
- **Positional encoding:** absoluto, relativo, sinusoidal, RoPE, inducido por masking.
- **Mecanismos de normalización:** LayerNorm, RMSNorm.
- **Mecanismos de atención:** atención cruzada, atención de Badhanau, scaled dot-product, Sliding window attention, atención diferencial.
- **Bloque feed forward:** función de activación, MoE.
- Conexiones residuales y dropout.
- **Arquitectura original:** seq2seq.

Entrenamiento y generación

Entrenamiento y generación

- Función de pérdida.
- **Técnicas de optimización:** Implementaciones eficientes, entrenamiento distribuido, cuantización, heurísticas.
- **Fine tuning:** LoRA, pruning, destilación, instruction fine tuning.
- Train set y evaluación.
- **Inferencia:** beam search, nucleus sampling.

Entrenamiento y generación

- Función de pérdida.
- **Técnicas de optimización:** Implementaciones eficientes, entrenamiento distribuido, cuantización, heurísticas.
- **Fine tuning:** LoRA, pruning, destilación, instruction fine tuning.
- Train set y evaluación.
- **Inferencia:** beam search, nucleus sampling.

Entrenamiento y generación

- Función de pérdida.
- **Técnicas de optimización:** Implementaciones eficientes, entrenamiento distribuido, cuantización, heurísticas.
- **Fine tuning:** LoRA, pruning, destilación, instruction fine tuning.
- Train set y evaluación.
- **Inferencia:** beam search, nucleus sampling.

Entrenamiento y generación

- Función de pérdida.
- **Técnicas de optimización:** Implementaciones eficientes, entrenamiento distribuido, cuantización, heurísticas.
- **Fine tuning:** LoRA, pruning, destilación, instruction fine tuning.
- Train set y evaluación.
- **Inferencia:** beam search, nucleus sampling.

Entrenamiento y generación

- Función de pérdida.
- **Técnicas de optimización:** Implementaciones eficientes, entrenamiento distribuido, cuantización, heurísticas.
- **Fine tuning:** LoRA, pruning, destilación, instruction fine tuning.
- Train set y evaluación.
- **Inferencia:** beam search, nucleus sampling.

Próxima clase

En la próxima clase.

- Propiedades emergentes.
- Scaling laws.
- Prompting.
- Consideraciones éticas.
- Algunos modelos tipo Transformer.

Próxima clase

En la próxima clase.

- Propiedades emergentes.
- Scaling laws.
- Prompting.
- Consideraciones éticas.
- Algunos modelos tipo Transformer.

Próxima clase

En la próxima clase.

- Propiedades emergentes.
- Scaling laws.
- Prompting.
- Consideraciones éticas.
- Algunos modelos tipo Transformer.

Próxima clase

En la próxima clase.

- Propiedades emergentes.
- Scaling laws.
- Prompting.
- Consideraciones éticas.
- Algunos modelos tipo Transformer.

Próxima clase

En la próxima clase.

- Propiedades emergentes.
- Scaling laws.
- Prompting.
- Consideraciones éticas.
- Algunos modelos tipo Transformer.

Modelos Generativos Profundos

Clase 7: Algunas cosas sobre la arquitectura Transformer y los LLMs I