LECTURE 10 | المحاضرة 10

Transformer Models and Attention

نماذج المحوّل وآلية الانتباه

Modern Sequence Modeling for Load, Renewable-Energy, and Grid-Event Forecasting

نمذجة حديثة للسلاسل الزمنية في الأحمال والطاقة المتجددة وأحداث الشبكة

Prepared by | إعداد: Dr.-Ing. Aouss Gabash

Learning Outcomes

مخرجات التعلم

1. Explain attention and apply transformer concepts to sequence data.

   شرح آلية الانتباه وتطبيق مفاهيم المحولات على البيانات المتسلسلة.

2. Interpret the governing mathematical or algorithmic principles.

   تفسير المبادئ الرياضية أو الخوارزمية الحاكمة.

3. Connect the topic to practical electrical-power-engineering applications.

   ربط الموضوع بتطبيقات عملية في هندسة الطاقة الكهربائية.

Prerequisites

المتطلبات السابقة

Lecture 09, vectors, matrices, and sequence modelling.

المحاضرة 09 والمتجهات والمصفوفات ونمذجة التسلسلات.

Transformer Models and Attention | نماذج المحوّل وآلية الانتباه

🎯 Learning Objectives

- Explain why recurrence is not required in Transformer models.
- Derive scaled dot-product attention from queries, keys, and values.
- Understand self-attention, multi-head attention, and positional encoding.
- Describe an encoder block and residual connections.
- Compare Transformers with RNN and LSTM for power-system time series.
- Recognize computational and interpretability limitations.

🎯 أهداف التعلم

- فهم سبب عدم حاجة Transformer إلى التكرار الزمني التقليدي.
- اشتقاق الانتباه النقطي المقاس من Q وK وV.
- فهم الانتباه الذاتي ومتعدد الرؤوس والترميز الموضعي.
- شرح كتلة Encoder والوصلات المتبقية.
- مقارنة Transformer مع RNN وLSTM.
- معرفة القيود الحسابية والتفسيرية.

1. Why Transformers?

All Time Steps → Attention Scores → Weighted Context → Forecast

> Attention allows distant time steps to interact directly instead of passing information sequentially through every intermediate step.

1. لماذا نستخدم Transformer؟

تسمح آلية الانتباه لكل خطوة زمنية بالنظر مباشرة إلى الخطوات الأخرى وتحديد أهميتها.

2. Query, Key, and Value

$$q_t = WQx_t, k_t = WKx_t, v_t = WVx_t$$

Query Q
What information is this step seeking?

Key K
What information does each step offer?

Value V
What content is transferred?

2. الاستعلام والمفتاح والقيمة

يحدد التشابه بين Q وK مقدار الاستفادة من V في بناء السياق.

3. Scaled Dot-Product Attention

$$S = QK^T$$

$$A = softmax(QK^T/\sqrt{dk})$$

$$Attention(Q, K, V) = AV$$

Each row of A sums to one and represents the attention distribution of one time step.

3. الانتباه النقطي المقاس

تُحسب درجات التشابه ثم تُقسم على $\sqrt{d_k}$ وتُمرر عبر Softmax للحصول على أوزان الانتباه.

4. Numerical Example

$$softmax([2, 1, 0]) \approx [0.665, 0.245, 0.090]$$

The first item contributes most strongly to the resulting context.

4. مثال عددي

تحول Softmax الدرجات إلى أوزان موجبة مجموعها واحد، ويكون العنصر الأول الأكثر تأثيرًا.

5. Multi-Head Attention

$$head_i = Attention(QW_iQ, KW_iK, VW_iV)$$

$$MultiHead = Concat(head_1, \dots, head_h)WO$$

Different heads can learn short-term variation, daily periodicity, and weather-related dependencies simultaneously.

5. الانتباه متعدد الرؤوس

يتيح تعدد الرؤوس تعلم أنواع مختلفة من العلاقات الزمنية في الوقت نفسه.

6. Positional Information

$$z_t = x_t + p_t$$

$$PE(pos, 2i) = sin(pos/100002i/d)$$

$$PE(pos, 2i + 1) = cos(pos/100002i/d)$$

Attention alone does not know sequence order, so positional information must be added.

6. المعلومات الموضعية

يُضاف تمثيل لموضع كل خطوة لأن الانتباه وحده لا يعرف ترتيب الزمن.

7. Transformer Encoder Block

Input + Position → Multi-Head Attention → Add & Normalize →

Feed-Forward → Add & Normalize

$$FFN(x) = W_2\varphi(W_1x + b_1) + b_2$$

Residual connections preserve information and improve gradient flow.

7. كتلة Encoder

تتكون من الانتباه متعدد الرؤوس والوصلات المتبقية والتطبيع وشبكة تغذية أمامية.

8. Causal and Non-Causal Attention

Causal Attention
Prevents access to future time steps.

Non-Causal Attention
Allows access to the full input window.

Forecasting systems must prevent future-information leakage.

8. الانتباه السببي وغير السببي

يمنع الانتباه السببي الوصول إلى المستقبل، وهو ضروري في التنبؤ الزمني الحقيقي.

9. Transformer vs RNN and LSTM

Property	RNN/LSTM	Transformer
Processing	Sequential	Parallel over sequence
Long-range dependencies	Possible but harder	Direct attention links
Memory cost	Moderate	Can grow quadratically
Small datasets	Often competitive	May overfit

9. مقارنة Transformer مع RNN وLSTM

يمتاز Transformer بالمعالجة المتوازية والروابط المباشرة، لكنه قد يحتاج إلى بيانات وحوسبة أكبر.

10. Applications in Power Systems

Load Forecasting
Short- and long-horizon demand.

Renewable Forecasting
PV and wind time series.

Event Detection
Disturbances and anomalies.

Price Forecasting
Market time-series modeling.

State Estimation
Temporal measurement relationships.

Asset Monitoring
Long-term sensor patterns.

10. التطبيقات في نظم القدرة

تشمل التنبؤ بالحمل والمتجددة والأسعار وكشف الأحداث ومراقبة الأصول وتقدير الحالة.

11. Limitations

- Attention cost can scale quadratically with sequence length.
- Large models may require substantial data and computation.
- Attention weights are not always complete explanations.
- Distribution shift can reduce performance after deployment.

11. القيود

قد تكون الكلفة الحسابية مرتفعة، كما أن أوزان الانتباه لا تعد دائمًا تفسيرًا كاملًا للقرار.

12. Engineering Workflow

Define Forecast → Prepare Sequences → Build Baseline →

Train Transformer → Validate & Monitor

12. سير العمل الهندسي

يجب بناء خط أساس، والحفاظ على التقسيم الزمني، ثم تقييم النموذج ومراقبته بعد التشغيل.

🧠 Review Questions

1. What are queries, keys, and values?
2. Why is attention scaled by $\sqrt{d_k}$?
3. Why is positional information necessary?
4. What is the purpose of multiple heads?
5. What is causal masking?
6. When might LSTM be preferable to Transformer?

🧠 أسئلة المراجعة

1. ما هي Q وK وV؟
2. لماذا نقسم على $\sqrt{d_k}$؟
3. لماذا نحتاج إلى معلومات الموضع؟
4. ما فائدة تعدد الرؤوس؟
5. ما المقصود بالقناع السببي؟
6. متى قد تكون LSTM أنسب من Transformer؟

References | المراجع

المراجع العامة المستخدمة في هذا المقرر

[1] A. Gabash, *Flexible Optimal Operations of Energy Supply Networks with Renewable Energy Generation and Battery Storage*. Saarbrücken, Germany: Südwestdeutscher Verlag für Hochschulschriften, 2014.

[2] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th Global ed. Harlow, U.K.: Pearson Education Limited, 2022.

[3] A. Gabash, “Energy Market Transition and Climate Change: A Review of TSOs-DSOs C+++ Framework from 1800 to Present,” *Energies*, vol. 16, no. 17, Art. no. 6139, 2023, doi: 10.3390/en16176139.

[4] A. Gabash, *AI Applications in Electrical Power Systems*, Version 1.0, 2026. [Online]. Available: https://aoussgabash.com

Recommended citation:

Gabash, A. (2026). Transformer Models and Attention. AI Applications in Electrical Power Systems, Lecture 10, Version 1.0. https://aoussgabash.com

المرجع المقترح:

أوس غباش، 2026، محاضرة 10، تطبيقات الذكاء الاصطناعي في أنظمة الطاقة الكهربائية، الإصدار 1.0، https://aoussgabash.com

© 2026 Dr.-Ing. Aouss Gabash. Educational use with attribution to the author and official website.