光华管理学院 Guanghua School of Management

因思想 而光华

从 Transformer 到通用时间序列预测
From Transformer to Universal Forecaster

李丰

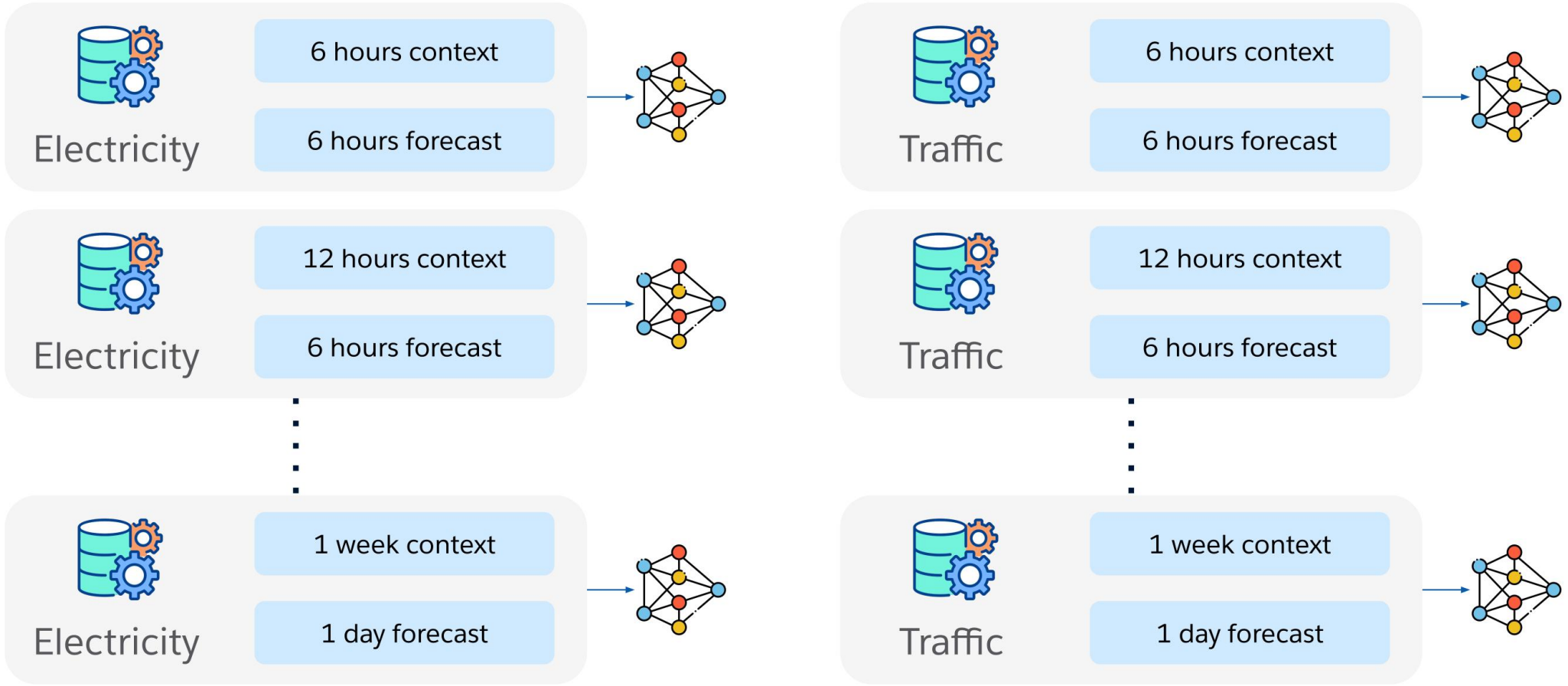
北京大学光华管理学院

https://feng.li/forecasting-with-ai

传统的预测范式

Existing forecasting paradigm

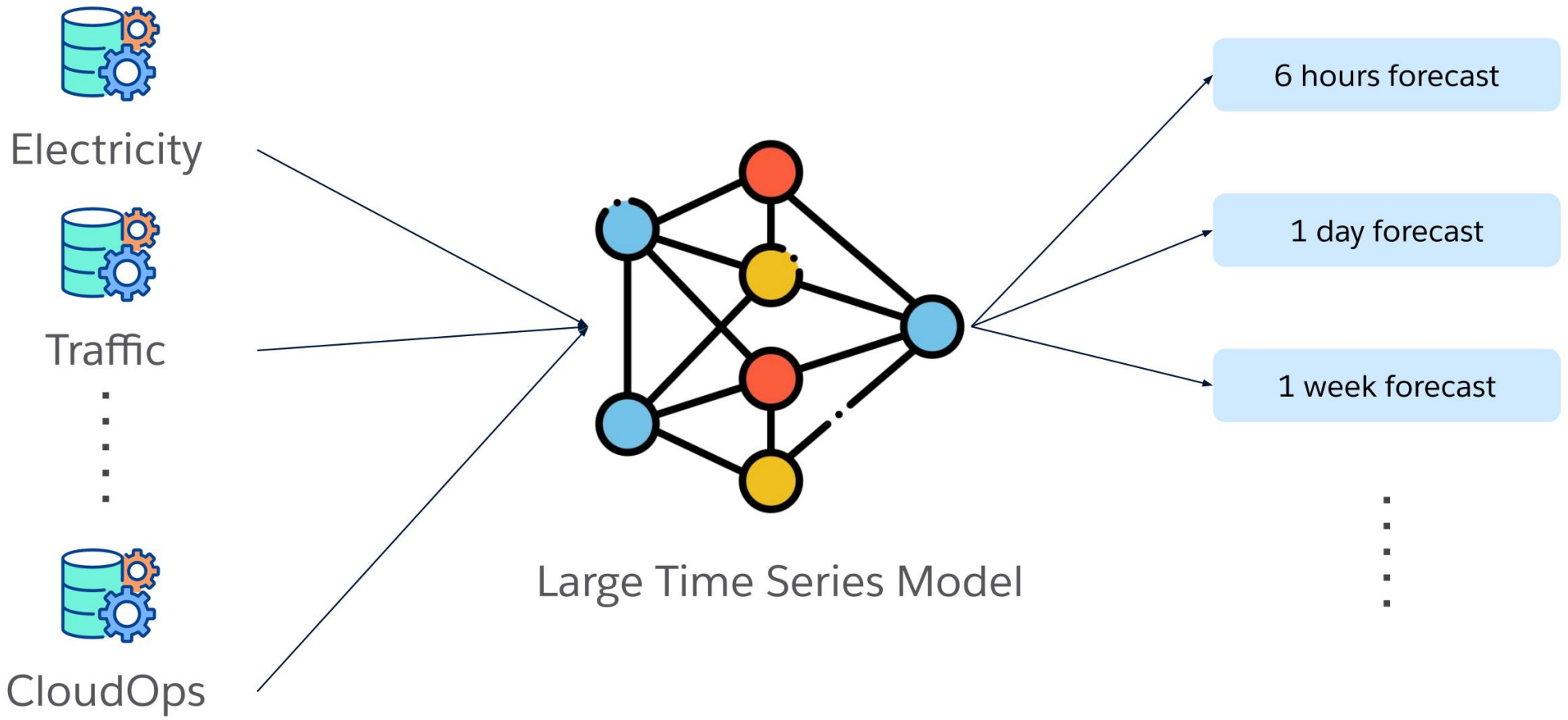
传统时间序列预测范式


Electricity
6 hours context
6 hours forecast
Traffic
6 hours context
6 hours forecast
Electricity
12 hours context
6 hours forecast
Traffic
12 hours context
6 hours forecast
Electricity
1 week context
1 day forecast
Traffic
1 week context
1 day forecast

局限与挑战

- 当前主流的深度预测方法通常遵循以下模式：
  - “一模型对应一数据集、一上下文长度、一预测长度”
  - 每个数据集都需要单独训练模型
  - 模型只能处理固定的输入窗口与固定的预测步长
  - 当数据集或任务设置发生变化时，模型需重新训练。
- 典型模型举例：ARIMA, DeepAR、N-BEATS、Temporal Fusion Transformer (TFT)、Informer
- 主要局限：
  - 可扩展性差（Scalability Issue）
  - 不同数据集、预测任务需重复训练，计算成本高。
  - 泛化能力弱（Poor Generalization）
  - 模型难以跨领域、跨时间尺度迁移。
  - 灵活性不足（Low Flexibility）
  - 资源浪费（Resource Inefficiency）
- 每次任务变化都要重新建模，参数无法共享。

全新预测范式：通用预测模型（Universal Forecaster）



通用预测模型

- **预训练+微调**（Pretrain & Fine-tune）
- 支持**多任务**与**可变上下文长度**
- 代表模型：
  - **PatchTST:** Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2023). A Time Series is Worth 64 Words: Long-term Forecasting with Transformers https://doi.org/10.48550/arXiv.2211.14730
  - **TimeGPT**: Garza, A., Challu, C., & Mergenthaler-Canseco, M. (2024). TimeGPT-1 https://doi.org/10.48550/arXiv.2310.03589
  - **Chronos:** Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., Zschiegner, J., Maddix, D. C., Wang, H., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., & Wang, Y. (2024). Chronos: Learning the Language of Time Series. https://doi.org/10.48550/arXiv.2403.07815
  - **Moirai:** Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., & Sahoo, D. (2024). Unified Training of Universal Time Series Forecasting Transformers. https://doi.org/10.48550/arXiv.2402.02592

Transformer 在通用预测模型中的主导地位

- Transformer 自注意力机制（Self-Attention）能捕捉长程依赖
- 输入长度灵活，天然适配变长序列
- 可并行计算，训练效率高
- 结构通用性强，语言、图像、时间序列均可适用。

模型	核心结构	特点
PatchTST (2023)	Vision Transformer (ViT) 思想	将时间序列切片为“时间块”作为Token
TimeGPT (Nixtla, 2024)	Decoder-only Transformer	基于大规模多领域时间序列预训练
Chronos (Amazon, 2024)	GPT架构 + Token化时间序列	使用离散化时间Token和概率建模
Moirai (Salesforce, 2024)	Sparse Transformer + Mixture-of-Experts (MoE)	支持可变步长和多频率输入
TimesFM (Google, 2024)	Encoder-only Transformer	在数百万序列上预训练

Transformer

Attention Is All You Need

从序列出发：语言与时间的共同逻辑

- 企业数据中处处是“序列”：
  - 客户行为日志（点击、购买、退货）
  - 财务指标（季度报表）
  - 市场舆情（时间演化）
- 传统方法（ARIMA, RNN、LSTM）的问题：
  - 无法并行 (scale law)
  - 记忆短期，忽视长期关系
- Transformer 突破点：同时关注“整个序列”
- **示例：**预测顾客是否复购，既要看“最新点击”也要看“历史购买模式”。

Transformer 的注意力机制

- **概念：**注意力机制（Attention）= 让模型学会“关注重点”

$$Attention(Q, K, V) = softmax(QK^T/\sqrt{d_k})V$$

- ***Q****uery*（问题）：你在问“谁最重要？”
- ***K****ey*（线索）：每个人说了什么
- ***V****alue*（答案）：对应的信息内容

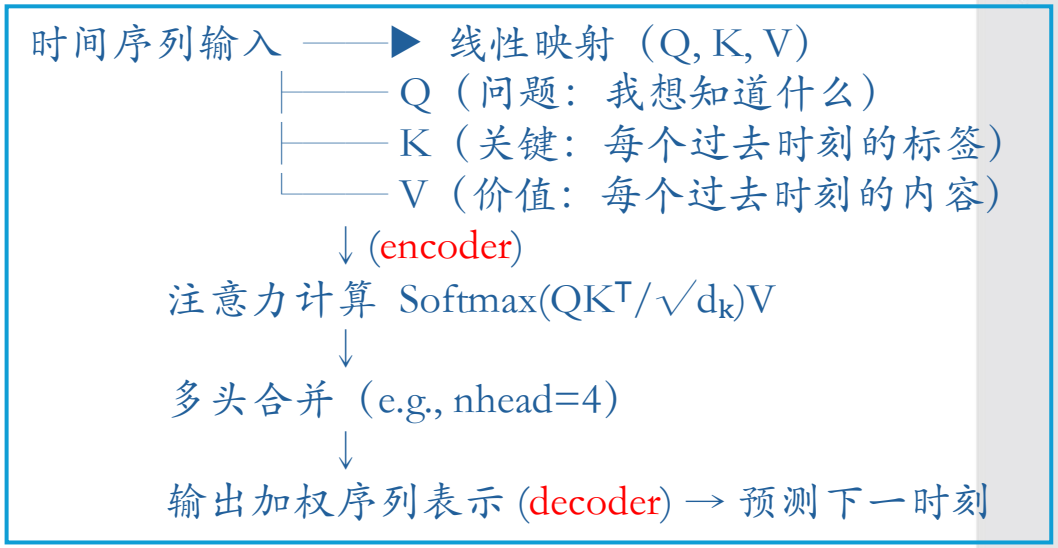
- **模型根据相关性决定“谁值得被听”**
  - QK^T：计算相似度: 每个 Query 与所有 Key 相乘得到“相似程度”。值越大表示 Query 对该 Key 越“关注”。
  - 除以 $\sqrt{d_k}$：缩放避免维度过大导致值过大，致使 softmax 输出过于极端（梯度不稳定）。
  - softmax：归一化注意力权重，把所有相似度转成概率权重（总和为 1）
  - 乘以 V：加权求和得到注意力输出。每个输出是所有 Value 的加权组合，权重来自 attention 分数。

Transformer softmax 的逐元素完整形式

- 注意力分数矩阵（未归一化）为：$S = \frac{QK^{\top}}{\sqrt{d_k}}$
  - 其第 i,j 个元素为：$s_{ij} = \frac{q_i \cdot k_j}{\sqrt{d_k}}$
  - 表示**第 i 个 Query 与 第 j 个 Key** 的相似度
- Softmax 在每一行上进行归一化（对所有 j）：$a_{ij} = \frac{\exp(s_{ij})}{\sum_{j'=1}^{L_k} \exp(s_{ij'})}$
  - 对每个 Query（第 i 行），将所有 Key 的相关性分数转化为概率权重
  - 所有权重之和等于 $\sum_{j=1}^{L_k} a_{ij} = 1$

Attention 层内部数学结构

- $Attention(Q, K, V) = softmax(QK^T/\sqrt{d_k})V$
- 其中 $Q = XW_Q, K = XW_K, V = XW_V$是线性投影
  - X 是 输入序列的特征表示矩阵，每个时间步经过 embedding/线性变换后的表示 $X = xW_{input} + b_{input} \in R^{L \times d_{model}}$。
  - 把**原始数据x**映射到 Transformer 内部隐藏维度d_{model}，b_{input}是偏置向量增加线性变换的灵活性。
  - 每个W_{input}, W_Q, W_K, W_V是可学习的参数矩阵。



概念	时间序列预测中的场景
Q（Query）	“当前分析师的问题”——我想知道哪些过去的事件会影响现在？
K（Key）	“历史记录的标签”——每条过去数据告诉我：我是节日、淡季还是促销期
V（Value）	“历史记录的内容”——每个时刻的真实销售表现
Attention	“权重分配机制”——哪些过去事件对当前预测最相关？

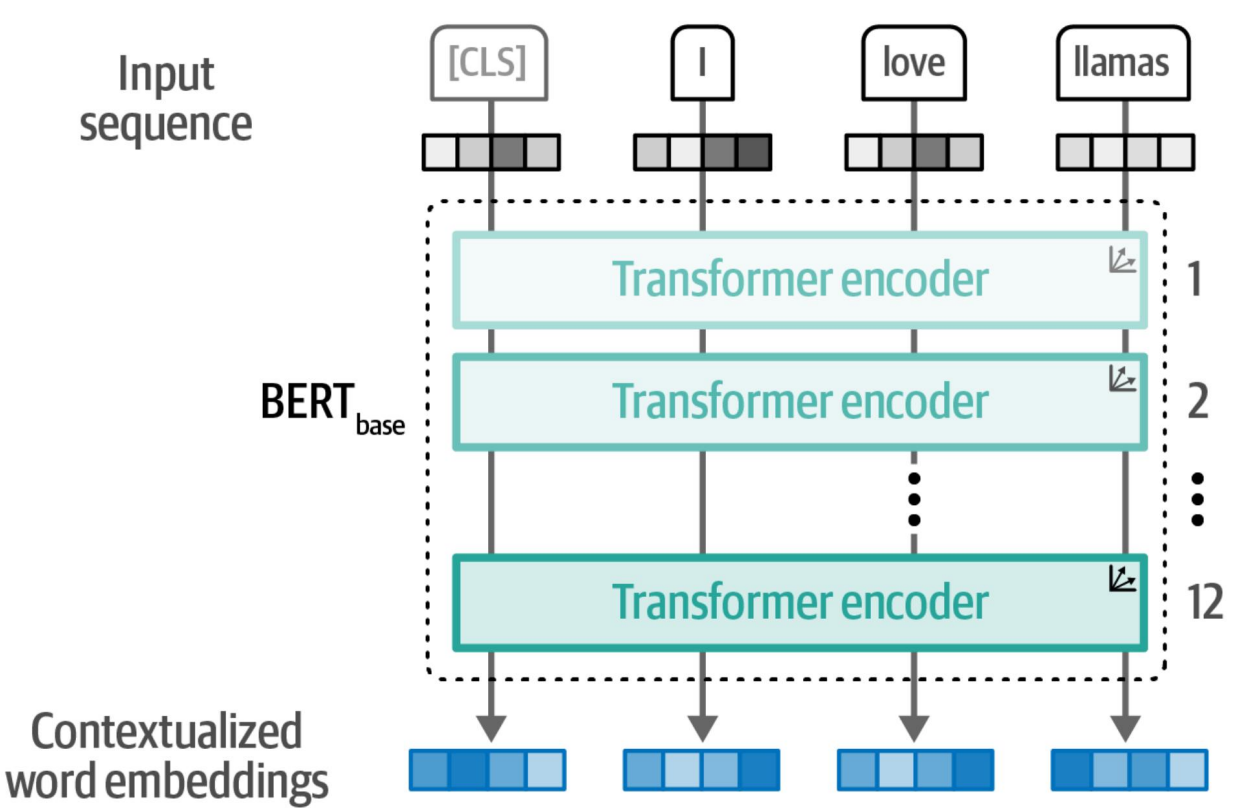
- **Transformer 模型的本质：**
- **自动学习“哪些历史时刻最重要”。**
- 这比 ARIMA, LSTM 那种固定记忆顺序的模型更灵活、更智能。

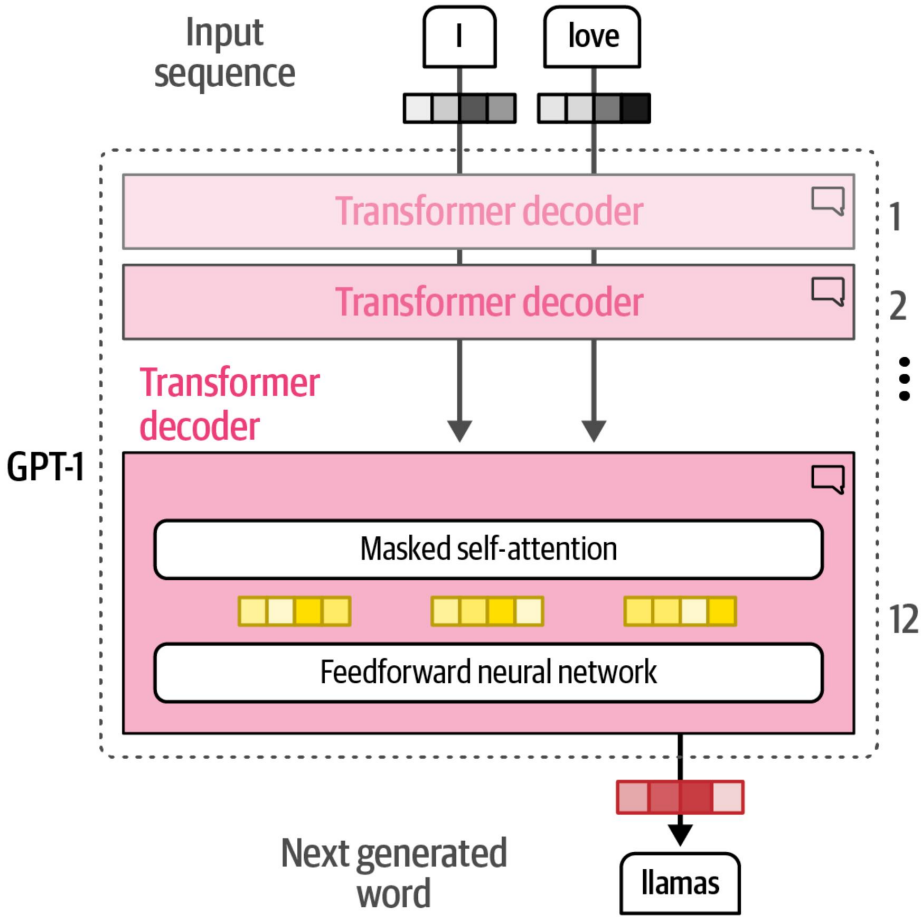
Transformer 的计算复杂度

操作	复杂度	含义
计算 Q, K, V（线性映射）	(O(Ld^2))	每个输入都要经过三个投影矩阵
计算 QK^T	(O(L^2 d))	所有时间步两两匹配
Softmax + V 加权	(O(L^2 d))	根据注意力权重取加权平均
合并多头结果	(O(L h d))	拼接所有头的输出

模型	计算复杂度	是否可并行性	长序列表现
RNN / LSTM	(O(Ld^2))	✖ 不可并行	记忆短
CNN (1D卷积)	(O(Lk d^2))	✔ 可并行	固定
Transformer	(O(L^2 d))	✔ 可并行	全局建模，但耗资源
Informer	(O(L log L))		稀疏注意力（只关注关键时间点）
TimeMixer / PatchTST	(< O(L))		分块建模时间序列

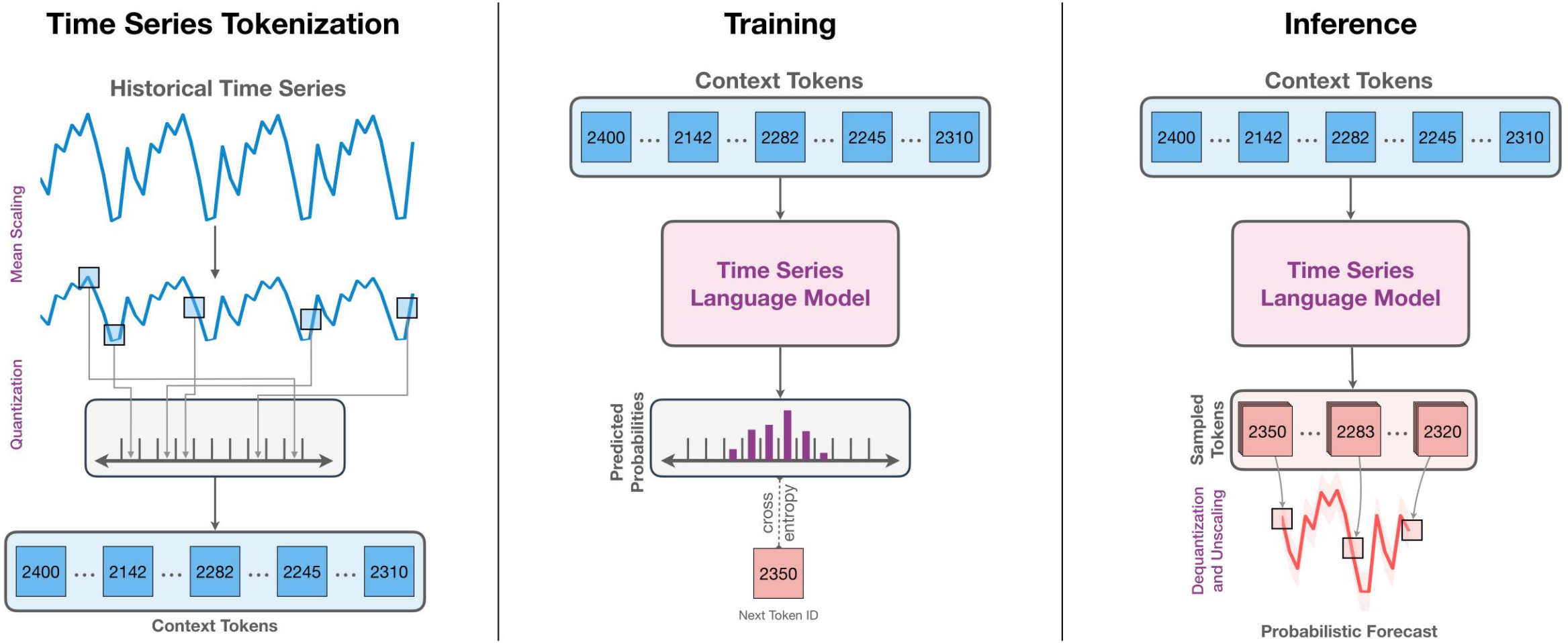
从 BERT 到第一代 GPT 模型





Alammar, J., & Grootendorst, M. (2024). *Hands-On Large Language Models: Language Understanding and Generation*. O'Reilly Media.

学习时间序列的语言



Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., Zschiegner, J., Maddix, D. C., Wang, H., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., & Wang, Y. (2024). *Chronos: Learning the Language of Time Series* https://doi.org/10.48550/arXiv.2403.07815

通用时间序列预测模型对比

模型	任意变量（零样本）	是否支持概率预测	分布灵活性	预训练数据（规模）	是否开源
MOIRAI	✓ 是	✓ 是	✓ 是	LOTSA（> 270亿条）	✓ 是
TimeGPT-1	✓ 是	✓ 是	✗ 否	未公开（约1000亿条）	✗ 否
ForecastPFN	✗ 否	✗ 否	-	合成数据（6000万条）	✓ 是
Lag-Llama	✗ 否	✓ 是	✗ 否	Monash 数据集（< 10亿条）	✓ 是
TimesFM	✗ 否	✗ 否	-	Wiki + Trends + 其他数据（> 1000亿条）	✓ 是
TTM	✗ 否	✗ 否	-	Monash 数据集（< 10亿条）	✓ 是
LLMTime	✗ 否	✓ 是	✓ 是	网络级文本（Web-scale Text）	✓ 是

- 这些通用时间序列模型的预训练数据规模差异巨大，从数千万到千亿级不等。
- 源涵盖合成数据、学术基准集、网络文本与多源趋势数据。
- 总体来看，数据多样性与规模是决定模型泛化能力与零样本预测性能的关键因素。

预测模型的性能比较

TimeGPT-1 与传统模型的预测性能比较

	Monthly		Weekly		Daily		Hourly	
	rMAE	rRMSE	rMAE	rRMSE	rMAE	rRMSE	rMAE	rRMSE
ZeroModel	2.045	1.568	6.075	6.075	2.989	2.395	10.255	8.183
HistoricAverage	1.349	1.106	4.188	4.188	2.509	2.057	2.216	1.964
SeasonalNaive	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Theta	0.839	0.764	1.061	1.061	0.841	0.811	1.163	1.175
DOTheta	0.799	0.734	1.056	1.056	0.837	0.806	1.157	1.169
ETS	0.942	0.960	1.079	1.079	0.944	0.970	0.998	1.009
CES	1.024	0.946	1.002	1.002	0.919	0.899	0.878	0.896
ADIDA	0.852	0.769	1.364	1.364	0.908	0.868	2.307	2.207
IMAPA	0.852	0.769	1.364	1.364	0.908	0.868	2.307	2.207
CrostonClassic	0.989	0.857	1.805	1.805	0.995	0.933	2.157	2.043
LGBM	1.050	0.913	0.993	0.993	2.506	2.054	**0.733**	**0.709**
LSTM	0.836	0.778	1.002	1.002	0.852	0.832	0.974	0.955
DeepAR	0.988	0.878	0.987	0.987	0.853	0.826	1.028	1.028
TFT	0.752	0.700	0.954	0.954	0.817	0.791	1.120	1.112
NHITS	0.738	0.694	0.883	0.883	**0.788**	**0.771**	0.829	0.860
TimeGPT	**0.727**	**0.685**	**0.878**	**0.878**	0.804	0.780	0.852	0.878

- 在所有频率上，TimeGPT 的 rMAE 与 rRMSE 均为最小值或接近最小值。
- 传统统计模型在低频数据上仍具竞争力。如 Theta 和 DOTheta 在月度、周度任务中表现优于其他传统方法。ETS、CES 在日度任务中也能保持较好性能。
- 深度模型内部差异显著。在深度模型中，NHITS 与 TFT 表现明显优于 LSTM 与 DeepAR。
- 整体趋势频率越高，误差越大。反映高频序列的噪声性和建模难度更高。

MOIRAI 与传统模型的预测性能比较

- 多种时间序列预测模型在不同数据集下的概率预测表现。
- 概率预测评估指标为 CRPS（连续分级概率评分，越低越好）与 MSIS（平均尺度化区间得分，越低越好）。
- 结果区分了 Zero-shot（零样本预测）、Full-shot（完全训练）与 Baseline（基线模型）三类。
- 主要结论
  - MOIRAI 系列模型在 Zero-shot 预测中表现最优
  - 传统方法在概率预测上已明显落后
  - **预训练+注意力架构**正成为时间序列预测的新范式

		Zero-shot			Full-shot				Baseline	
		MOIRAI$_{\text{Small}}$	MOIRAI$_{\text{Base}}$	MOIRAI$_{\text{Large}}$	PatchTST	TiDE	TFT	DeepAR	AutoARIMA	Seasonal Naive
Electricity	**CRPS**	0.072	0.055	0.050	0.052±0.00	**0.048±0.00**	0.050±0.00	0.065±0.01	0.327	0.070
	MSIS	7.999	6.172	5.875	5.744±0.12	**5.672±0.08**	6.278±0.24	6.893±0.82	29.412	35.251
Solar	**CRPS**	0.471	0.419	**0.406**	0.518±0.09	0.420±0.00	0.446±0.03	0.431±0.01	1.055	0.512
	MSIS	8.425	7.011	**6.250**	8.447±1.59	13.754±0.32	8.057±3.51	11.181±0.67	25.849	48.130
Walmart	**CRPS**	0.103	0.093	0.098	0.082±0.01	**0.077±0.00**	0.087±0.00	0.121±0.00	0.124	0.151
	MSIS	9.371	8.421	8.520	**6.005±0.21**	6.258±0.12	8.718±0.10	12.502±0.03	9.888	49.458
Weather	**CRPS**	0.049	**0.041**	0.051	0.059±0.01	0.054±0.00	0.043±0.00	0.132±0.11	0.252	0.068
	MSIS	5.236	5.136	**4.962**	7.759±0.49	8.095±1.74	7.791±0.44	21.651±17.34	19.805	31.293
Istanbul Traffic	**CRPS**	0.173	0.116	0.112	0.112±0.00	0.110±0.01	0.110±0.01	**0.108±0.00**	0.589	0.257
	MSIS	5.937	4.461	4.277	**3.813±0.09**	4.752±0.17	4.057±0.44	4.094±0.31	16.317	45.473
Turkey Power	**CRPS**	0.048	0.040	**0.036**	0.054±0.01	0.046±0.01	0.039±0.00	0.066±0.02	0.116	0.085
	MSIS	7.127	6.766	**6.341**	8.978±0.51	8.579±0.52	7.943±0.31	13.520±1.17	14.863	36.256

智能决策的新时代

- 企业竞争的核心：数据 + 算法 + 场景
- Transformer让企业从“预测”走向“理解”
- 决策范式转变：
  - 从“经验决策” → “数据驱动决策” → “智能决策”
- 管理者需要理解：
  - 如何在企业中落地小模型（私域知识）
  - 如何治理数据与算法的伦理问题