关于大模型体系-理解-结构-符号

1. 理解即是结构

- 语言不是由“结构”来实现的，结构是由语言和符号来表达的。
- 但是理解就是结构，大模型在前向传播的过程中，嵌入层，注意力层的计算都是线性变换，这其实是一个对事物进行特征提取的过程，它把一个事物用由很多特征的具体取值组成的数值向量来表示。
- 大模型的输入对象在经过注意力层的计算后生成的输出张量，在经过前馈神经网络的计算后，在激活函数层被非线性函数作用在其上后，就产生了化学变化，形成了结构。
- 不同的国家的人们虽然使用不同的语言与符号，但是在语言与符号背后所指代的事物对象是一致的，表示同一种事物，因此不同语言之间可以通过翻译这座桥梁得到等义转换。
- 在Transformer中，“理解”是由模型通过训练后自动构建的结构化表示空间体现出来的。这些理解结构体现在：嵌入矩阵(embedding matrix)最后训练出来的权重矩阵即是词向量转码结构的记录，注意力层最后训练得到的权重参数矩阵得到的特征信息与规律结构的记录，而前馈神经网络层的权重参数矩阵根据某种结构规律由一个对象预测下一个对象，并且建立这两个不同的对象之间的联系。前馈神经网络表达句相事物。本质要理解研究世界的结构，世界的结构如何组成，看维特根斯坦的逻辑哲学论。逻辑空间（维度更少的空间，决定人的认知水平，得到逻辑能力的认知，就是大模型的涌现能力的体现，从低维开始，不断的演化进化的过程，这个世界上发生的事实的总和，事实要符合逻辑空间，一个事实发生后，后，另一个矛盾的事实就不可能发生，逻辑不可违背，讨论的是世界的结构）具有抽象性，抽象的基础：归纳（人具有相同特征，就会做相同的事情，有这个特征就发生这个结果，没有这个特征就不会发生这个结果），演绎。图的概念，组合关系，维特提出图的概念的原因，拆解出来的可以用图表达，图就是组合的关系的表达，事物用更小的部分构成，每个部分形成关系，就是图，本质上组合关系，更小的事物和它们之间发生的事实。逻辑是从现实世界中抽象出来的世界本质规律的空间。组合关系由事物和事实构成。共性，抽象特征的词，事物之间有逻辑联系。逻辑空间可以由语言描述。事实是由逻辑空间的规律产生的。逻辑比事实少，事实是逻辑空间里的规律的重复。主题：语言；结构；做AI的，要了解人。为什么用二进制加减乘除就能表达这个世界，单位相同，世界是自相似的，类似的，单位相同的事情才有加法。世界的起源，结构。

	理解的类型，	隐含的结构形式，	例子
1.	句法理解，	注意力图谱中的依存结构，	主谓宾、修饰关系；
	语义理解，	向量空间中的聚类与方向，	同义词、实体聚类；
	推理与逻辑，	层之间的表示演化计算结构，	多跳问答、因果链；
	篇章与上下文关系，	跨句注意力连接形成的结构，	指代消解、话题迁移；

2.大模型的低能耗训练的可能性

- 大模型一定有一种低能耗，高效率的方式能理解描述我们这个世界，这个宇宙的运行方式。就像天才和普通人摄入同样的能量，产生的对这个世界的理解并不相同。
- 这个低能耗，高效率的理解过程实际上就是大模型在反向传播时，运用更加精确简洁的符号和语言，进行梯度计算和运用优化器更新权重参数的过程。
- 低能耗对应更简洁，高效率对应更精准。
- 另外在大模型对输入张量进行前向传播的过程中，会产生一些低能耗的数据计算流动过程：

稀疏激活 与 密集计算

- 普通Transformer 通常每个输入都要经过所有层、所有头，多数神经元一次都不“寂寞”——能量消耗高。
- Mixture-of-Experts (MoE) 框架下，每个样本只“唤醒”一小部分专家（子网络），其余“休眠”不计算。这样只花费 ≈10%–20% 的 算力，却保持或接近全模型性能。
- 天才在大脑中只调动关键联结，而非调动全脑，因而更高效。

量化 与 剪枝：打磨多余连接

- 网络剪枝（Pruning）：识别并移除那些权重值极小、对最终结果影响甚微的连接。
- 权重量化（Quantization）：把 32-bit 浮点参数压缩成 8-bit 或更低，既节省存储也加速推理。
- 如同大脑在熟练掌握一项技能后，“瘦身”大量不必要的突触，让关键通路更加高效。

检索增强 与 记忆网络

- 传统大模型“全靠内部权重”记忆事实；新的 Retrieval-Augmented Models 则把大部分外部知识存于检索库，只在需要时快速调用。
- 这样模型本体更“轻”，查询知识时才启动额外计算。

要让大模型像“天才”般低耗高效，关键在于：

A.激活稀疏（只唤醒必要部分）
B.剪枝量化（去除冗余、压缩表达）
C.外部记忆+检索（只在需要时调用大脑外的信息库）

校准输出，准确输入，校准我们的概念，符号系统。不精确的符号系统的使用无法实现增量学习，能帮助自成长，能精确的定位，描述，表达，精确的符号化。先定性，后定量。形成自洽的逻辑体系，是实现低能耗的训练的开始，清晰的前向传播，和反向传播的实现。由人推理大模型的训练。核心是可解释，语言能非常清晰的表达出来。

3.结构是能用语言和符号精确高效的表达出来的

- 大模型在进行反向传播的时候，就是形成规律，形成结构的过程。在这个过程中，大模型最后学会和存储的结构是可以用语言和符号精确的表达出来的，就是语言和符号承载了结构的具体意象与形式。
- 在大模型的各个层当中，前馈神经网络，和注意力层在内的数学符号是符号系的一种更加具体与精确的表达。

将模型拆解成符号化的函数块

一个典型的 Transformer 层可以抽象为以下几个子模块：

1.多头自注意力：

$$\mathrm{Attn}^l(X) = \mathrm{Concat}(\{\,\mathrm{softmax}(\tfrac{Q_i K_i^\top}{\sqrt{d}})V_i\}_{i=1}^{h})W^O$$

这里第l层的可训练参数包括：

- 查询、键、值的映射矩阵 $\{W_Q^l, W_K^l, W_V^l\}$
- 输出投影矩阵 W_O^l

2.前馈网络（Feed-Forward）：

$$\mathrm{FFN}^l(X) = \sigma(XW_1^l + b_1^l)\,W_2^l + b_2^l$$

其参数为 $\{W_1^l, b_1^l, W_2^l, b_2^l\}$。

3.归一化与残差连接：通常写作

$$X^{l+1} = \mathrm{LayerNorm}\big(X^l + \mathrm{Attn}^l(X^l)\big), \quad Y^{l+1} = \mathrm{LayerNorm}\big(X^{l+1} + \mathrm{FFN}^l(X^{l+1})\big).$$

LayerNorm 本身也有可学习的缩放和平移参数 $\{\gamma^l, \beta^l\}$。

将整套L层堆叠后，加上输入嵌入层和输出头（如语言建模的词表投影矩阵 $W_{\text{LM-head}}$），模型的全部可训练参数就符号化为：

$$\Theta = \big\{W_{\text{emb}}, \{W_Q^l, W_K^l, W_V^l, W_O^l, \gamma^l, \beta^l, W_1^l, b_1^l, W_2^l, b_2^l\}_{l=1}^{L}, W_{\text{LM-head}}\big\}.$$

符号化模型结构：用数学符号明确标出每个子模块的参数。
符号化梯度流： $\nabla_\theta \mathcal{L} \neq 0$ 表示参数在训练中被更新， $\nabla_\theta \mathcal{L} = 0$ 则表示被冻结。
反过来，把数学里的基本规律，换成赋予它的单位，和实际意义。