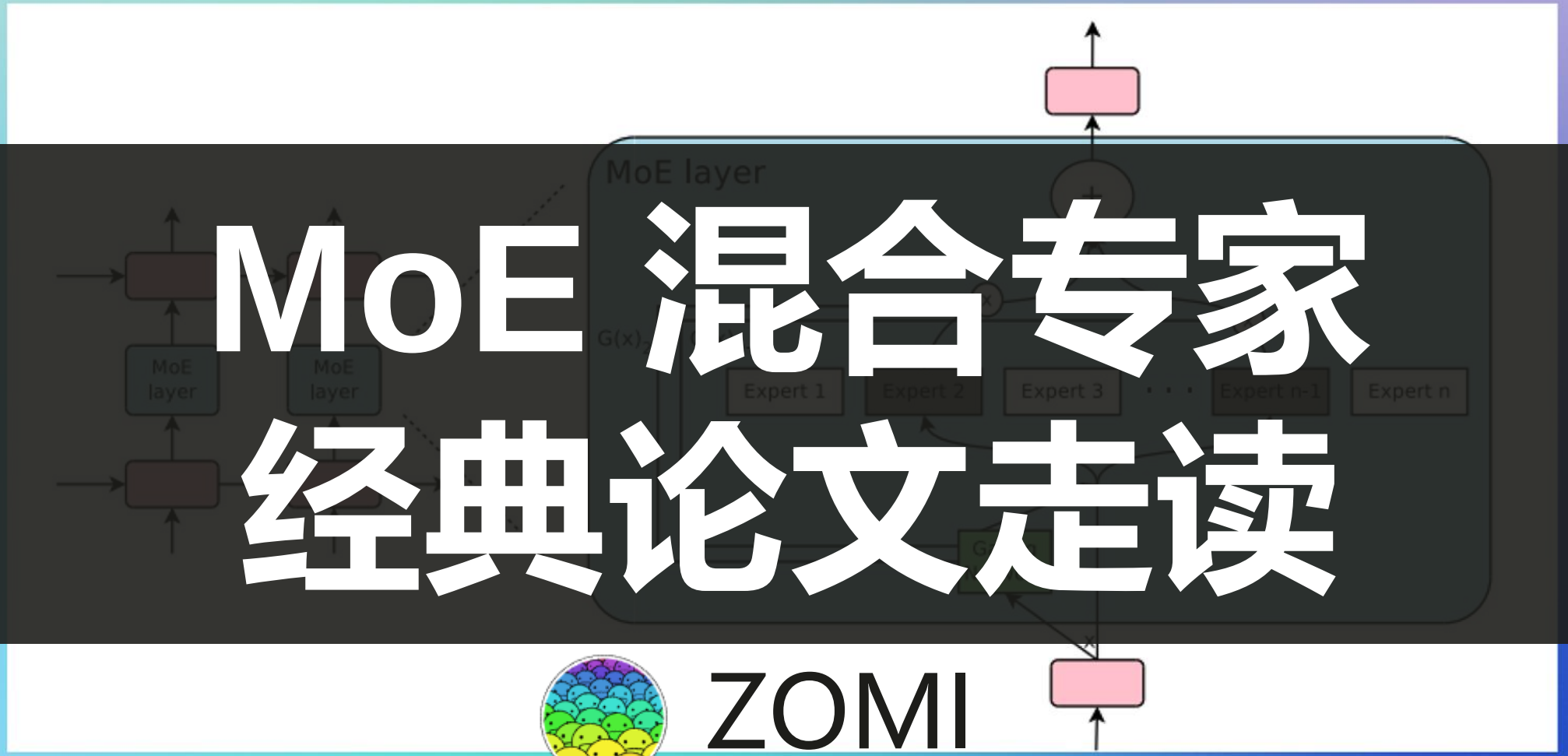
Mixture of Experts (MoE)



MoE 混合专家经典论文走读

ZOMI

Contents

视频目录大纲

- MoE 混合专家
  - MoE 基本介绍 (4)
  - MoE 前世今生 (5)
  - MoE 核心工作
    - 奠基工作：Adaptive Mixtures of Local Experts
    - LSTM MoE：Outrageously Large Neural Networks ...
    - GShard: Scaling Giant Models with Conditional ...
    - Switch Transformers: Scaling to Trillion Parameter Models ...
    - GLaM: Efficient Scaling of Language Models with Mixture ...
    - ST-MoE: Designing Stable and Transferable Sparse Exp
  - MoE 架构原理 (3)
  - 大模型遇上 MoE (4)
  - 手撕 MoE 代码 (5)
  - MoE 的未来 (2)

01

Adaptive Mixtures of Local Experts

思路

- 对于比较复杂的任务，一般可以拆分为多个子任务。比如要求计算输入文本中有多少个动词和名词，那就可以拆分为“数动词”和“数名词”这两个子任务。
- 而一个模型如果要同时学习多个子任务，多个子任务相互之间就会互相影响，模型的学习就会比较缓慢、困难，最终的学习效果也不好。
- 因此这篇文章提出了一种由多个分开的子网络组成的监督学习方法。这些分开的网络，在训练过程中，分别学习处理整个训练数据集中的一个子集，也就是一个子任务。这个思路就是现代MoE的思路，每个子网络（也就是一个expert）学习处理一部分内容。
- 文章里把这个MoE的方法应用于vowel discrimination task，即元音辨别任务，验证了MoE设计的有效性。元音辨别指的是语音学中区分不同元音的能力，在语音学中，模型需要学习辨别不同的元音因素，以便准确地理解和识别语音输入。通过让多个子模型分别学习分别学习不同元音（a、e、i、o、u）辨别的子任务，最终效果得到了提升。

模型设计

- 各个expert network和gating network接收同样的输入，每个expert给出各自的处理结果；而gating network输出每个expert的权重，就像一个开关一样，控制着每个expert对当前输入的打开程度，只是这个开关不是离散的，而是stochastic的，给出的不是true和false，而是权重。

损失函数优化

- 实际上，MoE这个idea在这篇文章之前就有了。如论文中所提，Jacobs和Hinton在1988就讨论过。但是之前的工作在loss的设计上，和ensemble更相近，多个expert之间更倾向于合作，每个expert会学习其他expert的residual部分。
- 是把期望输出和所有expert输出的混合结果进行比较。
- 这样做的结果是，在训练过程中，每个expert学习的其实是其他expert的组合结果所剩下的残差。这样的学习目标并不能很好迫使每个expert单独输出好的结果，因此不能得到稀疏的模型。
- 从另一个角度来看，这个损失计算把所有专家耦合在了一起。即当一个expert的输出发生了变化，所有expert的组合结果也会变化，其他所有的expert也需要做相应的改动来适应这个变化。因此各个expert之间更加倾向于合作，而不是相互竞争并单独给出好的结果，让gating network输出稀疏的结果。
- 虽然可以使用如增加辅助损失函数的做法，迫使模型给出稀疏激活的结果，但是这样相当于增加了很强的先验正则化，对模型最终效果也是有损害的。

损失函数优化

- 而Hinton和Jordan在这个工作里，提出更简单的做法是对loss计算进行修改，使得各个expert之间的关系从合作变成竞争。
- 在这个损失函数中，每个expert的输出结果会单独和期望结果进行对比，这就要求每个expert单独给出完整的结果，而不是仅学习其他expert的残差。
- 这样的loss计算具有localization的特性，即如果一个训练case错了，那么会被修改的主要是被gating network选中且出错的expert，以及负责分配权重的gating network，而不会很大地影响其他expert。
- 这样一来，不同的expert之间不会直接相互影响，虽然还是有间接的影响，比如某个expert的输出变了，gating network可能会分配新的权重，但是至少不会改变其他expert error的符号（+，-），即优化的方向。
- 最终的结果是，对于给定的输入，这样的系统会倾向于以高权重分配单一一个expert来预测结果。

实操技巧

- 相比原loss函数的导数，优化后的loss函数的导数，把当前第 ii 个expert的表现，和其他expert联系起来了。这样能够更好地衡量expert ii 对当前case的处理结果好坏。特别是在训练初期，gating network的权重是近似平均分配的，那么使用原loss函数的结果是，对当前case效果最好的expert，学习速度是最慢的（因为loss最小）；而优化的loss函数则可以让当前最好的expert的学习速度最快。相当于让“有天赋”的专家在对应的子任务上尽快提高水平。这样就强化了localization的特征，使得各个expert更快拟合到自己擅长的部分，加速训练。

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

GitHub https://github.com/chenzomi12/AIInfra

引用与参考

- https://mp.weixin.qq.com/s/6kzCMsJuavkZPG0YCKgeig
- https://www.zhihu.com/tardis/zm/art/677638939?source_id=1003
- https://huggingface.co/blog/zh/moe
- https://mp.weixin.qq.com/s/mOrAYo3qEACjSwcRPG7fWw
- https://mp.weixin.qq.com/s/x39hqf8xn1cUlnxEIM0_ww
- https://mp.weixin.qq.com/s/ZXjwnO103e-wXJGmmKi-Pw
- https://mp.weixin.qq.com/s/8Y281VYaLu5jHoAvQVvVJg
- https://blog.csdn.net/weixin_43013480/article/details/139301000
- https://developer.nvidia.com/zh-cn/blog/applying-mixture-of-experts-in-llm-architectures/
- https://www.zair.top/post/mixture-of-experts/
- https://my.oschina.net/IDP/blog/16513157

- PPT 开源：https://github.com/chenzomi12/AIInfra
- 夸克链接：https://pan.quark.cn/s/74fb24be8eff