

DYG

Contents

目录

1 团队介绍

2 赛题理解

3 特征工程

4 模型介绍

5 结果与分析

6 总结与思考



团队介绍



王贺

武汉大学 计算机硕士

算法工程师

2019腾讯广告算法大赛冠军

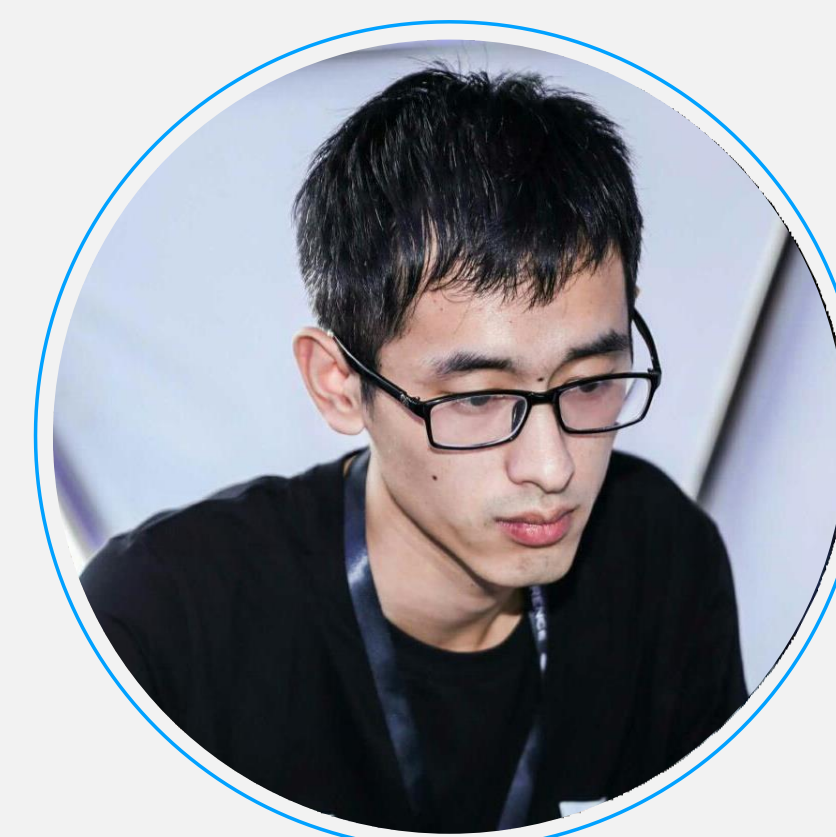


郭达雅

中山大学-MSRA 联合培养博士

2019腾讯广告算法大赛冠军

在国际顶级学术会议NeurIPS, AAAI,
ACL, EMNLP发表多篇一作论文



梁少强

从事NLP相关工作

国内NLP和数据挖掘比
赛中多次名列Top5



赛题理解

数据

- 点击日志：用户id、广告素材id、点击次数与时间
- 用户信息：年龄、性别
- 广告信息：广告id、产品id、类别id、广告主id、行业id

目标

通过用户的点击记录，预测该用户年龄和性别

评价指标

准确率 (accuracy) = 年龄准确率 + 性别准确率



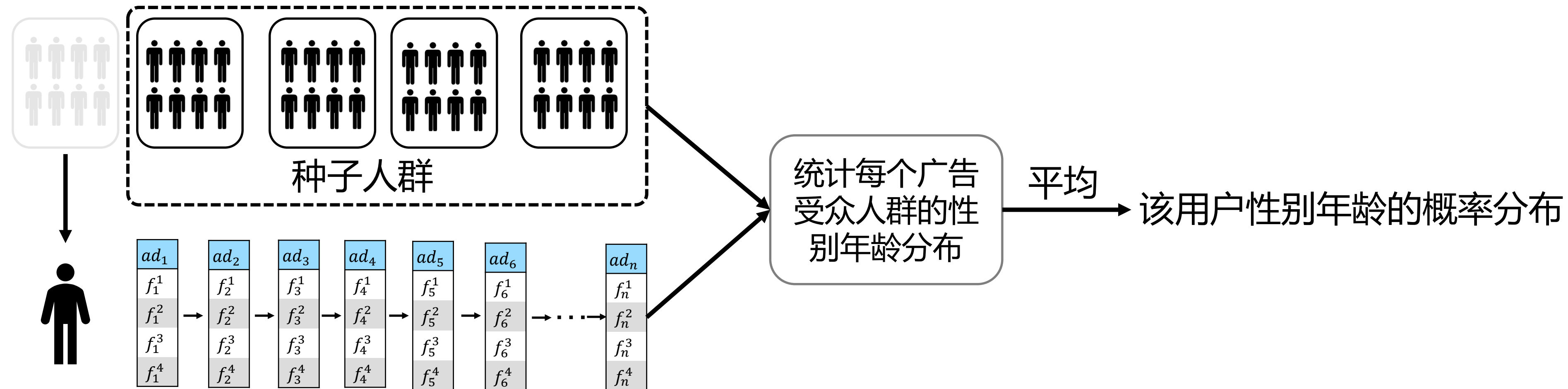
特征工程

特征工程

1. 统计特征

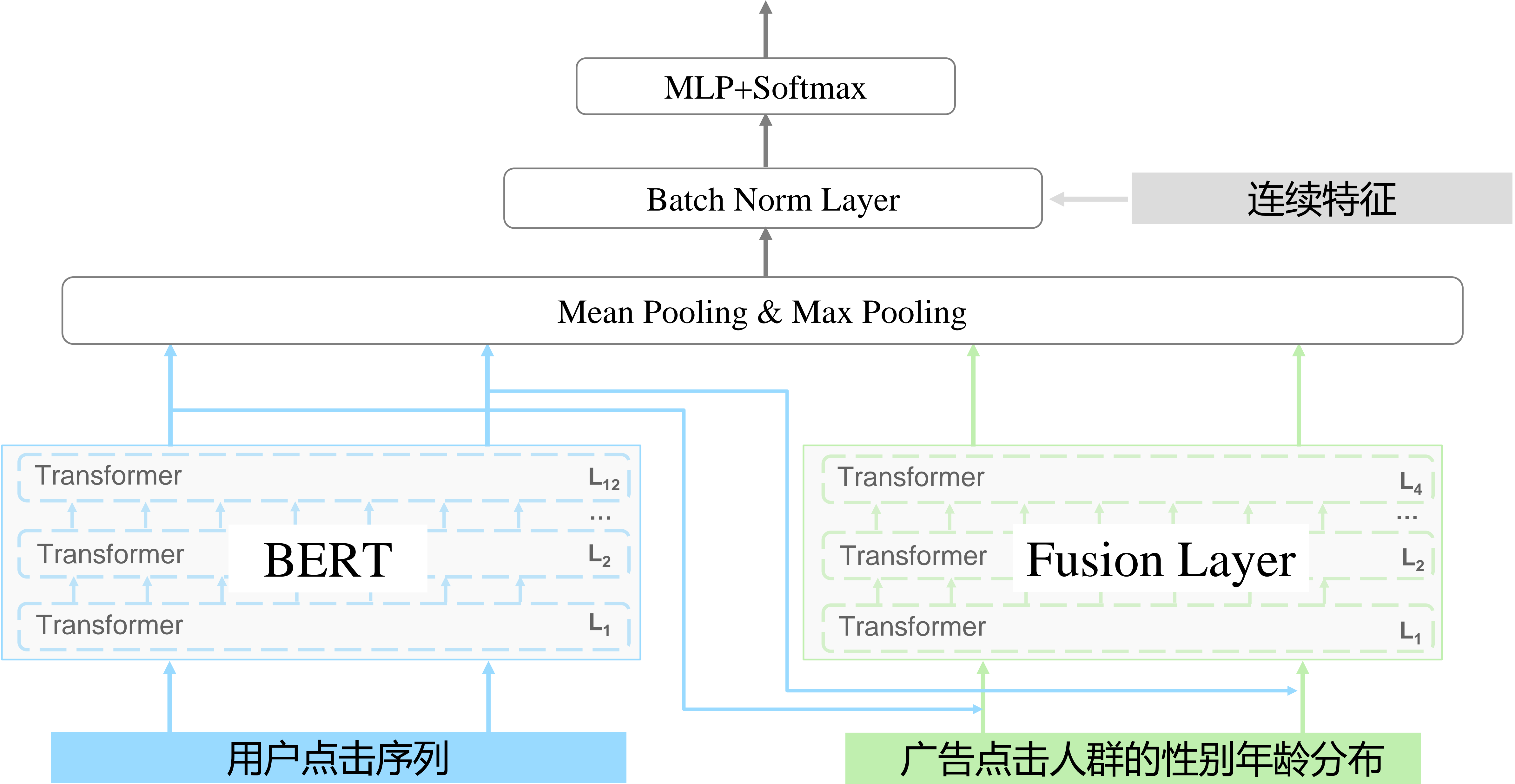
- 1) 用户出现的总次数和天数
- 2) 用户点击广告的总次数
- 3) 用户点击不同广告、产品、类别、素材、广告主的总数
- 4) 用户每天每条广告点击的平均次数，均值和方差

2. 概率分布特征

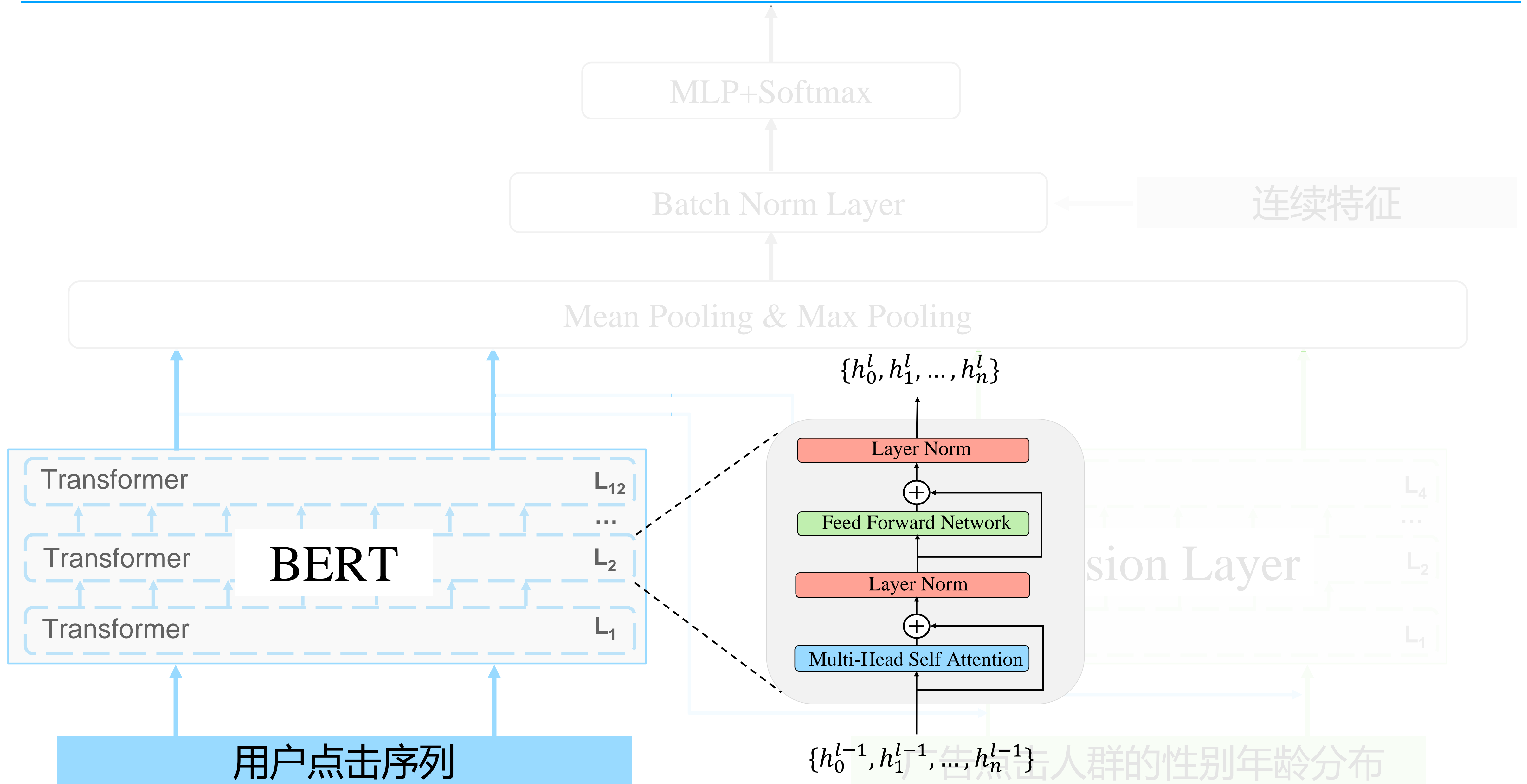




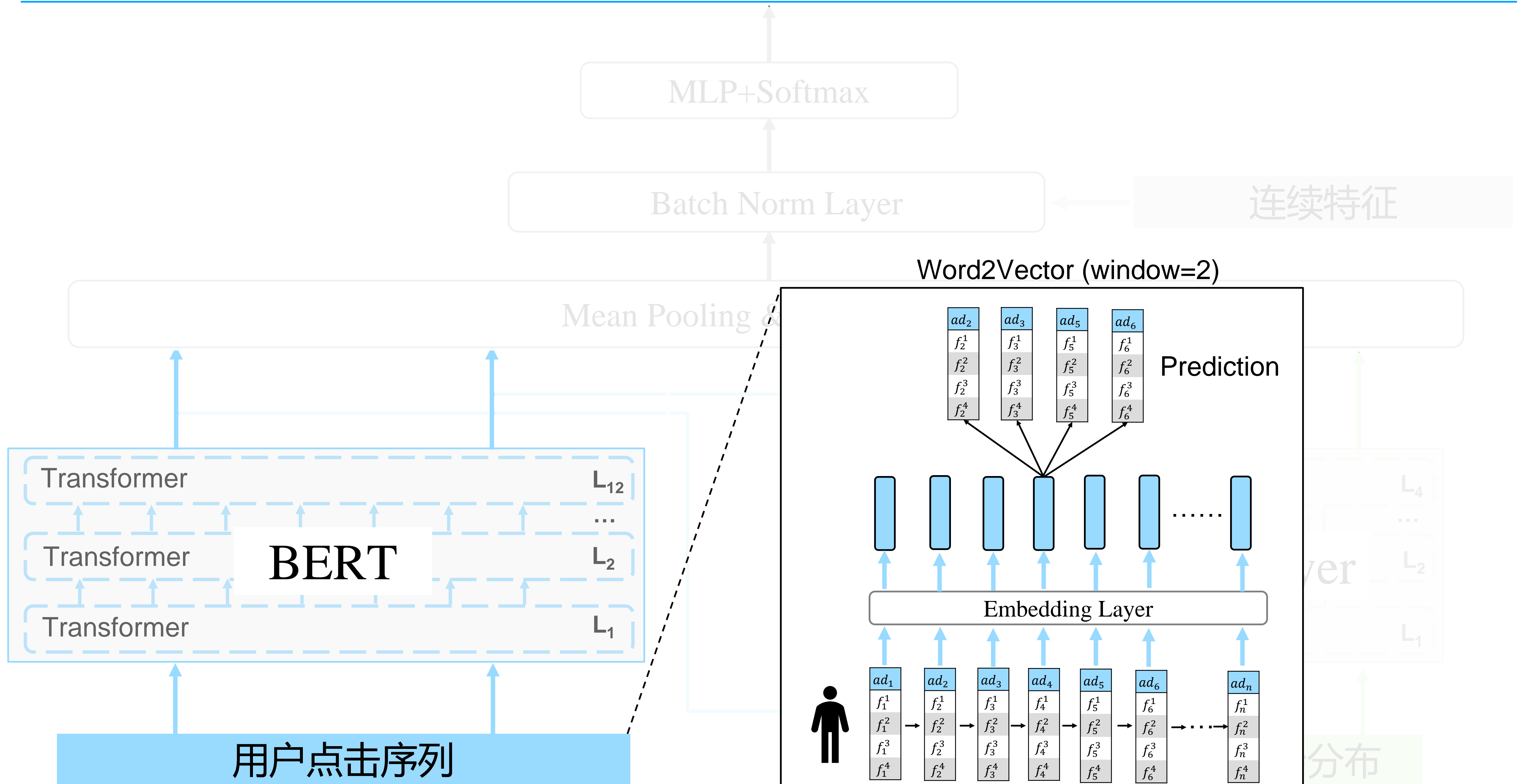
模型介绍



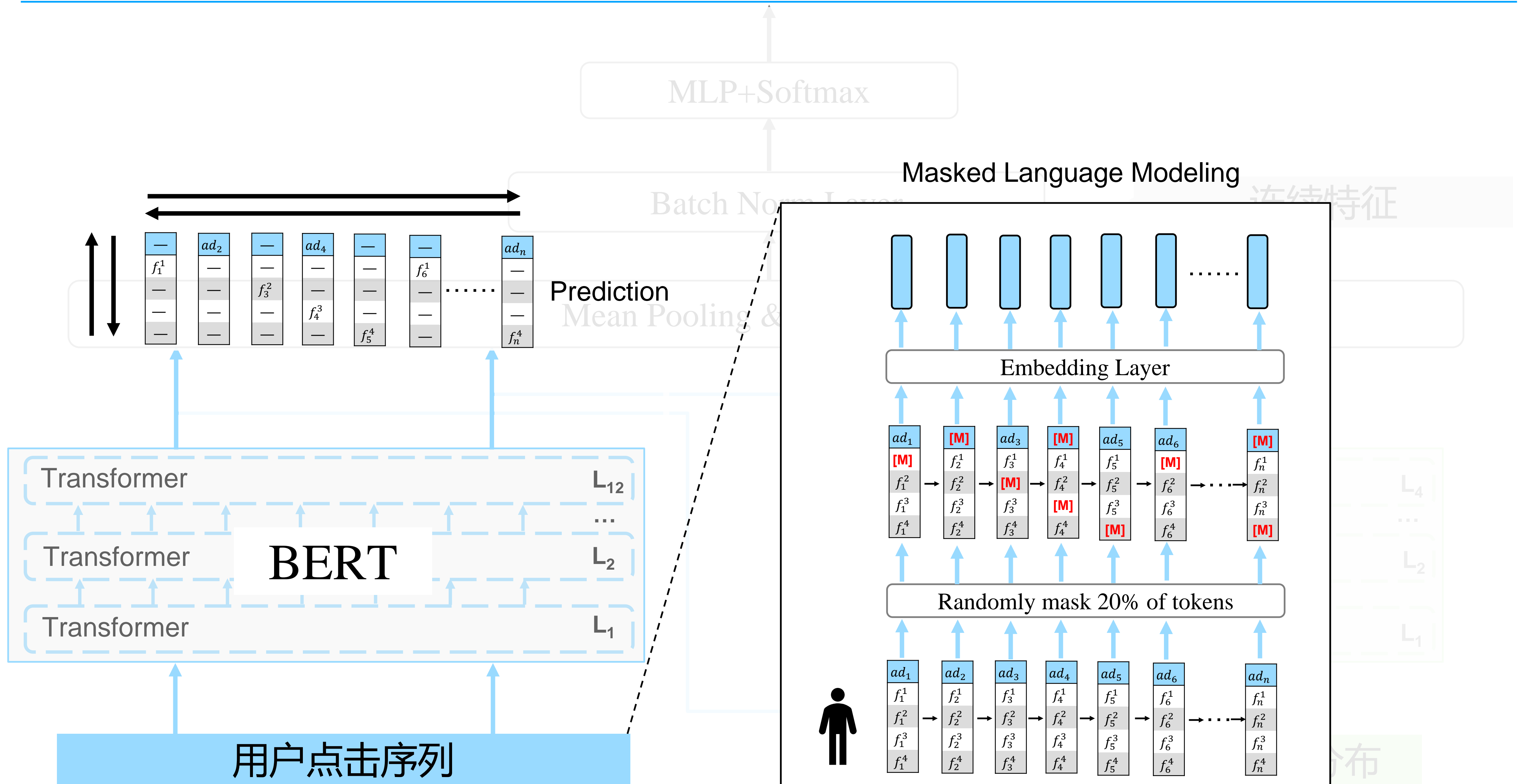
BERT—Bidirectional Encoder Representation From Transformer



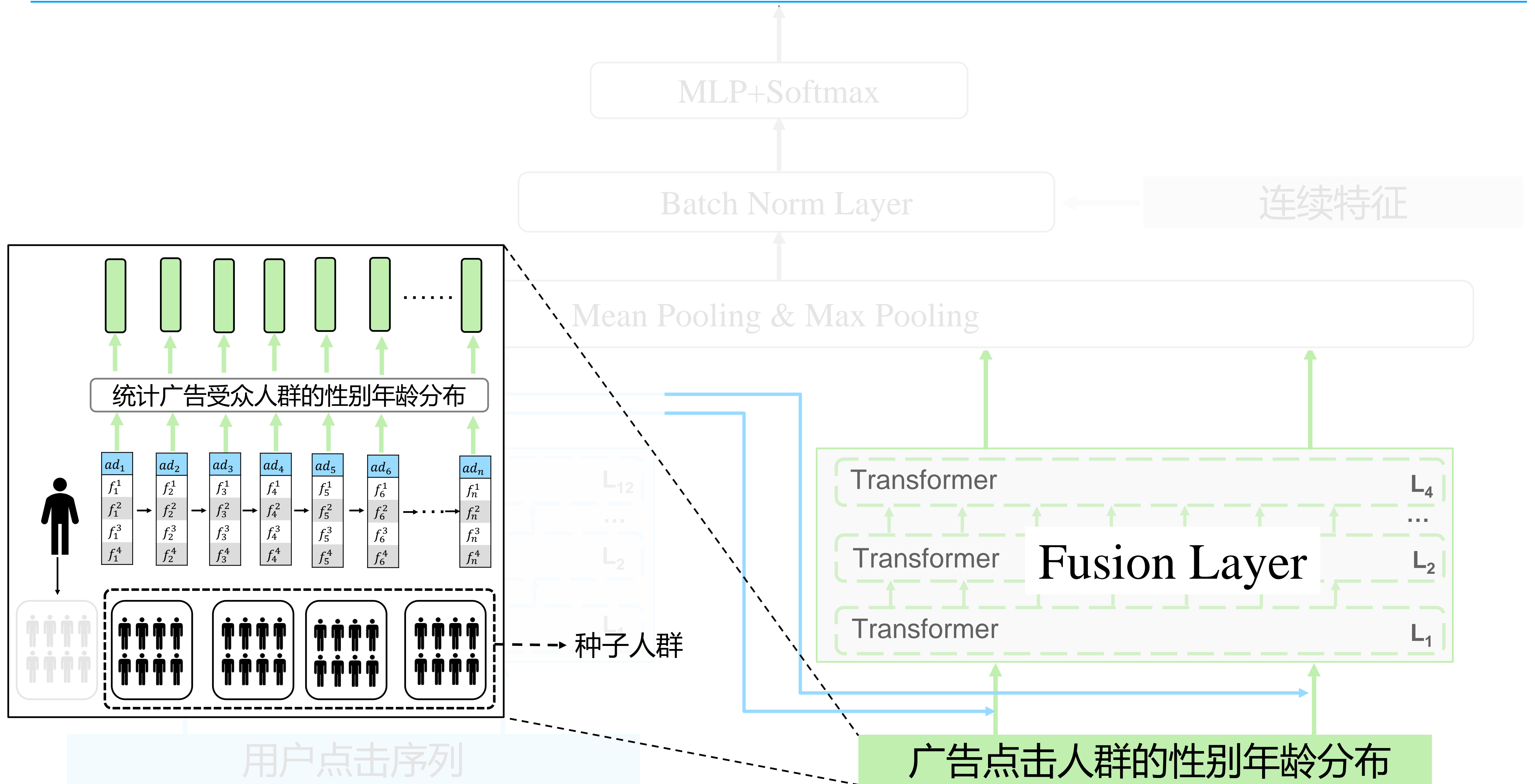
预训练—Word2Vector



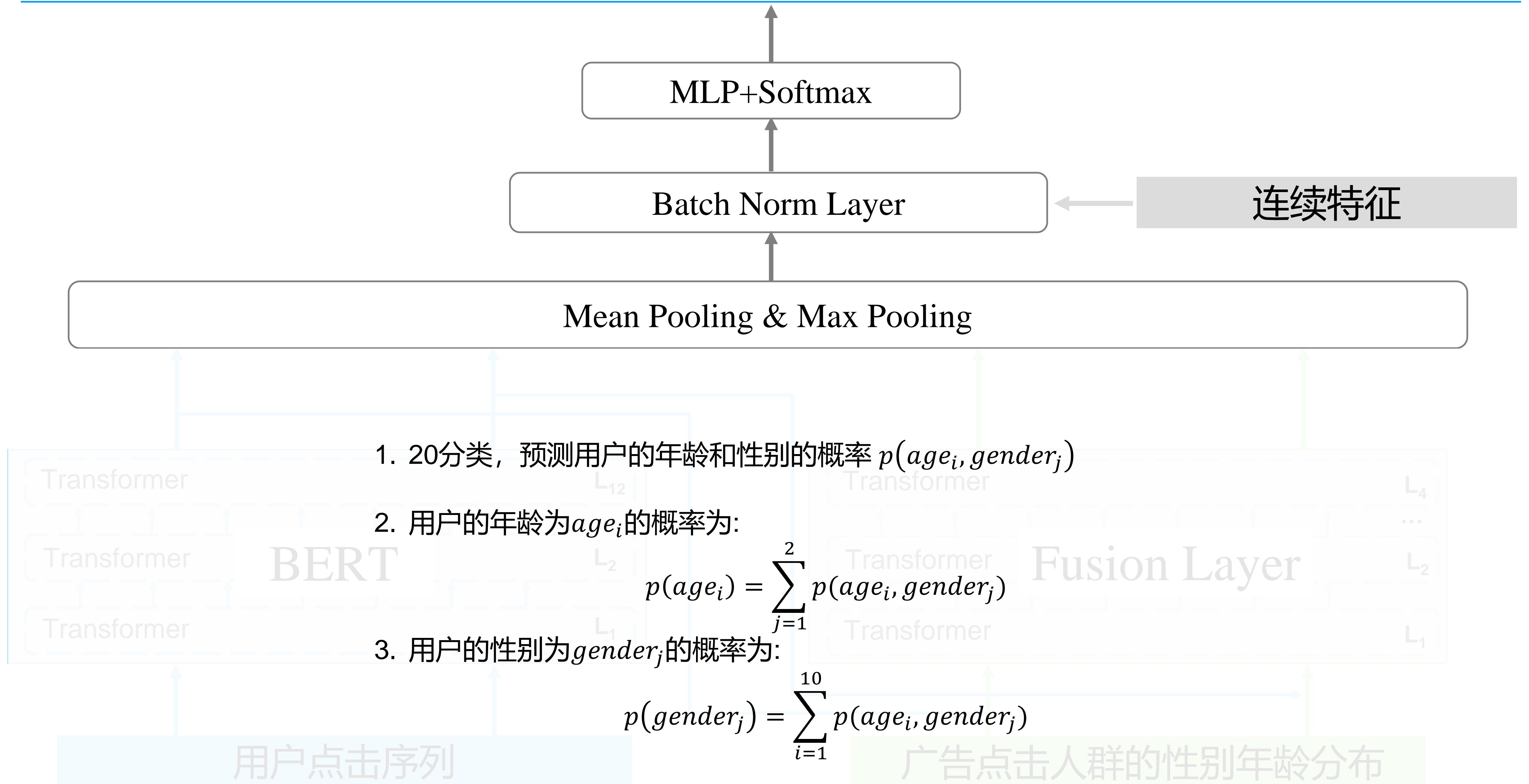
预训练—Masked Language Modeling (MLM)



融合层—Fusion Layer



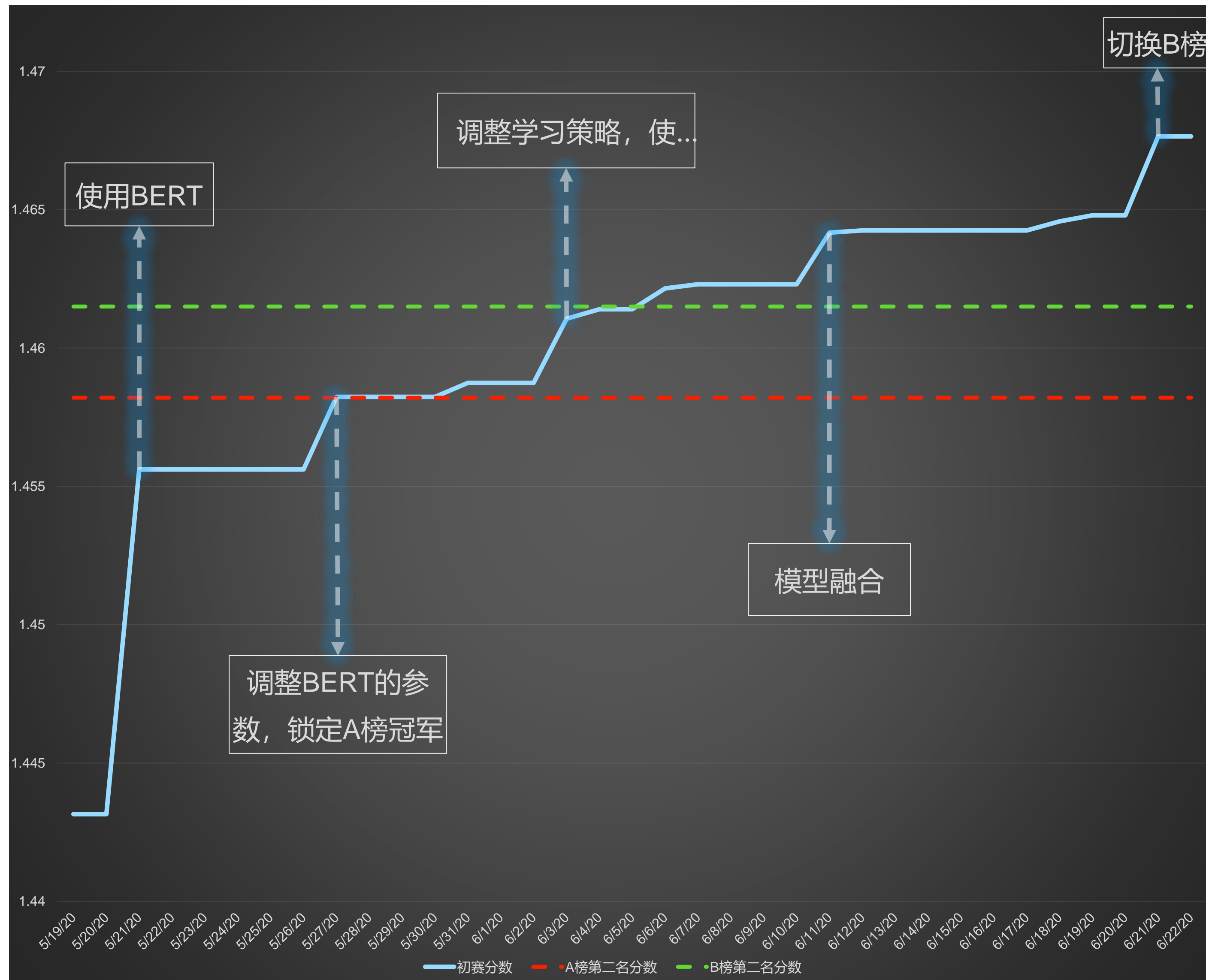
输出层





结果与分析

初赛成绩



2020-5-21: 使用BERT

2020-5-27: 调整BERT的参数, 锁定A榜冠军

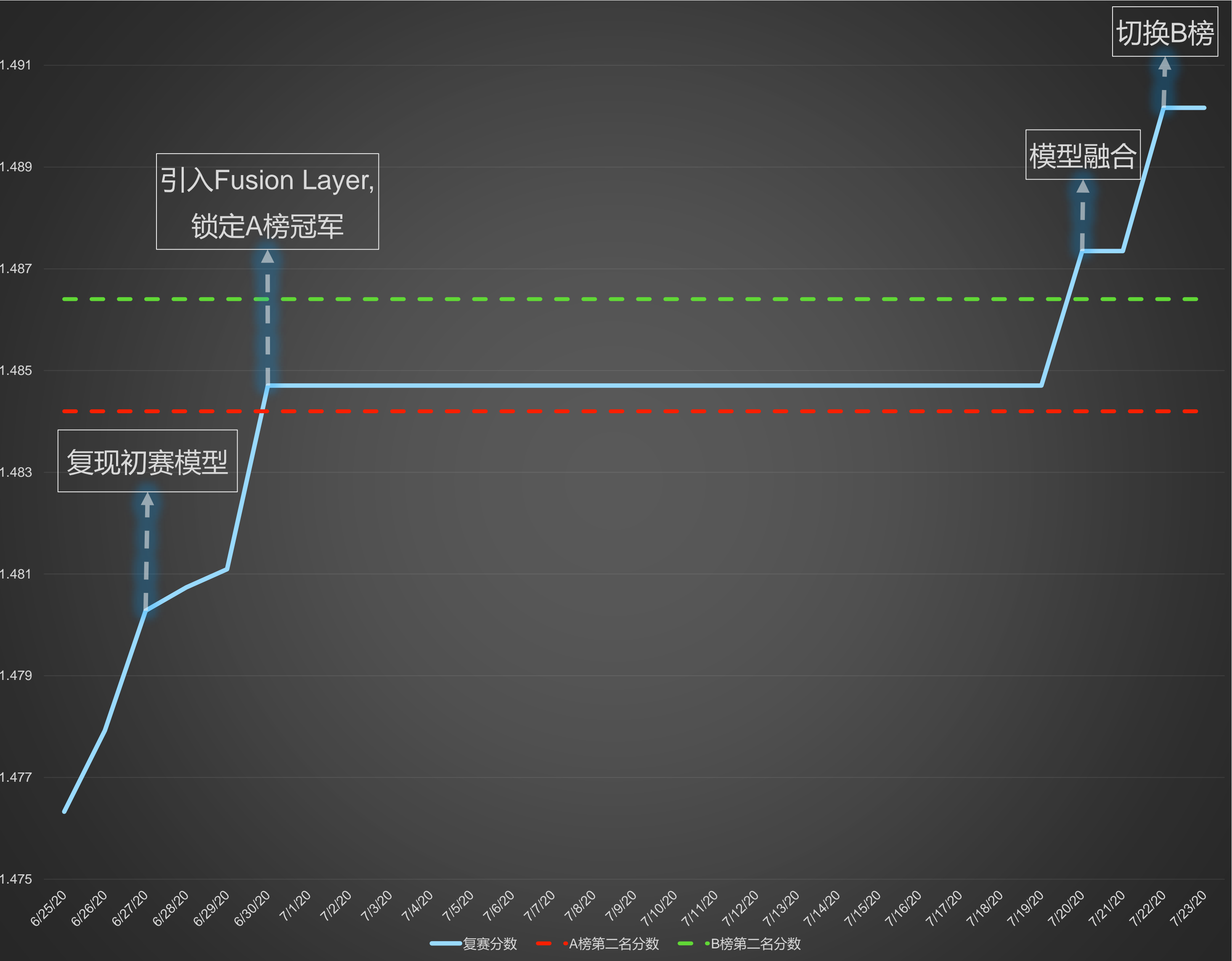
2020-6-03: 调整学习策略, 使用warmup调整学习率, 学习率先增后减

2020-6-12: 模型融合

2020-6-22: 切换B榜

| 队伍排名 | 队伍名称 | 排名变化 | 最佳成绩 |
|------|-----------|------|----------|
| 1 | DYG | 0 - | 1.464798 |
| 2 | 日晨 | 0 - | 1.458208 |
| 3 | 微醺 | 0 - | 1.457616 |
| 4 | 挥霍的人生 | 0 - | 1.457246 |
| 5 | 日晨 | 7 ↑ | 1.456464 |
| 6 | Lindada | 1 ↓ | 1.45584 |
| 7 | BANJITINO | 1 ↓ | 1.4556 |
| 8 | 安之 | 1 ↓ | 1.45534 |
| 9 | chizhu | 1 ↓ | 1.455072 |
| 10 | 天才抱大腿 | 1 ↓ | 1.454272 |

复赛成绩



- 2020-6-26: 复现初赛模型
- 2020-6-30: 引入Fusion Layer, 锁定A榜冠军
- 2020-7-21: 模型融合
- 2020-7-22: 切换B榜

| 队伍排名 | 队伍名称 | 排名变化 | 最佳成绩 |
|------|-----------|------|----------|
| 1 | DYG | 0 - | 1.490166 |
| 2 | 山有木兮 | 0 - | 1.486404 |
| 3 | 微醺 | 0 - | 1.485036 |
| 4 | □□□ | 0 - | 1.483378 |
| 5 | BANJITINO | 1 ↑ | 1.482612 |
| 6 | 正方形的圆 | 1 ↓ | 1.482324 |
| 7 | 天才抱大腿 | 1 ↑ | 1.481152 |
| 8 | 挥霍的人生 | 1 ↓ | 1.480908 |
| 9 | 玉古路38号_ | 5 ↑ | 1.480708 |
| 10 | 小太阳2020 | 27 ↑ | 1.480212 |



总结与思考

主要创新

- 改进BERT并运用到人口属性预测场景
 - 提出分阶段预训练并改进MLM预训练目标，从多维度学习广告及其属性的语义表示
 - 将BERT运用到人口属性预测的场景，从性能上验证了预训练在广告领域的潜力
- 提出融合后验概率分布的方法及模型
 - 利用种子人群求出每个广告的后验概率分布
 - 利用多层Transformer融合后验概率分布及BERT的语义表示，能有效提升性能

问题思考

- 预训练模型越大越好？

| Model | Age | Gender | Overall |
|--------------------------------|---------------|---------------|---------------|
| Bert-small (L=6, H=512,l=1024) | 0.5247 | 0.9507 | 1.4754 |
| Bert-base (L=12,H=512,l=2048) | 0.5268 | 0.9512 | 1.4780 |
| Bert-large (L=24,H=512,l=1024) | 0.5261 | 0.9511 | 1.4772 |
| Bert-xl (L=24,H=512,l=2048) | 0.5234 | 0.9512 | 1.4746 |

- 如何进一步改善预训练模型？
 - 预训练目标
 - 支持图输入的预训练模型
 - 广告稀疏性问题，例如在未脱敏的情况下使用BPE算法

总结

历时两个半月的腾讯广告大赛，非常感谢工作人员辛苦的答疑。感谢主办方提供真实的业务场景与数据，让我们能在比赛中学习到更多知识，在广告业务中做更多尝试。

THANKS

