ark互联网恶意评论识别器

一. ark出现的原因

> 你能看出下面的评论是恶意评论吗?
>
> - 你是一个退休的伐木工，无斧又无木
> - 初升东西
> - 真是宋朝的开始，唐完了
> - 一个人拍照就是全家福

1. 背景

不知大家有没有注意到，现在的互联网风气越来越恶劣，漫骂声，争吵声，网络暴力不绝于耳。越来越多的人开始吧生活中的情绪，工作上的压力，家庭里的不满在网络上倾泻而出，把压抑已久的恶意在陌生人面前展现的淋漓尽致。

起初各大平台为了防范用户间的恶意，采用关键词检索方法审核用户评论。但久而久之，用户们开始用更高超的技巧反制平台的审核，其中最常见的方法就是采用**谐音字代替关键字**以及**语义复杂化**。

2. 谐音字代替关键字

谐音字代替关键字的核心思想就是，通过换字的方式逃过审核时关键词检索，如**无斧又无木**实际上是说**无父又无母**，**初升**实际上说的**畜生**。这种方式简单粗暴，但往往简单的东西反而皮实，为了制服这种谐音字评论，原始的关键词检索只能不断扩充它的词表，但这种方式治标不治本，只要再换个谐音字，关键词检索就又失效了。

3. 语义复杂化

这是一个更加高级的反制手段，同样想要使用这个手段的门槛也更高，它需要用户有深厚的语文功底，将恶意隐匿于无形，同时又不能太晦涩难懂，导致被骂的人浑然不知。如**唐完了**，**唐**指唐氏儿综合征，比起简单直当的**智障**两字，一句**唐完了**更显优雅。至于**一个人拍照就是全家福**，当然说的就是**孤儿**了。显而易见，语义复杂化后对关键词检索如同降维打击，傻傻的审核机制完全看不懂这到底在说些什么。

4. 何为ark

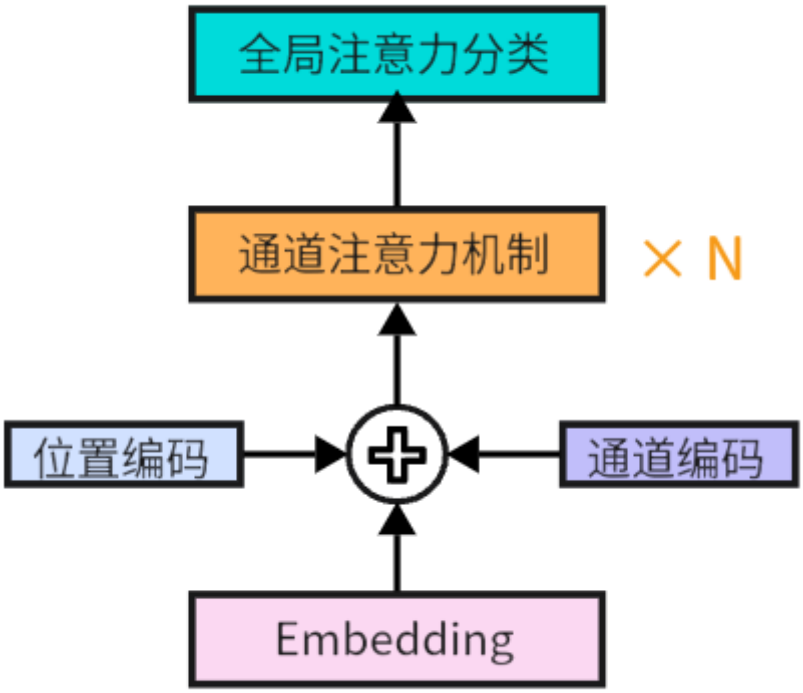
ark是一个基于多头注意力机制深层时序网络模型，目的是为了提出一种新的审核机制，以控制恶意的传播。

- 在数据处理上，ark将一个文本，分为**原始文本**，**拼音信息**，**声母信息**三通道，以应对谐音字代替关键字情况
- 在模型框架上，ark选择**多头注意力机制**，以捕捉道原语中携带的复杂信息。
- 在模型结构上，ark采用独特的**通道注意力机制**，每个单通道将会学到它与所有通道之间的关系。
- ark的名字来源于诺亚方舟$(Noah's\ Ark)$，寓意是带领人们走出网络恶意的方舟。

二. ark的模型结构设计

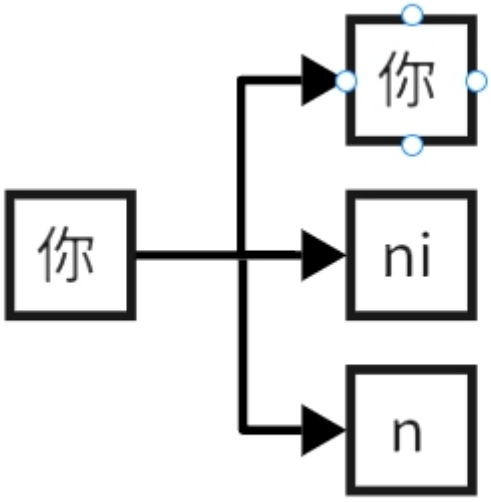
> ark主要由三大层组成

- **Embedding层**负责提取输入携带的原始信息。
- **通道注意力层**通过原始信息提取更加深层的信息。
- **全局注意力层**通过通道注意力得到的深层信息，将全部信息整合后分类。

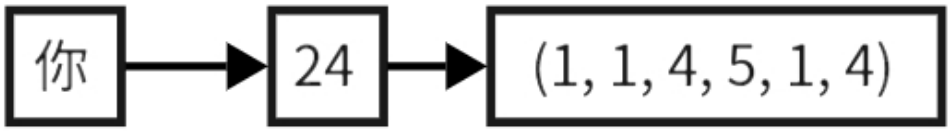


1. Embedding层

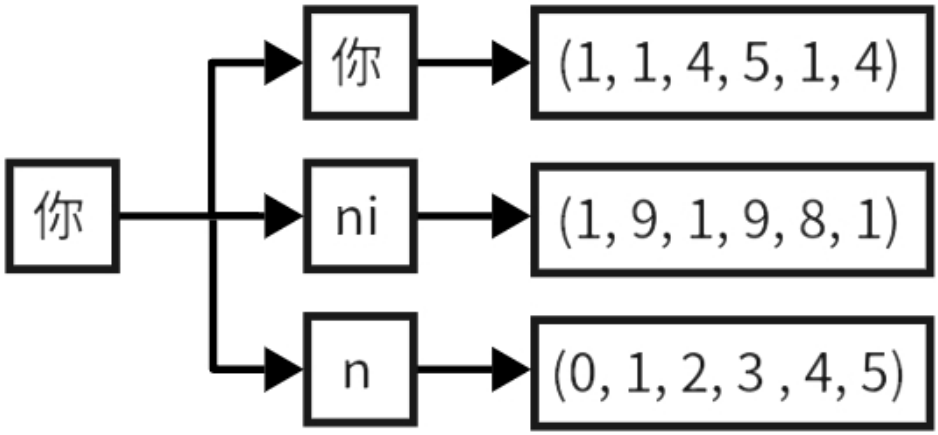
> 每个词元可以被分为三个通道，分别代表**原词，拼音，声母**



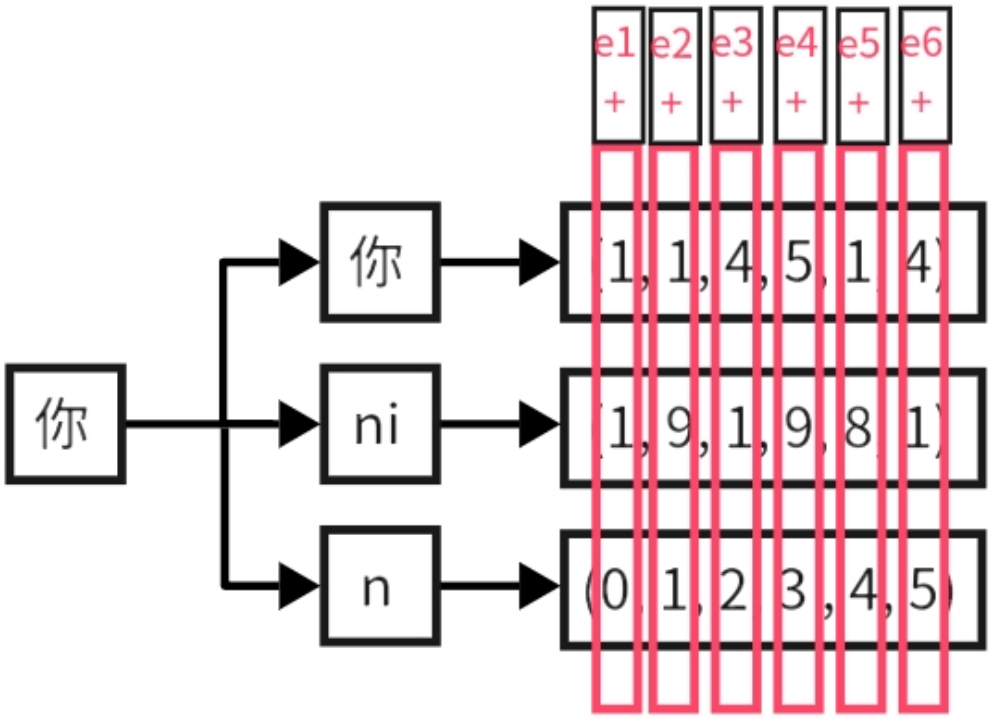
> 每个词元可以用唯一ID表示，并可以根据ID转化为**词向量**



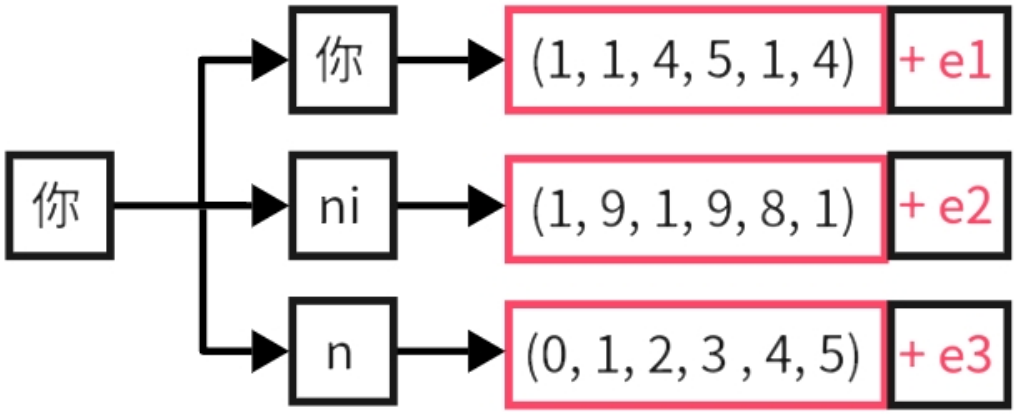
一个词元有多个通道，便有多个词向量表示不同维度的特征



为了保留位置信息，对词向量添加位置编码



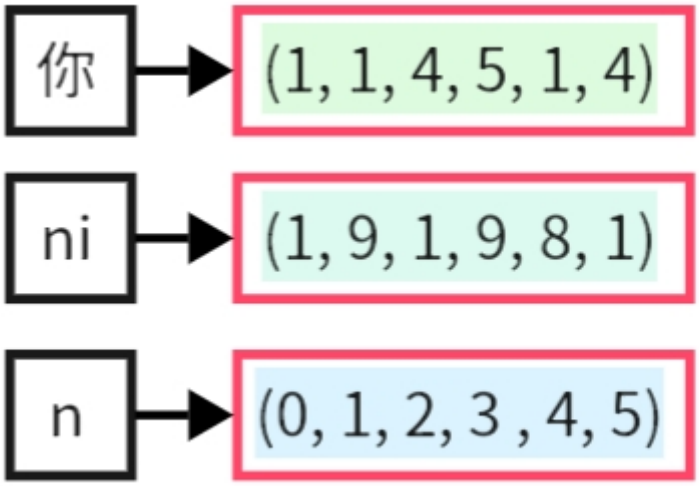
为了保留通道信息，对词向量添加通道编码



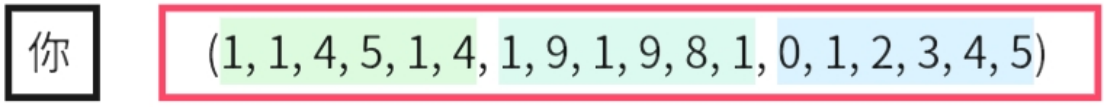
2. 通道注意力层

(1). 当前层的输入为**X**(通道数, 批量大小, 句子长度, 隐藏层大小)

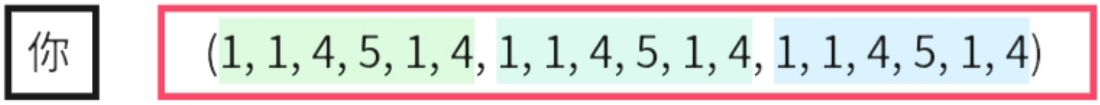
(2). 将多通道在句子长度的维度拼接得到**Q**(批量大小, 通道数 × 句子长度, 隐藏层大小)，即**每个通道的每个词元都会在多头注意力机制里查询一次**



(3). 将多通道在隐藏层维度拼接得到**KV**(批量大小, 句子长度，通道数 × 隐藏层大小)，即**每个词元此时都包含三通道的完整信息**



(4). 复制**Q**的隐藏层大小至通道数 * 隐藏层大小，使其维度与**KV**相同，以计算注意力权重，每次计算注意力权重时相当于一个通道的信息与三个通道的信息计算关联信息。



(5). **Q与KV进入**
$SelfAttention \rightarrow AddNorm \rightarrow multiHeadAttention \rightarrow AddNorm \rightarrow PositionWiseFFN$
，**得到结果O**$(批量大小, 通道数 \times 句子长度, 通道数 \times 隐藏层大小)$

(6). **将O展平为**$(批量大小, 通道数, 句子长度, 通道数, 隐藏层大小)$

(7). **将O的第1维度信息汇聚得到y**$(批量大小, 句子长度, 通道数, 隐藏层大小)$

(8). **将y形状重排为**$(通道数, 批量大小, 句子长度, 隐藏层大小)$, **重复操作**(1)

3. 全局注意力层

(1). **当前层的输入为X**$(通道数, 批量大小, 句子长度, 隐藏层大小)$

(2). **将第0维度和第2维度信息汇聚得到GX**$(批量大小, 1, 隐藏层大小)$

(3). **GX与X进入**
$SelfAttention \rightarrow AddNorm \rightarrow multiHeadAttention \rightarrow AddNorm \rightarrow PositionWiseFFN$
，**得到结果PY**$(批量大小, 1, 隐藏层大小)$

(4). **PY进入线性层分类得到结果Y**$(批量大小, 类别数)$

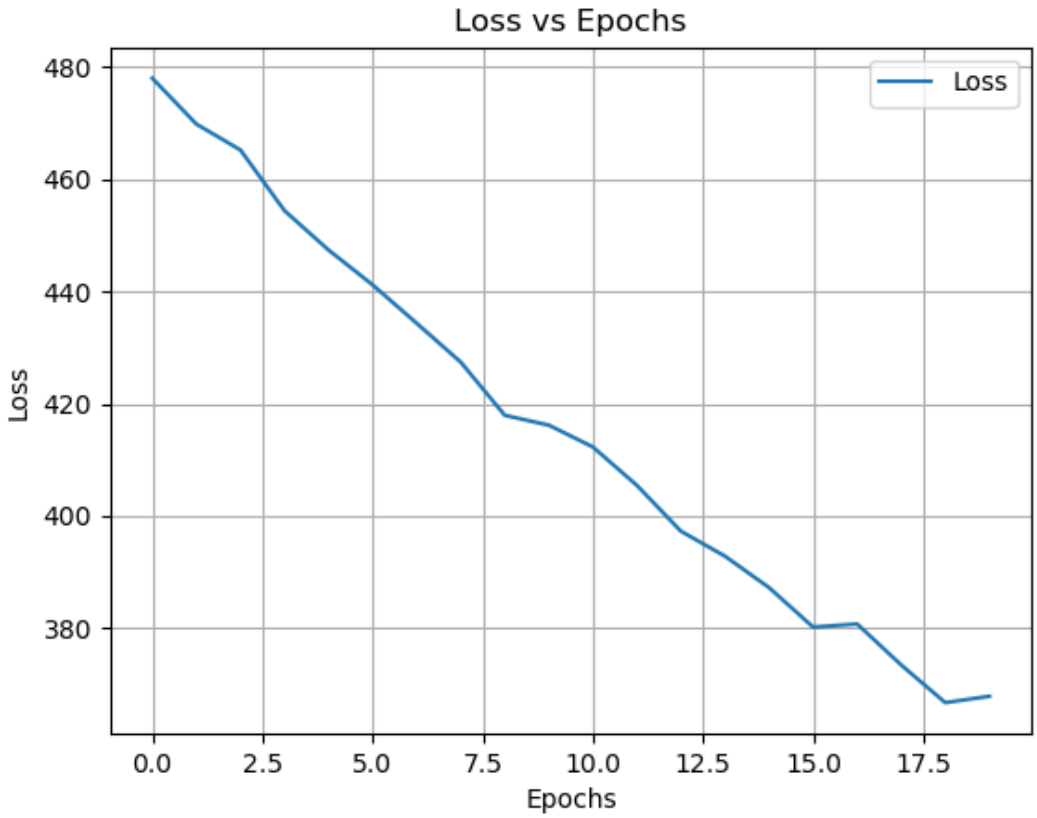
三. ark的训练结果与数据

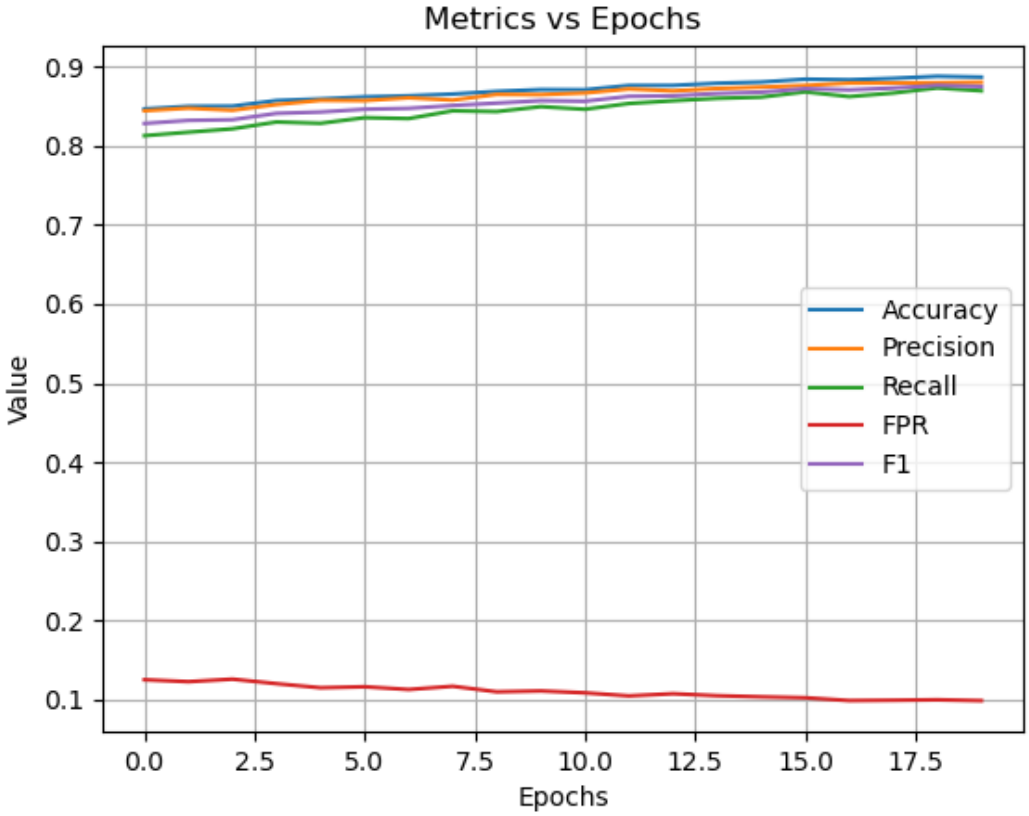
```
Epoch 1, Train A Epoch Total Loss: 478.0040
Epoch 1, Accuracy  0.846695, Precision  0.844362, recall  0.812835, FPR
0.125046, F1-score  0.828298

Epoch: 1, ValidMetrics:
Epoch 1, Accuracy  0.852516, Precision  0.869844, recall  0.815526, FPR
0.113177, F1-score  0.841810

Epoch 2, Train A Epoch Total Loss: 469.8315
Epoch 2, Accuracy  0.850049, Precision  0.847645, recall  0.817279, FPR
0.122601, F1-score  0.832185

Epoch: 2, ValidMetrics:
Epoch 2, Accuracy  0.853067, Precision  0.791934, recall  0.870675, FPR
0.159187, F1-score  0.829440

Epoch 3, Train A Epoch Total Loss: 465.2153
Epoch 3, Accuracy  0.850233, Precision  0.844963, recall  0.821521, FPR
0.125804, F1-score  0.833077

Epoch: 3, ValidMetrics:
Epoch 3, Accuracy  0.856651, Precision  0.795906, recall  0.875042, FPR
0.156148, F1-score  0.833600

Epoch 4, Train A Epoch Total Loss: 454.4017
Epoch 4, Accuracy  0.857294, Precision  0.852299, recall  0.830174, FPR
0.120072, F1-score  0.841091

Epoch: 4, ValidMetrics:
2Epoch 4, Accuracy  0.860786, Precision  0.821570, recall  0.863242, FPR
0.141063, F1-score  0.841891
```

```
Epoch 5, Train A Epoch Total Loss: 447.3882
Epoch 5, Accuracy  0.859346, Precision  0.857506, recall  0.828491, FPR
0.114902, F1-score  0.842749

Epoch: 5, ValidMetrics:
Epoch 5, Accuracy  0.859821, Precision  0.860373, recall  0.834123, FPR
0.117814, F1-score  0.847045

Epoch 6, Train A Epoch Total Loss: 441.1767
Epoch 6, Accuracy  0.861904, Precision  0.857217, recall  0.835628, FPR
0.116166, F1-score  0.846285

Epoch: 6, ValidMetrics:
Epoch 6, Accuracy  0.864921, Precision  0.859151, recall  0.844191, FPR
0.117482, F1-score  0.851605

Epoch 7, Train A Epoch Total Loss: 434.3322
Epoch 7, Accuracy  0.863283, Precision  0.860715, recall  0.834517, FPR
0.112710, F1-score  0.847414

Epoch: 7, ValidMetrics:
Epoch 7, Accuracy  0.863680, Precision  0.871372, recall  0.833918, FPR
0.109778, F1-score  0.852234

Epoch 8, Train A Epoch Total Loss: 427.4127
Epoch 8, Accuracy  0.865504, Precision  0.857827, recall  0.844281, FPR
0.116784, F1-score  0.851000

Epoch: 8, ValidMetrics:
Epoch 8, Accuracy  0.862440, Precision  0.869844, recall  0.832700, FPR
0.111053, F1-score  0.850867

Epoch 9, Train A Epoch Total Loss: 417.9241
Epoch 9, Accuracy  0.868873, Precision  0.865062, recall  0.843305, FPR
0.109787, F1-score  0.854045

Epoch: 9, ValidMetrics:
Epoch 9, Accuracy  0.860372, Precision  0.888176, recall  0.817952, FPR
0.098892, F1-score  0.851619

Epoch 10, Train A Epoch Total Loss: 416.1656
Epoch 10, Accuracy  0.871125, Precision  0.864766, recall  0.849567, FPR
0.110883, F1-score  0.857099

Epoch: 10, ValidMetrics:
Epoch 10, Accuracy  0.866023, Precision  0.848152, recall  0.853891, FPR
0.124126, F1-score  0.851012

Epoch 11, Train A Epoch Total Loss: 412.2815
Epoch 11, Accuracy  0.870926, Precision  0.866869, recall  0.846234, FPR
0.108467, F1-score  0.856427

Epoch: 11, ValidMetrics:
Epoch 11, Accuracy  0.866437, Precision  0.853346, recall  0.851005, FPR
0.120816, F1-score  0.852174
```

```
Epoch 12, Train A Epoch Total Loss: 405.4339
Epoch 12, Accuracy  0.876333, Precision  0.871917, recall  0.853540, FPR
0.104645, F1-score  0.862631

Epoch: 12, ValidMetrics:
Epoch 12, Accuracy  0.866437, Precision  0.825542, recall  0.871613, FPR
0.137425, F1-score  0.847952

Epoch 13, Train A Epoch Total Loss: 397.2579
Epoch 13, Accuracy  0.876424, Precision  0.869479, recall  0.857008, FPR
0.107371, F1-score  0.863199

Epoch: 13, ValidMetrics:
Epoch 13, Accuracy  0.867677, Precision  0.852429, recall  0.853994, FPR
0.121113, F1-score  0.853211

Epoch 14, Train A Epoch Total Loss: 392.8095
Epoch 14, Accuracy  0.879151, Precision  0.872519, recall  0.860005, FPR
0.104870, F1-score  0.866217

Epoch: 14, ValidMetrics:
Epoch 14, Accuracy  0.869745, Precision  0.865872, recall  0.848503, FPR
0.112133, F1-score  0.857100

Epoch 15, Train A Epoch Total Loss: 387.2073
Epoch 15, Accuracy  0.880637, Precision  0.874240, recall  0.861553, FPR
0.103437, F1-score  0.867850

Epoch: 15, ValidMetrics:
Epoch 15, Accuracy  0.861199, Precision  0.888482, recall  0.819155, FPR
0.098516, F1-score  0.852411

Epoch 16, Train A Epoch Total Loss: 380.1710
Epoch 16, Accuracy  0.884313, Precision  0.876308, recall  0.868254, FPR
0.102285, F1-score  0.872262

Epoch: 16, ValidMetrics:
Epoch 16, Accuracy  0.867126, Precision  0.863123, recall  0.845555, FPR
0.114461, F1-score  0.854249

Epoch 17, Train A Epoch Total Loss: 380.7224
Epoch 17, Accuracy  0.883531, Precision  0.879374, recall  0.862261, FPR
0.098716, F1-score  0.870733

Epoch: 17, ValidMetrics:
Epoch 17, Accuracy  0.865472, Precision  0.868622, recall  0.838891, FPR
0.111226, F1-score  0.853497

Epoch 18, Train A Epoch Total Loss: 373.4108
Epoch 18, Accuracy  0.885354, Precision  0.879475, recall  0.866772, FPR
0.099137, F1-score  0.873078

Epoch: 18, ValidMetrics:
```

```
Epoch 18, Accuracy  0.867402, Precision  0.813932, recall  0.882996, FPR
0.143700, F1-score  0.847059

Epoch 19, Train A Epoch Total Loss: 366.7032
Epoch 19, Accuracy  0.887989, Precision  0.879664, recall  0.873237, FPR
0.099699, F1-score  0.876438

Epoch: 19, ValidMetrics:
Epoch 19, Accuracy  0.869056, Precision  0.868928, recall  0.845171, FPR
0.110283, F1-score  0.856885

Epoch 20, Train A Epoch Total Loss: 367.8412
Epoch 20, Accuracy  0.886932, Precision  0.880390, recall  0.869600, FPR
0.098603, F1-score  0.874962

Epoch: 20, ValidMetrics:
Epoch 20, Accuracy  0.867264, Precision  0.882371, recall  0.833237, FPR
0.101610, F1-score  0.857100
```





四. ark的不足

1. ark的数据集$(7w+)$仍需要完善，数据面不够涵盖多方面的知识，好的数据集仍然是好模型的基础
2. ark的上下文只支持128，对于长序列ark暂时无法支持