江苏省人工智能学会
JiangSu Association of Artificial Intelligence

HUAWEI

DIGIX 极客 | 算法精英大赛路演

用户人口属性预测

Nurbs

Background

> 参赛同学根据华为提供数据构建预测模型进行年龄段预估，在测试数据集上给出预估结果。

25G

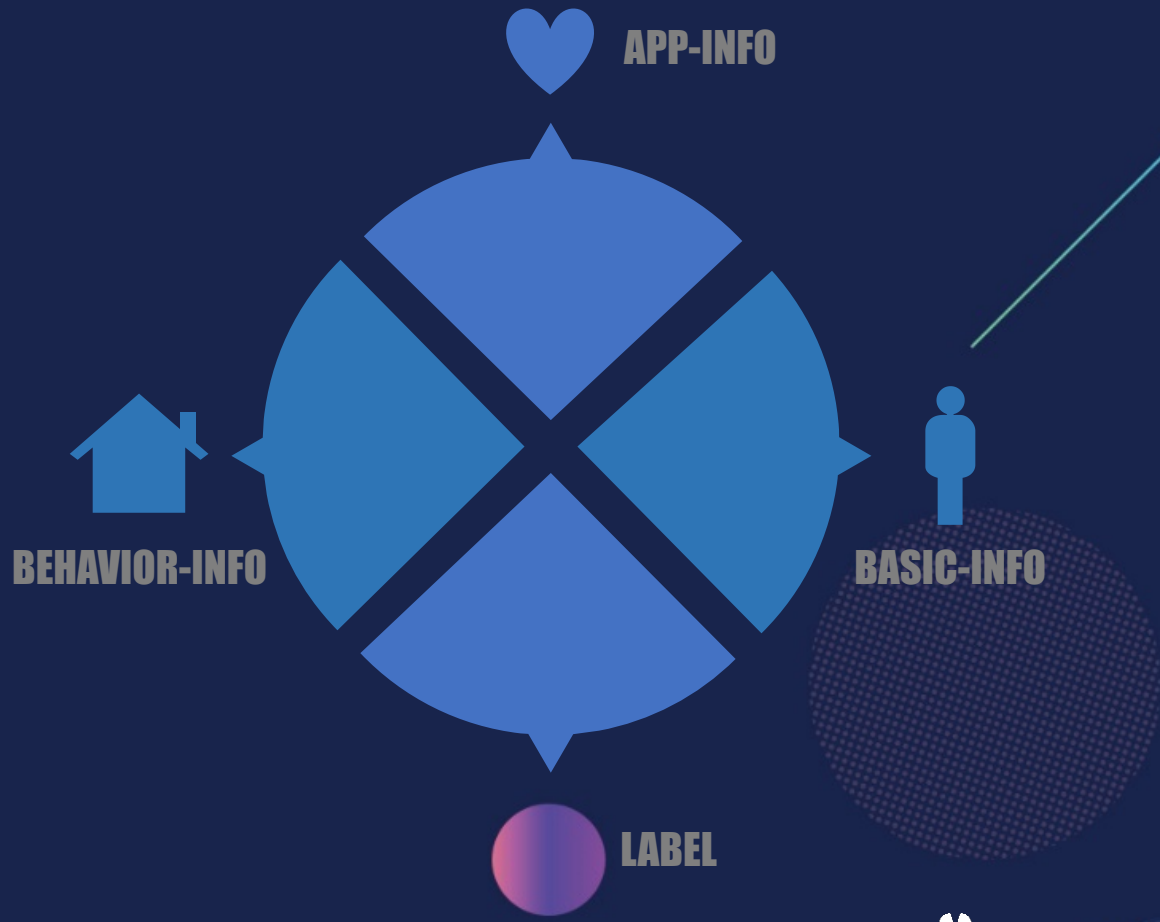
Table - USAGE

30天内按天统计每个用户对具体某个app的累计打开次数和使用时长

1.0G

Table - ACTIVE

用户的激活APP列表文件，每一行代表一条用户激活app的记录



Summary

MULTI-CLASS
含单调性的多分类

MULTI-TABLE
多源联合的表格数据

MULTI-MODAL
多模态的应用场景

END-TO-END Solution

Maximize business value, Extremely fast commercial landing, Avoid over-fitting caused by model stacking.

Solution

Traditional
LR SVR BAYES TREE etc.

NLP-Deep
TextCNN Capsule BiLSTM-Att

CTR-Deep
xDeepFM FGCNN DIEN

Rich semantic mining. Refinement of APP Implied Dependencies. Prolific information on different modeling angles.

推荐、分类

对不同的业务逻辑思考，抽象出可能产生商业价值的场景

Recommend
基于华为设备的社交网络做到同年龄阶层社交推荐、垂直行业目标客户的精准营销

Huawei-based social networking to achieve social networking recommendations of the same age, and precise marketing of target customers in different industries

Classification
手机用户群画像分析，帮助厂商了解产品的人群定位，深度挖掘APP与用户的交互关系

Analysis of mobile phone user group images helps manufacturers understand the crowd positioning of products and deeply explore the interaction between APP and users.

Traditional-Text FE

Sparse Matrix FE

TFIDF Sparse Matrix
COUNT Sparse Matrix
HASH Sparse Matrix

Topic FE

LDA LSI pLSA
SVD NMF
Word2Vec Fasttext Glove

Match FE

V2V Cosine
MV-LSTM
ESIM

Input sparse matrix input linear model learning
Enter the topic model into the tree model to learn nonlinear expression

Traditional-Text FE

Acc 53%

Only Use TFIDF

NGRAM = (1,1)

Acc 59%

Only Use COUNT

NGRAM = (1,1)

Acc 60%

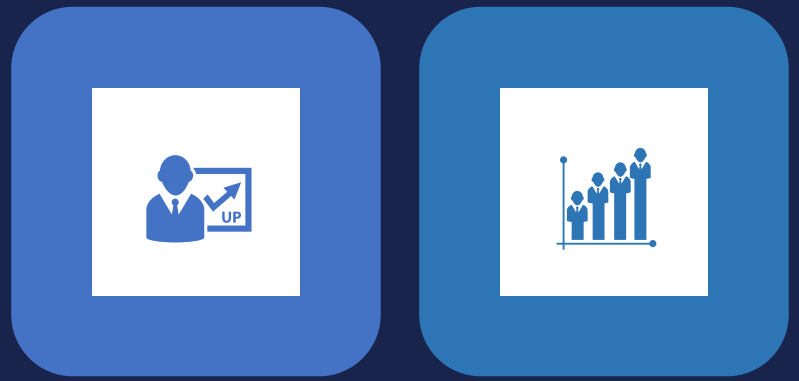
COUNT concat TFIDF

NGRAM = (1,1)

Traditional-Tabular FE

Category Encoder

- Count Encoder
  - transform(count)
- Rank Encoder
  - Rank(cat)
- Cross Combine
  - Label Encoder Cross
  - Count Encoder Cross



USER base

- 第一次/最后一次使用时间
- APPID Sequence Feature:
  - Gini、Entropy
  - Nunique、Count
- TIMES/DURATION/USE_DATE序列统计值
  - Mean、STD、Min、Max
  - Skew、Kurt、Mad

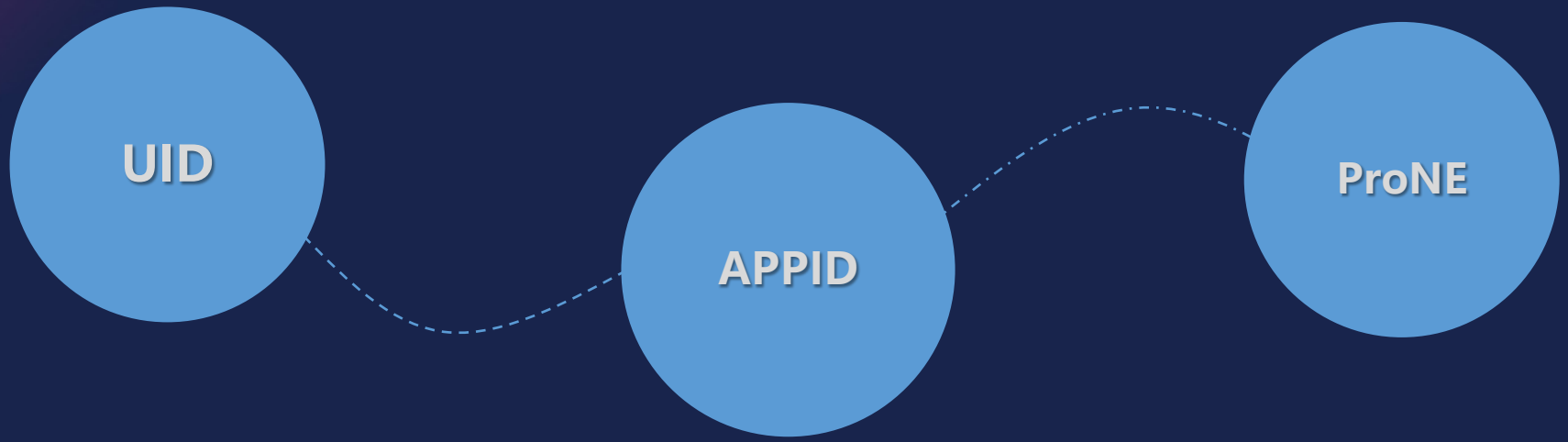
APPID base

- Transform category info
- Count编码后，高低频刻画
- 小众App的差异性捕捉

Graph base

- UID-APPID Bipartite (SIGIR 2018 Lei)
- UID-APPID Sequence (KDD 2018 Alibaba)
  - Diff2Vec (CCN 2018 Benedek)
  - ProNE (IJCAI 2019 Tsinghua)
- PageRank

Graph-Base FE



BUILD GRAPH

1. Bipartite Graph
2. Time based APPID-Graph

Graph-Base FE

EVCENT

Calculate the vector centrality of the nodes in the graph.

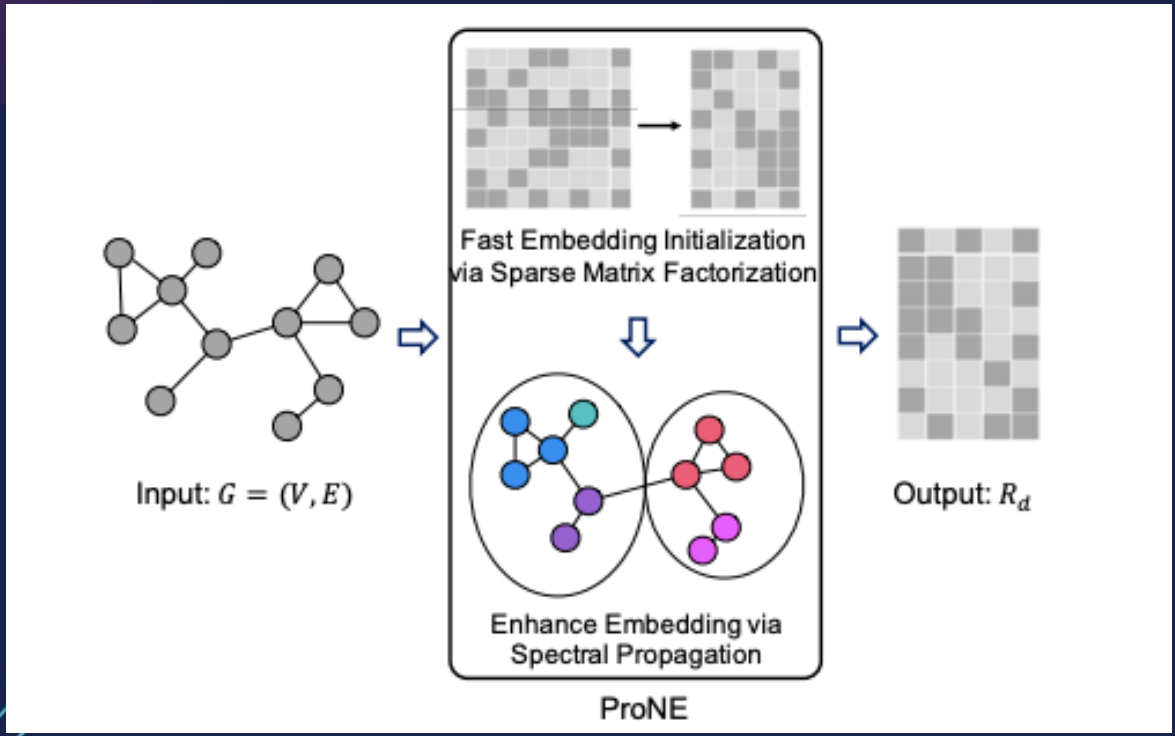
SHELL INDEX

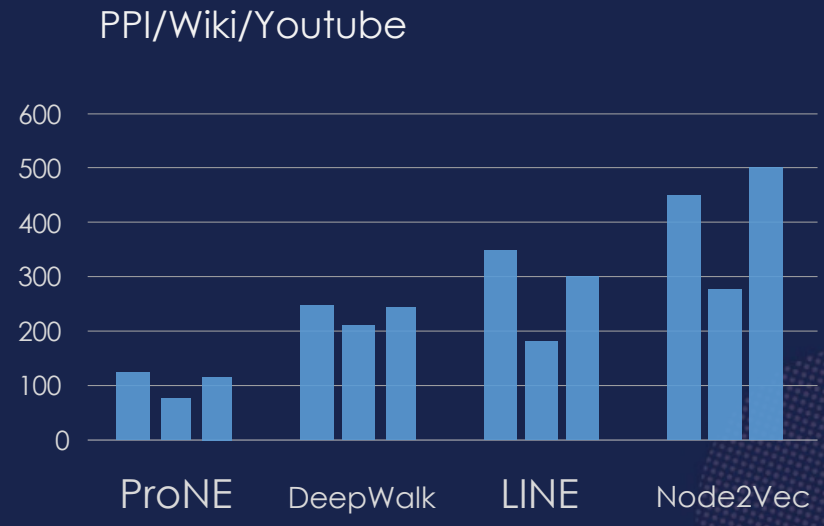
Calculate the largest subgraph with a node degree of at least K in the graph

PAGERANK

Calculate the PageRank value of the node in the graph

ProNE Embedding





Traditional-Setting

Feature	CV Score
Raw Feature	0.523
+ Sparse Matrix FE	0.633 +
+ Topic FE	0.642 +
+ Match FE	0.645 +
+ Category Encoder	0.645 -
+ USER base	0.647 +
+ APPID base	0.647 +
+ Graph base	0.649 +
+ Hyper Param	0.649 +

NLP-Network

Text CNN

We show that a simple CNN with little hyperparameter tuning and static vectors achieves excellent results on multiple benchmarks. Learning task-specific vectors through fine-tuning offers further gains in performance. We additionally propose a simple modification to architecture to allow for the use of both task-specific static vectors.

Capsule Network

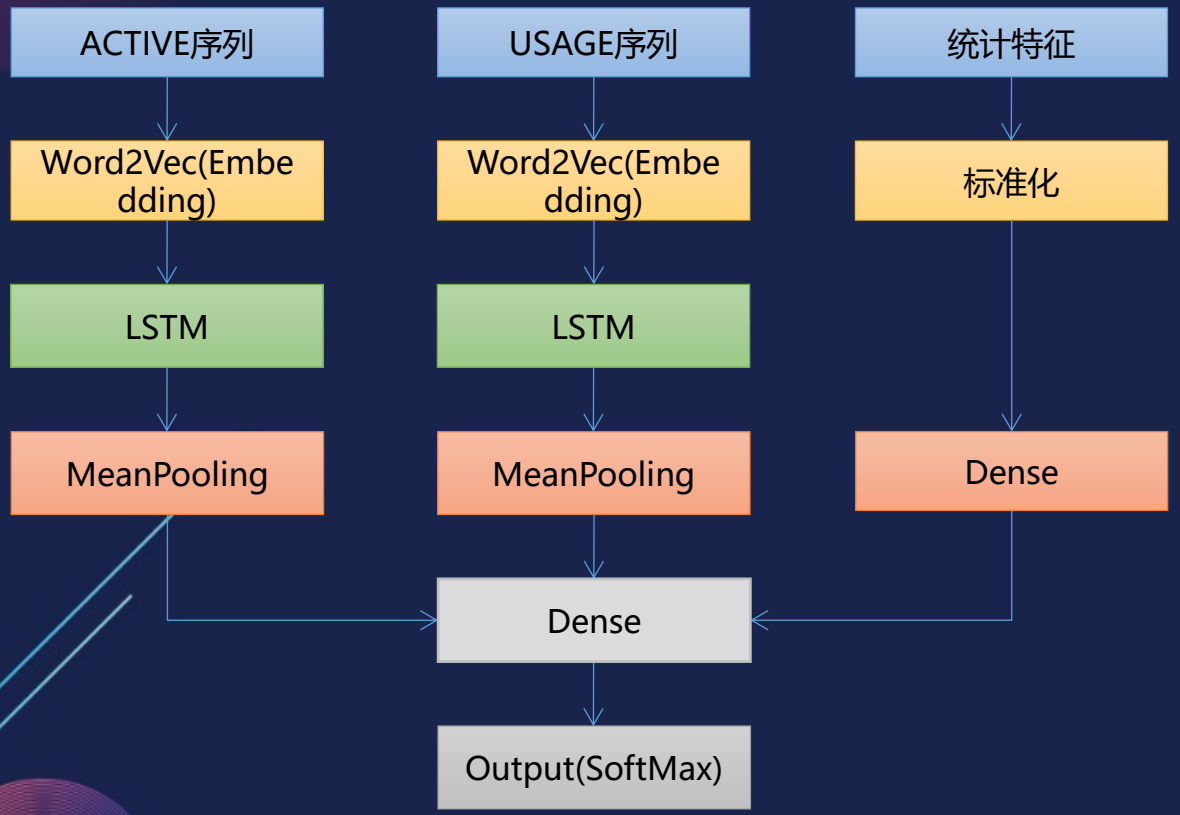
A capsule is a group of neurons whose activity vector represents the instantiation parameters of a specific type of entity such as an object or an object part. We use the length of the activity vector to represent the probability that the entity exists and its orientation to represent the instantiation parameters

Bi-LSTM With Attention

We examine a BiLSTM architecture with attention mechanism (BiLSTM-Attention) and a LSTM architecture with attention mechanism (LSTM-Attention), and try different dropout rates based on these two models. We then exploit an ensemble of these methods to give the final prediction which improves the model performance.

NLP-Network

Multi Context Network

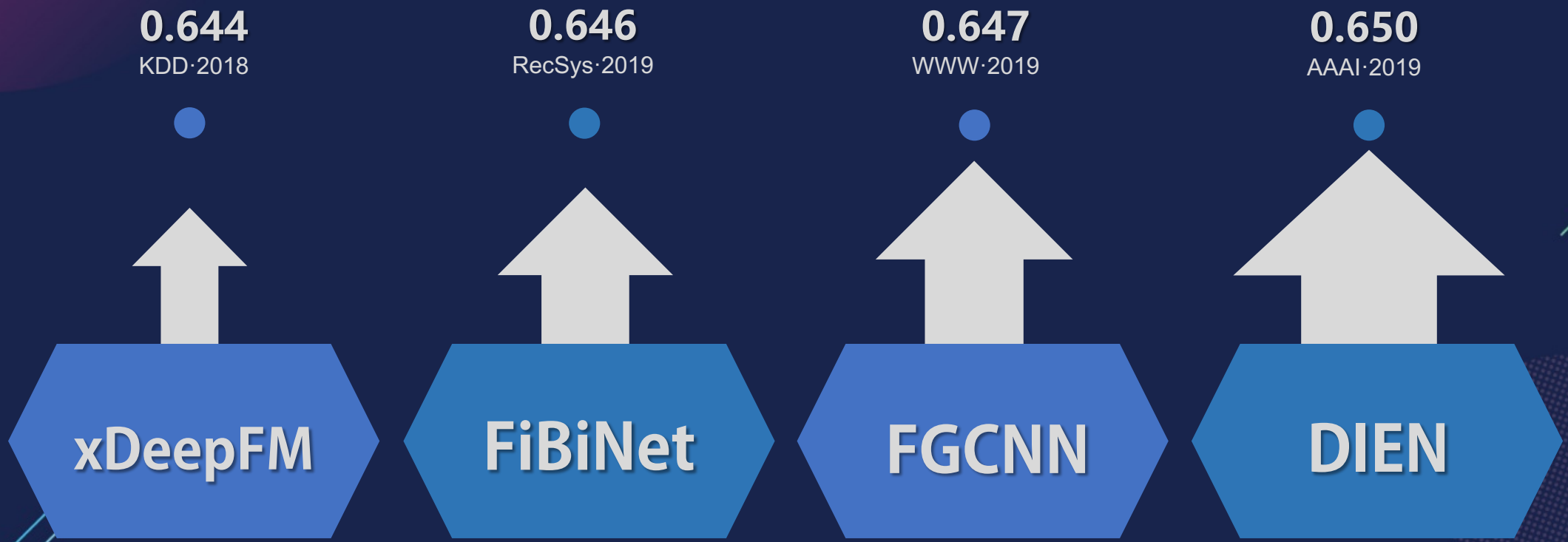


优点及创新：

1、手工特征少

2、借鉴nlp建模思路，将app序列和使用表序列看成一篇文档作词嵌入

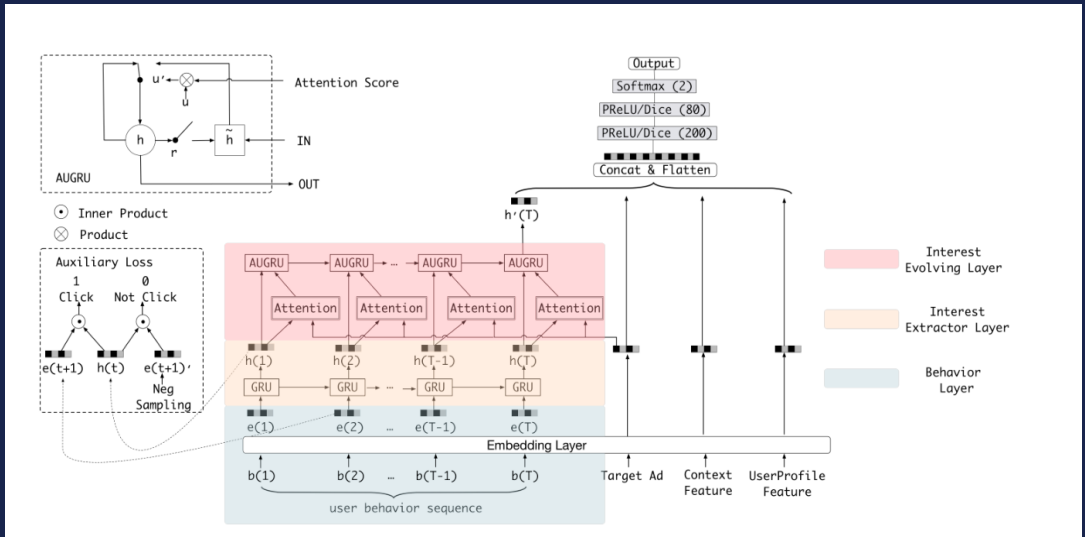
CTR-Network


0.644
KDD·2018
0.646
RecSys·2019
0.647
WWW·2019
0.650
AAAI·2019
xDeepFM
FiBiNet
FGCNN
DIEN

CTR-Network

DIEN

Interest



兴趣发展过程

细粒度时序捕捉兴趣

辅助损失

Parameter-Tuning

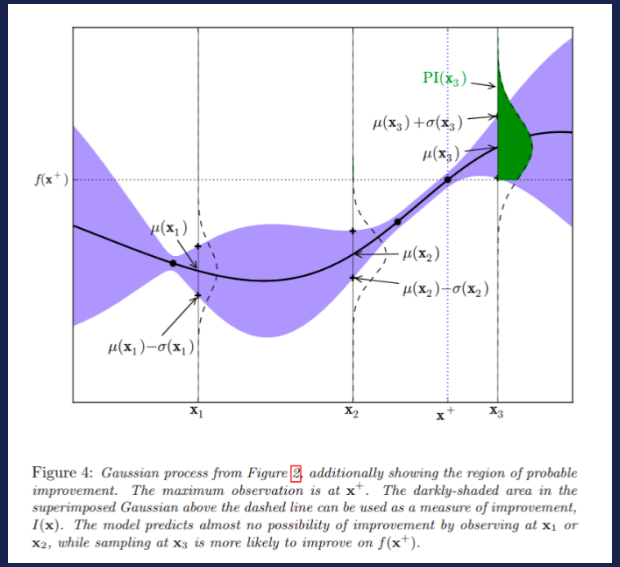


Figure 4: *Gaussian process from Figure 2, additionally showing the region of probable improvement. The maximum observation is at $\mathbf{x}^+$. The darkly-shaded area in the superimposed Gaussian above the dashed line can be used as a measure of improvement, $I(\mathbf{x})$. The model predicts almost no possibility of improvement by observing at $\mathbf{x}_1$ or $\mathbf{x}_2$, while sampling at $\mathbf{x}_3$ is more likely to improve on $f(\mathbf{x}^+)$.*

Bayesian optimization for machine learning tuning is proposed by J. Snoek (2012). The main idea is that given the objective function of the optimization (generalized function, you only need to specify the input and output, without knowing the internal structure and the mathematical properties.)

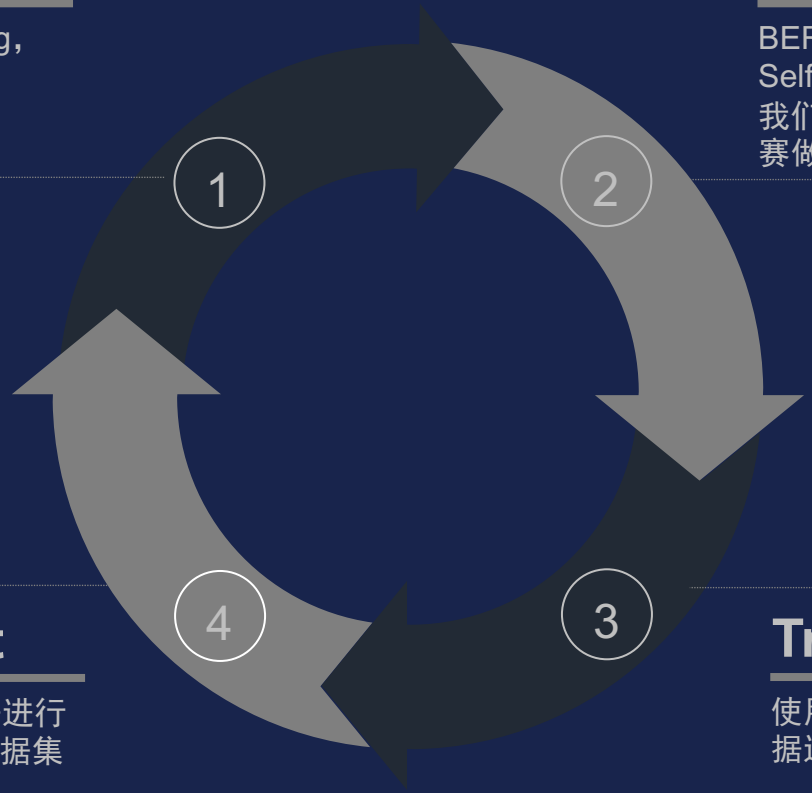
FINE-Tuning

Try learning BERT

尝试去训练基于APPID语料的Embedding，但由于效率问题，没有尝试成功.

Thinking about BERT

BERT本质上是多层预训练好的Multi-Head Self-Attention对新CASE去FINE-Tuning，那我们也可以基于初赛数据的领域知识，对复赛做FINE-Tuning.



FINE-Tuning B Dataset

对A Dataset的网络及其权重对复赛数据进行微调，较大避免了同时训练A、B两个数据集分布的差异性导致模型学习有偏.

Train A Dataset

使用 LSTM-GRU 双层Attention模型对初赛数据进行训练，得到网络及其权重.

Meta-Model

GroupBY

大量损失信息

Why Not Train

使用有监督的学习代替无监督的聚合方法

NN - GroupKFold

使用GroupbyKFold防止过拟合,在展开的表上进行训练.可以大量构建基于APPID的特征

Siamese Network

在ACTIVE和USAGE借鉴孪生网络，共享Embedding层参数

COMPLEXITY-Optimization

IGRAPH

使用IGRAPH提取图特征，效率较NetWorkX提升7倍

Generating

使用生成迭代器的方式优化内存使用直接拟合容易导致内存溢出

Streaming

流式的存储向量化、序列化特征，大大降低每次I/O的所需时间

CuPy

使用CUDA上实现CuPy代替Numpy，其存储及矩阵运算上效率更快

NETWORK-Optimization

In addition to explaining why we should use warmup, we also propose RAdam, a theoretically sound variant of Adam.

RAdam[3]

We empirically demonstrate Lookahead can significantly improve the performance of SGD and Adam.

Lookahead[2]

Training with cyclical learning rates instead of fixed values achieves improved classification accuracy without a need to tune and often in fewer iterations.

CyclicLR[1]

Input the APPID sequence input of the Active and Usage tables into two-channel information.

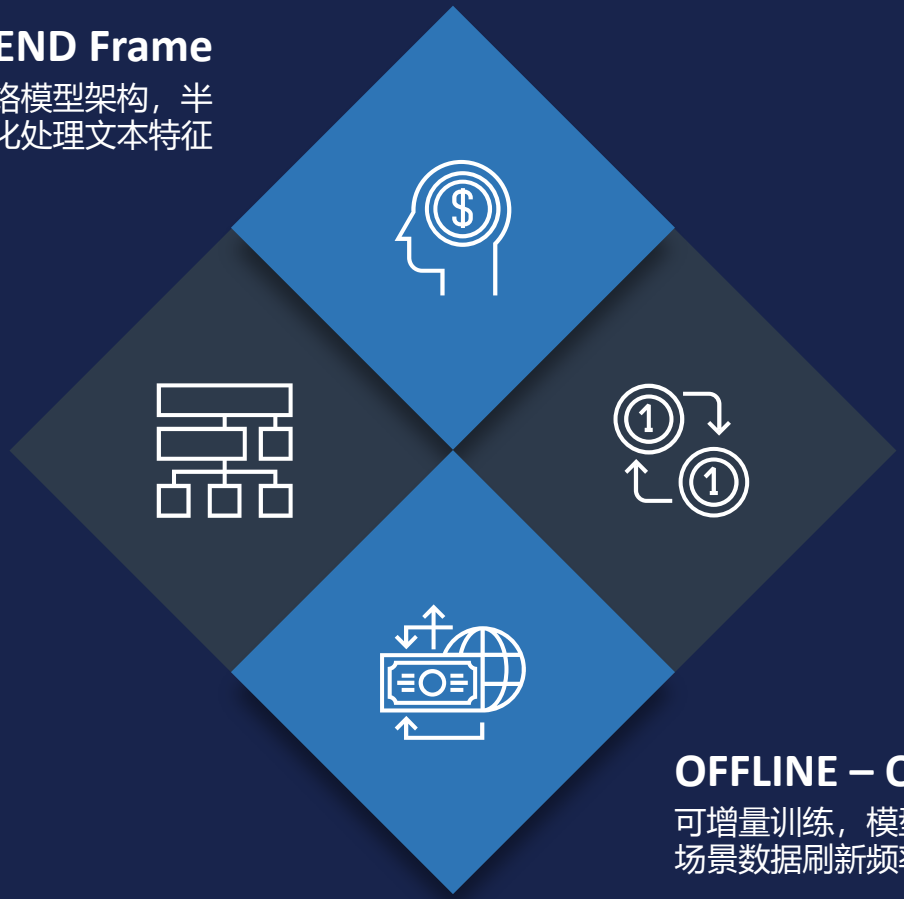
Multi Input

1. https://arxiv.org/abs/1506.01186
2. https://arxiv.org/abs/1907.08610
3. https://arxiv.org/abs/1908.03265

Advantage

END - END Frame

实现端到端的网络模型架构，半自动流程化处理文本特征



Multi-Model Learning

CTR、 NLP、 GBDT的多模态学习，多尺度的特征抽取

Transfer Learning

使用迁移学习的思想，对数据集进行微调，降低分布不一致的影响

OFFLINE – ONLINE

可增量训练，模型训练速度匹配业务场景数据刷新频率

HUAWEI

请评委老师批评指正