Tencent Advertisement Algorithm Competition
Deep Interest FFM and Gradient Boosting Approach

He Jiang[1], Yu Shi[2], Hongyi Huang[2]

[1]Department of Automation, Tsinghua University [2]IIIS, Tsinghua University

Introduction

Lookalike problem aims at discovering potential customs for products. In this Tencent Advertisement Algorithm competition, we are supposed to predict whether a user is interested in an advertisement. This problem is indeed reduced to a click-through-prediction problem.

Prevalent models for this problem include Logistic Regression(LR), Gradient Boost Decision Tree(GBDT), Factorization Machine(FM), Field-aware Factorization Machine(FFM) and different types of neural networks(NN). We create a new model by introducing attention mechanism to a combination of them.

Our model successfully passes preliminary contest (200 teams out of more than 1000 teams) and currently ranks the 13th place in the final competition.

Problem

We are given millions of users and hundreds of advertisements with their features, such as age, gender of a user and product type of an advertisement. Single feature field may have multiple feature values(e.g. a user can have many interests).

Training data is a list of triplets (user,advertisement,label). All feature values are categorical and encrypted. There are more than $450,000$ different feature values and many of them are of very low frequency.

Final contest becomes much more difficult as the amount of data increases (Table 1).

Table: Sizes of dataset.

	users	ads	training	test
preliminary	$1.1 * 10^7$	173	$8.8 * 10^6$	$2.3 * 10^6$
final	$4.4 * 10^7$	832	$4.6 * 10^7$	$1.2 * 10^7$

Our goal is predict the label given (user, advertisement) pair. The final result is evaluated by AUC metric.

Method

We tried GBDT and create a combination model of Deep FFM/FM and Deep Interest Network named Deep Interest FFM.

Deep Interest FFM

One challenge in this competition is that each user can have multiple values in a single feature, e.g. a user can have 5 different key words and hundreds of apps recently installed. Each feature value is categorical and requires embedding in FFM. To get the final embedding vector of a multi-value feature, a naive way is doing mean pooling over embedding vectors of different values. We propose to apply attention module of Deep Interest Network to get a weighted sum of these embedding vectors. We call our model Deep Interest FFM.

We always adopt binary cross entropy loss or its variants, like weighted version or focal loss.

$$L = \sum_n \log(1 + exp(-y_n f(u_n, a_n))) \quad (1)$$

FM tries to capture second-order interaction between features. Features are represented by a vector in an embedding space. The weight of second-order term is modeled by inner product of two vectors. It provides an efficient algorithm and addresses the problem of sparsity of second-order terms.

$$f(u, a) = \sum_{i<j} \langle v_i, v_j \rangle x_i x_j + \sum_i w_i x_i + w_0 \quad (2)$$

FFM[2] introduces the concept of field-awareness, enlarging the capacity of FM. Field-awareness simply means when interacted with features of different fields (like age and gender), a certain feature is represented differently.

$$f = \sum_{i<j} \langle \mathbf{v_{i,(field_j)}}, \mathbf{v_{j,(field_i)}} \rangle x_i x_j + \sum_i w_i x_i + w_0 \quad (3)$$

DeepFM[1] model combines an FM model with a neural network model. The vector embedding are shared and two models are trained jointly. We replace the FM module in DeepFM with FFM. For multi-value features, before feeding its embedding vectors to DeepFFM, we use an attention module to combine the multiple embedding vectors as in Deep Interest Network(DIN)[3]. Deep Interest Network(DIN)[3] learns the vector representations of users' interests with attention mechanism. The attention is a function of raw embedding of advertisements' features and users' interests.

$$v_u = \sum_i w_i v_i = \sum_i g(v_i, v_a) v_i \quad (4)$$

Our best single neural network model is a combination of DeepFFM and DIN (see Figure 1). We call it Deep Interest FFM.

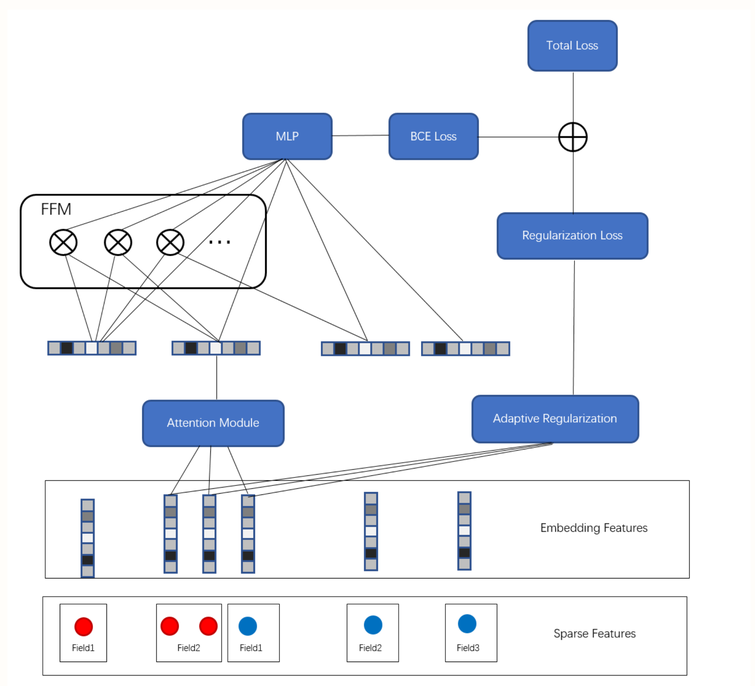

Figure: Deep Interest Neural FFM.

Regularization in such network is difficult. Since we can have many embedding vectors corresponding to large number of categorical values. To avoid overfitting, we use the adaptive regularization of Deep Interest Network. At each mini-batch, only the norm of embedding vectors of categorical values appearing in this mini-batch is penalized.

$$L_2(W) \approx \sum_{j=1}^{K} \sum_{m=1}^{B} \frac{\alpha_{mj}}{n_j} \|w_j\|_2^2 \quad (5)$$

Gradient Boosted Decision Trees

We use the GBDT toolkit LightGBM. GBDT requires more feature engineering work. Our features can be divided into threes categories:

1. One-hot encoding of raw features.
2. Combined one-hot feature for each pair of user-ad features.
3. Click through rate (CTR) for each pair of user-ad features.

By combining the three features above, we get more than 100,000 sparse features, and hundreds of dense features. In the preliminary contest, the CTR features cause overfitting easily. In the final competition, we divide the dataset into 4 parts and calculate CTR in a cross-validation fashion. In this way, the CTR of each training sample will not use its own label, thus reduces the risk of overfitting.

Results and Discussions

In the preliminary contest, we show improvements of models in Table 2.

Table: Different model's performance on preliminary contest.

Model	AUC
baseline1 (MLP)	0.724
baseline2 (MLP)	0.727
Deep FM	0.732
add orthogonal initialization	0.736
add focal loss	0.737
add low-frequncy filter	0.738
add attention and adaptive regularizatoin	0.745
add learning rate decay	0.748
add BN	0.749
train on more data	0.751
LightGBM add cross features	0.743
LightGBM ensemble	0.748
LightGBM and NN	0.751
bagging	0.755
Deep Interest Neural FFM	0.752

In the final round, we mainly use Deep Interest Neural FFM, because it outperforms other models by a large margin. We train this model on different subsets of datasets with different subsets of features and then average the results. GBDT will be used as auxiliary module for our final ensemble. We also show the effectiveness of adaptive regularization in Figure 2.

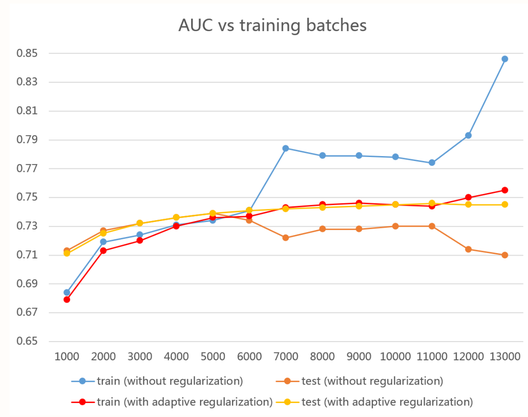

Figure: Regularization.

Summary and Conclusions

We design a new model Deep Interest FFM for advertisement prediction. The model handles multi-value features using attention, which will focus on only properties of a user related to the advertisement. Adaptive regularization is applied to alleviate overfitting. We also train a GBDT model for ensemble. Currently we rank 13^{th} on the leaderboard.

References

[1] H. Guo, R. Tang, Y. Ye, Z. Li, and X. He.
Deepfm: A factorization-machine based neural network for ctr prediction.
arXiv preprint arXiv:1703.04247, 2017.

[2] Y. Juan, Y. Zhuang, W.-S. Chin, and C.-J. Lin.
Field-aware factorization machines for ctr prediction.
In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 43–50. ACM, 2016.

[3] G. Zhou, C. Song, X. Zhu, X. Ma, Y. Yan, X. Dai, H. Zhu, J. Jin, H. Li, and K. Gai.
Deep interest network for click-through rate prediction.
arXiv preprint arXiv:1706.06978, 2017.