

Multi-Level Deep Cascade Trees for Conversion Rate Prediction in Recommendation System

Hong Wen,¹ * Jing Zhang,^{2,3} * Quan Lin,¹ Keping Yang,¹ † Pipei Huang¹

¹ Alibaba Group

² Hangzhou Dianzi University

³ University of Technology Sydney

{qinggan.wh, pipei.hpp}@alibaba-inc.com

{tieyi.lq, shaoyao}@taobao.com, jing.zhang@uts.edu.au

Abstract

Developing effective and efficient recommendation methods is very challenging for modern e-commerce platforms. Generally speaking, two essential modules named “Click-Through Rate Prediction” (*CTR*) and “Conversion Rate Prediction” (*CVR*) are included, where *CVR* module is a crucial factor that affects the final purchasing volume directly. However, it is indeed very challenging due to its sparseness nature. In this paper, we tackle this problem by proposing multi-Level Deep Cascade Trees (*ldcTree*), which is a novel decision tree ensemble approach. It leverages deep cascade structures by stacking Gradient Boosting Decision Trees (*GBDT*) to effectively learn feature representation. In addition, we propose to utilize the cross-entropy in each tree of the preceding *GBDT* as the input feature representation for next level *GBDT*, which has a clear explanation, i.e., a traversal from root to leaf nodes in the next level *GBDT* corresponds to the combination of certain traversals in the preceding *GBDT*. The deep cascade structure and the combination rule enable the proposed *ldcTree* to have a stronger distributed feature representation ability. Moreover, inspired by ensemble learning, we propose an Ensemble *ldcTree* (*E-ldcTree*) to encourage the model’s diversity and enhance the representation ability further. Finally, we propose an improved Feature learning method based on *EldcTree* (*F-EldcTree*) for taking adequate use of weak and strong correlation features identified by pre-trained *GBDT* models. Experimental results on off-line data set and online deployment demonstrate the effectiveness of the proposed methods.

Introduction

With the explosive growth of information available online, Recommender System (*RS*), as a useful information filtering tool, is used for estimating users’ preferences on items they have not seen and guides them to discover products or services they might be interested in from massive possible options. In general, recommender systems are classified into the following three categories based on the forms of recommendations (Balabanović and Shoham 1997): Collaborative recommendations, Content-based recommendations

*The co-first authors contributed equally.

†Corresponding author.

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

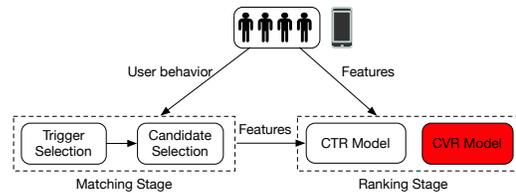


Figure 1: The framework for online recommendation in our E-commerce platform.

and Hybrid recommendations. Collaborative recommendations make users recommended items that people with similar tastes preferred in the past. Content-based recommendations make users recommended items similar to the ones the user preferred in the past. Hybrid recommendations integrates two or more types of recommendation strategies, which helps to avoid certain limitations of single strategy (Adomavicius and Tuzhilin 2005).

Recommender System increasingly plays an essential role in industry area, which promotes services for many applications (Gomez-Uribe and Hunt 2016; Davidson and Liebold 2010). In addition, in order to help customers find exactly what they need, recommendation techniques have been studied and deployed extensively on E-commerce platforms, which provides good user experience and promotes incredible increment in revenue. Usually, the deployed framework for our online E-commerce platforms is illustrated in Fig. 1. Specifically, when a user visits it through a terminal, such as smart phones, the system firstly analyzes his/her long and short term behaviors and then his/her interested items, called *Triggers*, are selected. Then, massive items closely related with *Triggers* are generated. Further, top K (e.g., 500) of them (based on the “*matching score*”), along with extra information (e.g., user, item, user-item cross features), are delivered to the next *Ranking* stage where it mainly contains two core modules, namely *CTR* and *CVR*. Finally, the recommendation results are generated and displayed to the user. In this paper, we mainly focused on the *CVR* module.

In the past few decades, deep learning has been witnessed the tremendous successes in many application areas (Zhang, Yao, and Sun 2017; Liu et al. 2017), such as image classification (Krizhevsky, Sutskever, and Hinton 2012;

Simonyan and Zisserman 2014), speech recognition (Deng et al. 2010; 2013) and object detection (Girshick 2015; Ren et al. 2015). Meanwhile, recent studies also demonstrate its efficiency and effectiveness in coping with recommendation tasks (Elkahky, Song, and He 2015; Wang, Wang, and Yeung 2015; Chen et al. 2017; He et al. 2017; Huang et al. 2015; Yang et al. 2017; Guo et al. 2017). Though deep learning has been partially overcoming obstacles of conventional models and gaining momentum due to its state-of-the-art performances, it has apparent deficiencies, such as a huge amount of data and powerful computational facilities required for training, more importantly many hyper-parameters to be tuned. Recently, *gcForest* (Zhou and Feng 2017), an alternative to *DNN*, is proposed, which generates a deep forest ensemble, with a cascade structure to do representation learning. In addition, it achieves highly competitive performance compared with *DNN* for various domains' tasks while having fewer hyper-parameters.

In this paper, partially inspired by *gcForest* (Zhou and Feng 2017), we firstly propose a multi-Level Deep Cascade Trees model (short as *ldcTree*) to cope with the essential task *CVR prediction* in the "also view" module. *ldcTree* is another alternative to *DNN* and encourages to do representation learning by a level-by-level cascade structure. Specifically, it takes a multi-dimensional representational feature vector from preceding level and outputs its processing results to the next level by employing Gradient Boosting Decision Trees (*GBDT*) models (Friedman 2001), where a new tree is created to model the residual of previous trees during each iteration, and a traversal from root node to a leaf node represents a combination rule of certain input features. One step further, we propose to utilize the cross-entropy value of each leaf node on the trees in the preceding level *GBDT* to construct the feature representation for the next level *GBDT*, which results a clear explanation of the splitting node in the next level *GBDT*, i.e., a traversal from root to a leaf node in next level *GBDT* indicates a combination rule of certain paths on the trees from preceding level *GBDT*. Then, inspired by the idea of ensemble learning, we proposed Ensemble *ldcTree* (short as *E-ldcTree*), which encourages the model's diversity and enhances the representation ability.

Furthermore, it is noteworthy that a small number of raw features contributes the majority of explanatory power while the remaining features have only a marginal contribution in *GBDT* models (He et al. 2014), which results in the importance of certain raw features can't be demonstrated. Therefore, we further proposed an improved Feature learning method based on the above *EldcTree*, named *F-EldcTree*, which takes more adequately use of weak and strong correlation features identified by pre-trained *GBDT* model at corresponding levels. The key contributions of this paper are:

- We propose a novel model *ldcTree* and its extension *EldcTree*, which are decision tree ensemble methods by exploiting the deep cascade structures and using a cross-entropy based feature representation. It exhibits a strong feature representation ability and has a clear explanation.
- To the extent of our knowledge, our proposed *F-EldcTree* is the first recommendation work which adequately takes

into account weak and strong correlation features identified by pre-trained *GBDT* model at corresponding levels, and contributes more excellent ability for representation learning.

- We have successfully deployed the proposed methods to the recommendation module in our E-commerce platform, and carry out online experiments with more than 100M users and items, furthermore achieves 12 percent *CVR* improvement compared with the baseline model.

The rest of this paper is organized as follows. Section 2 briefly reviews existing related work. Section 3 describes the proposed approach in detail, followed by presenting experimental results on both off-line evaluation and online applications in Section 4. We conclude the paper in Section 5.

Related work

Conversion Rate Prediction

Conversions are very rare events and only a very small portion of the users will eventually convert after clicking or being shown, resulting in extremely challenging for building thus models (Mahdian and Tomak 2007; Chapelle, Manavoglu, and Rosales 2015; Rosales, Cheng, and Manavoglu 2012; Oentaryo et al. 2014). In addition, it can be broadly divided to two categories of Post View Conversion (*PVC*) and Post Click Conversion (*PCC*), which means conversion after viewing an item without having clicked it itself and conversion after having clicked, respectively. In the context of this paper, conversion refers to the purchase event that occurs after a user clicking an item, i.e. post click conversions (*PCC*) (Rosales, Cheng, and Manavoglu 2012).

Tree based Feature Representation

GBDT follows the Gradient Boosting Machine (*GBM*) (Friedman 2001), which produces competitive, highly robust, interpretable procedures for both regression and classification, especially appropriate for mining less than clean data. In literature (He et al. 2014), a hybrid model, combining *GBDT* with Logistic Regression (*LR*), outperforms either of these methods on their own. It treats each individual tree as a bin feature and takes the index of the leaf node an instance ends up falling in as value. Therefore, it converts a real-valued vector into a compact binary-valued vector, further included into the next linear model, i.e. *LR*. Compared with (He et al. 2014), our proposed method employs the cross-entropy based feature representation in a deep cascaded structure, which results in strong and explainable representation ability, i.e., a traversal from root to a leaf node in next level *GBDT* indicates a combination rule of certain paths on the trees from preceding level *GBDT*.

gcForest, an alternative to deep neural networks for many tasks, is proposed in literature (Zhou and Feng 2017), which employs deep forest structure to do representation learning. Specifically, it takes a multi-dimensional class vector from preceding level, together with the original feature vector, as the inputs of next level. Our proposed methods mainly have two significant differences from *gcForest*:

- In *gcForest*, the class-specific prediction probabilities form a feature vector, further included into the next level forest after concatenating it with the original features. However, We employ *GBDT* as the base unit in the proposed deep cascade structure. In addition, we use the cross-entropy in each tree of the preceding *GBDT* as the feature representation for next level *GBDT*. The aforementioned two points lead to a more explainable feature representation ability of the proposed method, e.g., a traversal from root to a leaf node in next level *GBDT* indicates a combination rule of certain paths on the trees from preceding level *GBDT*.
- Compared with *gcForest*, Our method takes into account mutual complementarity between strong correlation features and weak correlation features for better representation learning.

The Proposed Approach

In this section, we firstly proposed a novel multi-Level Deep Cascade Trees (*ldcTree*) to tackle the *CVR* prediction problem in recommendation. Specifically, the base structure of the *ldcTree* is constructed by stacking several *GBDT*s sequentially, and the cross-entropy of each leaf node in the preceding *GBDT* is calculated and used to form the input feature representation for the next *GBDT*. Moreover, inspired by the idea of ensemble learning, an improved structure of the *ldcTree* is proposed, named Ensemble *ldcTree* (*EldcTree*), which further encourages the model's diversity and enhances the representation ability. Finally, we notice that a small number of features contributes the majority of explanatory power while the remaining features have only a marginal contribution in *GBDT* models (He et al. 2014), which leads to the importance of weak correlation features, especially the combination of weak correlation features, can't be revealed. Therefore, based on *EldcTree*, We propose an improved Feature learning method, named *F-EldcTree*. We will present the details in the following parts.

Representation Learning based on Cascade Trees

Inspired by representation learning in deep neural network which mostly relies on the level-by-level abstraction of features, we propose a novel method named *ldcTree* by employing the deep cascade tree structure. In *ldcTree*, level-by-level greedily learning towards the final target is carried out. Specifically, each level with a certain number of trees included in *GBDT*, receives the outputted features from its preceding level. Referring to Fig. 2, it is an illustration of *ldcTree*, where it contains two levels, with each level three trees and two trees respectively. To facilitate the narration, some mathematical notations are defined as follows:

- S_{ijk} : the cross-entropy of the k -th leaf node of the j -th tree at level i .
- I_{ijk} : the number of instances falling in the k -th leaf node of the j -th tree at level i .
- δ_{ijk} : the split threshold for the k -th node of the j -th tree at the level i .
- L_{ij} : the number of leaf nodes of the j -th tree at level i .

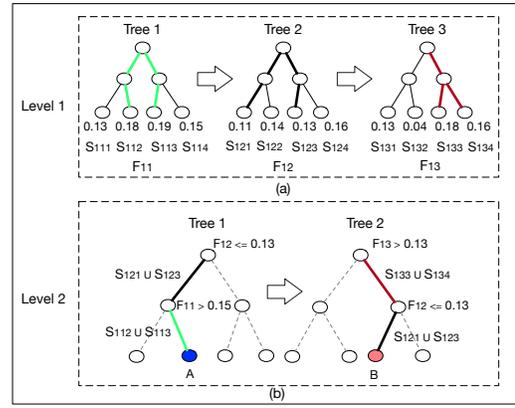


Figure 2: Illustration of feature representation in the *ldcTree*. (a) an exemplar *GBDT* of three trees at level 1. (b) an exemplar *GBDT* at the next level of two trees. Each splitting node at the level 2 corresponds a split of paths in a certain tree at level 1. Please note the corresponding colors of paths at the two levels.

- F_{ij} : the feature value at the j -th dimension of the feature at level i .
- N_i : the number of trees for the *GBDT* model at level i .
- $h_{ij}(x^n)$: the predicted probability of the n -th instance x^n on the j -th tree at the level i .
- y_n : the ground truth label of the n -th instance x^n : 0 or 1 in our two class *CVR* problem.

Given an instance, according to the principle of *GBDT* model, each individual tree will produce a path from the root node to a leaf node. Instances are split into different paths and each leaf node will gather a certain part of them. Then, we define the cross-entropy S_{ijk} at each leaf node as:

$$S_{ijk} = -\frac{1}{I_{ijk}} \sum_{m=1}^{I_{ijk}} [y_m \log(h_{ij}(x^m)) + (1 - y_m) \log(1 - h_{ij}(x^m))] \quad (1)$$

Therefore, there are L_{ij} cross-entropy values on the j -th tree on the i -th level, i.e., $S_{ij1}, S_{ij2}, \dots, S_{ijL_{ij}}$ and each of them is a possible instantiation of F_{ij} . For all the trees in a *GBDT* at the level i , we denote the feature representation as $[F_{i1}, F_{i2}, \dots, F_{iN_i}]$, and use it as the inputs of the *GBDT* at the level $i + 1$.

Without loss of generality, assuming the feature F_{ij} with the best *gini* value is chosen for splitting instances on the q -th node of the p -th tree at the level $i + 1$. The split threshold $\delta_{(i+1)pq}$ divides the set of $S_{ij1}, S_{ij2}, \dots, S_{ijL_{ij}}$, into two subsets S_{ij}^+, S_{ij}^- :

$$S_{ij}^+ = \bigcup_{p=1}^{L_{ij}} I(S_{ijp} > \delta_{(i+1)pq}) S_{ijp} \quad (2)$$

$$S_{ij}^- = \bigcup_{p=1}^{L_{ij}} I(S_{ijp} \leq \delta_{(i+1)pq}) S_{ijp} \quad (3)$$

where $I(z)$ is an indicator function that outputs 1 if z is true and zero otherwise.

Remark 1. By using cross-entropy as the basic feature representation for leaf nodes, the proposed *ldcTree* has a clear explanation: i.e., a traversal from root to leaf nodes in the next level *GBDT* corresponds to the combination of certain traversals in the preceding *GBDT*, which also leads to a distributed feature representation ability.

Explanation: Usually, $\delta_{(i+1)pq}$ in Eq.(2) and Eq.(3) takes a value from the set of $S_{ij1}, S_{ij2}, \dots, S_{ijL_{ij}}$. The splitting process is carried out repeatedly until the stop rule for leaf nodes holds (We show an example in Fig. 2). It can be seen that, each element in S_{ij}^+ or S_{ij}^- corresponds a path on the j -th tree of level i . Therefore, each splitting node at the level $i + 1$ corresponds a split of paths in a certain tree at the preceding level i . Consequently, a path at the level $i + 1$ corresponds a union of several paths at the preceding level i . It indicates the nature of the proposed *ldcTree*: a clear explanation and a distributed feature representation ability. As shown in Fig. 2, an instance is represented as a three-dimensional feature vector $[F_{11}, F_{12}, F_{13}]$ after level 1 with three trees, where F_{11} takes a value from the set of 0.13, 0.18, 0.19, 0.15, analogously for F_{12} and F_{13} . Given $[F_{11}, F_{12}, F_{13}]$ as the inputs at level 2, F_{12} is chosen as the split feature with the split threshold 0.13 on Tree 1, which leads to $S_{121}, S_{123} \subset S_{12}^-$ and $S_{122}, S_{124} \subset S_{12}^+$. And analogously, $S_{112}, S_{113} \subset S_{11}^+$ and $S_{111}, S_{114} \subset S_{11}^-$. Consequently, the union of paths at leaf node A can be represented as $S_{12}^- \cap S_{11}^+$, i.e. $(S_{121} \cup S_{123}) \cap (S_{112} \cup S_{113})$, and analogously for Tree 2.

Moreover, inspired by the ensemble learning idea, we propose an Ensemble *ldcTree* named *EldcTree* by constructing several parallel *ldcTrees* to enhance the representation ability further. As shown in Fig. 3, each horizontal part of the ensemble structure is a single *ldcTree*, and each vertical part of the ensemble structure are parallel *GBDT*s. Their initial input features for the first level of each *ldcTree* are different and chosen randomly from a raw feature pool which encourages the model’s *diversity*. The last level’s outputted features from each *ldcTree* are concatenated together and used as the input features of the last *GBDT* for final prediction. Here, the *diversity* in our model not only represents the nature of diverse abstracted high-level features, but also the value of inherent ensemble learning. For example, each individual *Horizontal* structure randomly choose a subset of features from the feature pool as input, which are then abstracted to different high-level features through the cascaded structure. It inherits the idea in feature engineering by learning high-level features from different combinations of low-level raw features. Then, the final *GBDT* model, concatenating all the high-level features from preceding *ldcTree* as inputs, indeed follows the ensemble learning idea for further enhancing the performance of the entire model. It is noteworthy that such an ensemble structure is naturally apt to parallel implementation and have the potential for incremental learning, e.g., the idea used in broad learning system (Chen and Liu 2018). We leave it as a future work.

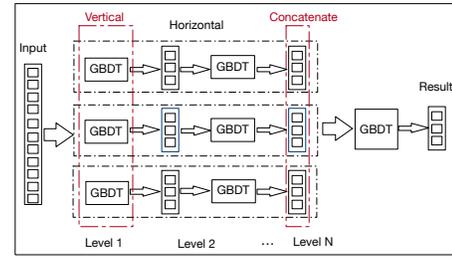


Figure 3: Illustration of the structure of *EldcTree*, where each horizontal part of the ensemble structure is a single *ldcTree*, and each vertical part of the ensemble structure are parallel *GBDT*s. The input features of the first level *GBDT*s are different and randomly chosen from a feature pool, which encourages the model’s diversity. The outputted features from each *ldcTree* are concatenated together and used as the input features of the last *GBDT* for final prediction.

Cascade Trees Associated With Weak Correlation Features

Though features can be chosen randomly from a given feature pool and used as the input features of each *ldcTree*, features indeed have different importance for prediction. In this paper, we use the *statistic Boosting Feature Importance* (He et al. 2014), which aims to capture the cumulative loss reduction attributable to a feature, to measure feature’s importance. More specifically, a best feature is selected and split to maximize the squared error reduction during each tree node construction. Therefore, the importance of each feature is determined by summing the total reduction for itself across all the trees. Typically, a small number of features contributes the majority of explanatory power while the remaining features have only a marginal contribution in *GBDT* models. Here, we regard the features contributing the majority of explanatory power as **SCF**, i.e. “Strong Correlation Features”, while features having only a marginal contribution as **WCF**, i.e. “Weak Correlation Features”.

To take adequate use of *WCF*, we proposed an improved Feature learning method based on the above *EldcTree*, named *F-EldcTree*, whose structure is showed in Fig. 4. In Fig. 4(a), a separated *GBDT* is pre-trained to identify the importance of all the initial raw features, and these features are further split into two subsets, namely *WCF* and *SCF*. Due to our practical lessons, single feature from *WCF* contributes little to final prediction results, while the combinations of these features from *WCF* not. Moreover, in *GBDT* models, a traversal from root to a leaf node in each tree indicates a combination rule of certain raw features. Therefore, we can resort to *GBDT* models for uniting certain features from *WCF* to take full advantage of them and collaborate with *SCFs* to further improve the prediction accuracy.

The strategies of randomly selecting features in *gcForest* (Zhou and Feng 2017) convinces us its effectiveness in ensemble learning. Therefore, in Fig. 4(b), features are also randomly chosen from *WCF*s during the first level *GBDT* training stage of each *ldcTree*, which not only can learn

the combination of certain *WCFs*, but also encourages the model’s diversity and enhances the representation ability further. For *GBDTs* at the remaining levels of each *ldcTree*, their input features consists of two parts: one is the representational features from the preceding level, the other is randomly chosen from the *SCFs*. In this way, the *F-EldcTree* starts learning features from *WCFs* and combines the learned features (i.e., the combinations of *WCFs*) with *SCFs* little by little. Lastly, an additional *GBDT* model concatenates all the representational features from *ldcTrees* as the inputs for final results.

Experimental results

To evaluate the effectiveness of the proposed method, we conducted extensive experiments including off-line and on-line evaluations. First, we present the evaluation settings including data set preparation, evaluation metrics, a brief description of related comparison methods, and the hyper-parameters settings of the proposed method. Then, we present the experimental results of different methods on the off-line data set along with the analysis. Finally, we present experimental results of different methods for online deployment through A/B test.

Evaluation settings

Data set preparation The off-line benchmark data set was constructed from the real click and purchase logs of our recommendation module in several consecutive days of December, 2017. And it consists of more than 100M instances, each of which contains user/item features and label (here, an individual id for each instance, while label is positive if purchase after clicking, or negative if no purchase after clicking). When preparing off-line benchmark data set, we divided the whole benchmark data into three disjoint parts according to the id, i.e., 40, 20 and 40 percent of the whole benchmark data for training data, validation data and test data respectively. Additionally, we extract hundreds of raw features including user features, item features and user-item cross features for each instance. For example, user features include users’ ages, genders and purchasing powers, etc. Item features include items’ prices, historical *CVRs* and Click-through Rate (*CTRs*), etc. User-item cross features include uses’ historical *CVRs*, preference scores, etc, on items.

Evaluation metrics In order to comprehensively compare the performance of the propose method with other related methods, we adopt two commonly used metrics namely *AUC* (Area Under Curve) and F_1 score based on precision and recall. Specifically, denoting all ground truth positive instances as T and all predicted positive instances as P , then precision and recall are defined as follows:

$$precision = \frac{|P \cap T|}{|P|} \quad (4)$$

$$recall = \frac{|T \cap P|}{|T|} \quad (5)$$

Then, the F_1 score is defined as:

$$F_1 = \frac{2 * precision * recall}{precision + recall} \quad (6)$$

Table 1: Hyper-parameters in *ldcTree* and *EldcTree*

Parameters Name	Value
the type of loss function	logistic loss
minimum instance numbers when node split	20
sampling rate of train set for each iteration	0.6
sampling rate of features for each iteration	0.6
the tree depth	8
the number of trees	150
learning rate	0.01

Related comparison methods In the following experiments, we compare the proposed method with other related methods including:

- *Naive GBDT*: We refer to a single *GBDT* model without level-by-level learning as *Naive GBDT*.
- *GBDT + LR*: First, feature representation is trained from a *GBDT* model. Then, it is used for *CVR* prediction by a Logistic Regression (LR). For feature representation, it calculated a bin feature for each individual tree by taking the index of the leaf node which an instance ends up falling in as the corresponding feature value. Therefore, it converts a real-valued raw feature vector into a compact binary-valued feature vector (He et al. 2014).
- *DNN*: Referring to (He et al. 2017), we design a *DNN* structure including three hidden layers and a prediction layer, where ReLU is used as the activation function for each hidden layer. We choose the hyper-parameters on the validation set. For instance, the number of units for each hidden layer is set as 128. Dropout rate is set as 0.5. We use the cross-entropy loss and *SGD* algorithm to train this *DNN* model.
- *gcForest*: Following our practical experience, we replace the Forests in the original *gcForest* (Zhou and Feng 2017) with *GBDTs*. Since the *CVR* prediction is a binary classification problem, a two-dimensional class-specific feature vector is obtained from each *GBDT*. Then, it is used along with the raw feature vector as the inputs of next level *GBDT* for learning deep feature representation further.

Moreover, we also evaluate the performance among our proposed *ldcTree*, *EldcTree* and *F-EldcTree* methods for demonstrating the effectiveness of ensemble learning and the adequate use of *WCF* and *SCF* at corresponding levels. It should also be noted that all methods use the same raw features as inputs if not specified (e.g., here, user, item and user-item cross features are included.)

Hyper-parameters settings We choose the hyper-parameters of the proposed methods according to the *AUC* metric on the validation set. And the main hyper-parameters of the proposed methods we used in all the following concern experiments are shown in Tab. 1. Here, we take a critical parameter “the tree depth” as example to illustrate the process of parameter selection in *ldcTree* model. After sampling a small subset from the whole data, we train

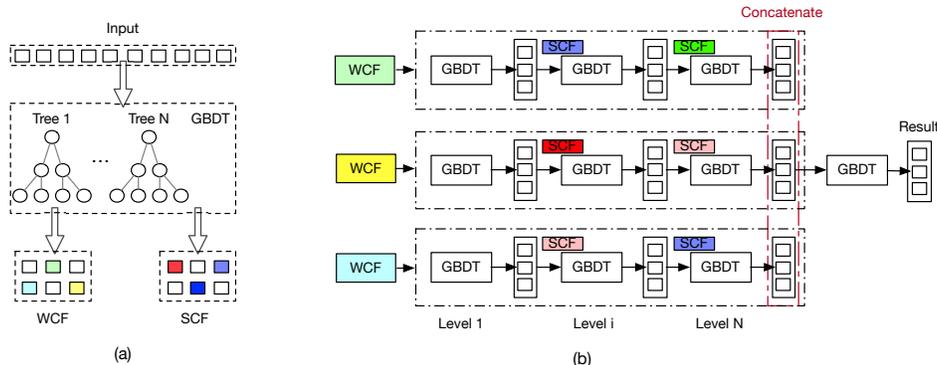


Figure 4: Illustration of the proposed *F-EldcTree* method, which utilizes weak correlation features and strong correlation features in a coordinated manner.

Table 2: The metric on Time Cost and the AUC in *ldcTree*

Metric Name	Depth 4	Depth 6	Depth 8	Depth 10
<i>Time Cost(s)</i>	5	7	15	26
<i>AUC</i>	0.783	0.789	0.794	0.797

different *ldcTree* models by changing the tree depth while fixing other settings. Results are shown in Tab. 2. It can be seen that the time cost for each iteration increases consistently with the growth of the tree depth, and the corresponding *AUC* on the validation set also increases. In addition, the time cost increases significantly when the depth increases from 8 to 10, while for the *AUC* value, it grows at a snail’s pace. For example, marginal gains of 0.003 are achieved by increasing the depth from 8 to 10. Therefore, to achieve a trade-off between model capacity and complexity, we set the hyper-parameter “the tree depth” as 8. Moreover, *EldcTree* and *F-EldcTree* can refer to the same hyper-parameters as *ldcTree*. Finally, aforementioned naive experiments also taught us that the performance of models can’t be boosted further while increasing the tree depth blindly.

Comparison results on off-line dataset

We report the *AUC* values and F_1 scores of different methods on the off-line test set. The results are shown in Tab. 3 and Tab. 4, respectively. Referring to Tab. 3, it can be seen that the *GBDT + LR* method achieves a gain of 0.0039 compared to *Naive GBDT* due to the additional *LR* classifier. *DNN* and *gcForest* achieve better results than *GBDT + LR*. It convinces us that strong representation features are learned due to their deep structures. Our proposed method *ldcTree* achieves higher *AUC* values than certain related methods (such as *Naive GBDT*, *GBDT+LR*, *DNN*). Moreover, the proposed *EldcTree* achieves better *AUC* result than *ldcTree* due to the idea of ensemble learning. Finally, the further proposed *F-EldcTree* achieves the best result than all the other competitive methods by taking full use of *WCF* and *SCF* little by little, together with the idea of ensemble learning. And the gain is nearly 0.063 compared to the initial

Table 3: Comparison *AUC* results for all competitors.

Method Name	<i>AUC</i>
<i>Naive GBDT</i>	0.7692
<i>GBDT + LR</i>	0.7731
<i>DNN</i>	0.7793
<i>gcForest</i>	0.7854
<i>ldcTree</i>	0.7942
<i>EldcTree</i>	0.8121
<i>F-EldcTree</i>	0.8315

baseline *Naive GBDT*. According to our practical lessons, it should be noted that a gain of 0.01 in off-line *AUC* can lead to big increment in revenue in our online recommendation system. In conclusion, the significant gain over initial *Naive GBDT* convinces the effectiveness of the proposed deep cascade structure for stronger feature representation, and the gain over *gcForest* convinces the effectiveness of level-by-level learning, for example, taking the outputs of preceding level as the inputs of the next level. Moreover, compared with *EldcTree*, results of *F-EldcTree* convince the idea by taking full use of weak and strong correlation features.

As for the F_1 score, we report several values by setting different thresholds. First, we sort all the instances in a descending order according to the predicted score. Then, we choose 3 thresholds namely top@10%, top@20% and top@50% to split the predictions into positive and negative groups accordingly. Finally, we calculate the Precision, Recall and F_1 scores of these predictions at different thresholds. Results are showed in Tab. 4. As can be seen, the proposed method achieves the best performance which is consistent with Tab. 3.

Online evaluation results

Next, we firstly present the effectiveness of level-by-level learning through online contrastive experiments. Then, we further demonstrate the effectiveness of *F-EldcTree* by taking full advantage of *WCF* and *SCF* compared with other competitors. It should also be noted that all comparison methods use the same input features if not specified (e.g.,

Table 4: The Precision, Recall and F_1 score for all the competitors.

Method Type	Method Name	top@10 percent			top@20 percent			top@50 percent		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Compare Methods	<i>Naive GBDT</i>	5.75%	36.39%	9.93%	4.16%	51.69%	7.70%	2.42%	76.57%	4.68%
	<i>GBDT + LR</i>	6.39%	37.40%	10.92%	4.63%	54.16%	8.52%	2.68%	78.71%	5.19%
	<i>DNN</i>	6.79%	38.32%	11.54%	4.92%	55.51%	9.03%	2.85%	80.96%	5.51%
	<i>gcForest</i>	6.82%	39.03%	11.61%	4.94%	56.52%	9.08%	2.87%	82.14%	5.54%
Our Methods	<i>ldcTree</i>	6.85%	39.86%	11.69%	4.96%	57.72%	9.13%	2.88%	82.92%	5.56%
	<i>EldcTree</i>	7.44%	41.49%	12.62%	5.39%	61.81%	9.91%	3.13%	87.31%	6.04%
	<i>F-EldcTree</i>	8.16%	43.91%	13.76%	5.91%	63.58%	10.81%	3.43%	92.40%	6.61%

Table 5: The Effectiveness of level-by-level learning.

Model Name	Day 1	Day 2	Day 3	Day 4
<i>Naive GBDT</i>	100%	100%	100%	100%
<i>ldcTree</i>	104.3%	104.1%	103.8%	103.9%
<i>EldcTree</i>	107.1%	107.4%	106.3%	106.8%

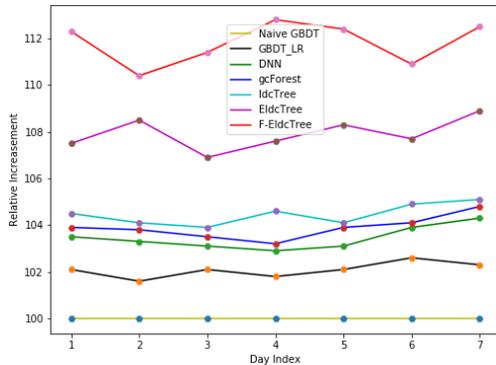


Figure 5: The online A/B test results on CVR. The initial baseline model marked in yellow is based on *Naive GBDT*. Here, all the comparison methods resort to the same features including user, item, and user-item cross features.

here, user, item and user-item cross features are included.). In addition, we fix all the other online recommendation modules unchanged except the CVR module.

The Effectiveness of level-by-level Learning For demonstrating the effectiveness of the proposed level-by-level learning, we implement *ldcTree* and *EldcTree* methods with the same features from *Naive GBDT*. After deploying them in the online recommendation system, we record four days’ purchase logs and calculate the relative increase in CVR. The A/B test results are showed in Tab. 5. It can be seen the proposed *ldcTree* method achieves more than 4% gain of CVR averagely, while *EldcTree* more than 7%. In addition, after analyzing the difference between the structures of the two methods, we find that the gain mainly comes from the stronger feature representation ability of the proposed deep cascade structure in *EldcTree*. It is consistent with the experimental results on the aforementioned off-line data set.

The Effectiveness of *F-EldcTree*. After demonstrating the effectiveness of utilizing level-by-level learning through

online A/B experiments, we employed *F-EldcTree* for online environment. In addition, the features for other competitive methods are exactly the same with *F-EldcTree*. The A/B test results are showed in Fig. 5, where *gcForest* and *DNN* achieve better results than *Naive GBDT* due to their deep feature representation abilities. As for the proposed method, it achieves the best result, i.e. 12 percent increment in CVR among all the methods.

In a nutshell, considering the experimental results from both off-line and online tests, we conclude that the proposed method has a stronger feature representation ability due to its deep cascade structure and the adequate use of *WCFs* and *SCFs* at corresponding levels. Moreover, these two distinct characteristics enables the learned features to have a clear explanation as depicted in Section 3.1.

Conclusions and future work

In this paper, we introduce effective and efficient distributed feature learning methods *ldcTree* and its extension *EldcTree*, which have a deep cascade structure by stacking several *GBDT* units sequentially. By using a cross-entropy based feature representation, it leads to a clear explanation and a distributed feature representation ability. Moreover, after taking into account mutual complementarity between strong correlation features and weak correlation features under the ensemble learning framework, the proposed method *F-EldcTree* achieves the best performance in both off-line and on-line experiments. Specifically, we successfully deploy the proposed method online in our E-commerce platform, it achieves a significant improvement compared with the previously baseline, i.e. 12 percent increment in CVR. Our methods have small training cost and are naturally apt to parallel implementation. In addition, it is promising to be applicable for other online advertising scenarios.

Future work may include the following two directions: 1) Incorporating more features, such as information from the parent nodes and sibling nodes for learning stronger feature representation. 2) Studying the end-to-end training method for jointly feature learning and classifying based on the proposed deep cascade tree structure.

Acknowledgment

This work was partly supported by the National Natural Science Foundation of China (NSFC) under Grants 61806062 and 61751304.

References

- Adomavicius, G., and Tuzhilin, A. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering* (6):734–749.
- Balabanović, M., and Shoham, Y. 1997. Fab: content-based, collaborative recommendation. *Communications of the ACM* 40(3):66–72.
- Chapelle, O.; Manavoglu, E.; and Rosales, R. 2015. Simple and scalable response prediction for display advertising. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5(4):61.
- Chen, C. P., and Liu, Z. 2018. Broad learning system: an effective and efficient incremental learning system without the need for deep architecture. *IEEE transactions on neural networks and learning systems* 29(1):10–24.
- Chen, J.; Zhang, H.; He, X.; Nie, L.; Liu, W.; and Chua, T.-S. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, 335–344. ACM.
- Davidson, J., and Liebald, B. 2010. The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, 293–296. ACM.
- Deng, L.; Seltzer, M. L.; Yu, D.; Acero, A.; Mohamed, A.-r.; and Hinton, G. 2010. Binary coding of speech spectrograms using a deep auto-encoder. In *Eleventh Annual Conference of the International Speech Communication Association*.
- Deng, L.; Li, J.; Huang, J.-T.; Yao, K.; Yu, D.; Seide, F.; Seltzer, M.; Zweig, G.; He, X.; Williams, J.; et al. 2013. Recent advances in deep learning for speech research at microsoft. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 8604–8608. IEEE.
- Elkahky, A. M.; Song, Y.; and He, X. 2015. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *Proceedings of the 24th International Conference on World Wide Web*, 278–288. International World Wide Web Conferences Steering Committee.
- Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* 1189–1232.
- Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.
- Gomez-Uribe, C. A., and Hunt, N. 2016. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)* 6(4):13.
- Guo, H.; Tang, R.; Ye, Y.; Li, Z.; and He, X. 2017. Deepfm: A factorization-machine based neural network for ctr prediction. *arXiv preprint arXiv:1703.04247*.
- He, X.; Pan, J.; Jin, O.; Xu, T.; Liu, B.; Xu, T.; Shi, Y.; Atallah, A.; Herbrich, R.; Bowers, S.; et al. 2014. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, 1–9. ACM.
- He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; and Chua, T.-S. 2017. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, 173–182. International World Wide Web Conferences Steering Committee.
- Huang, W.; Wu, Z.; Chen, L.; Mitra, P.; and Giles, C. L. 2015. A neural probabilistic model for context based citation recommendation. In *AAAI*, 2404–2410.
- Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.
- Liu, W.; Wang, Z.; Liu, X.; Zeng, N.; Liu, Y.; and Alsaadi, F. E. 2017. A survey of deep neural network architectures and their applications. *Neurocomputing* 234:11–26.
- Mahdian, M., and Tomak, K. 2007. Pay-per-action model for online advertising. In *Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising*, 1–6. ACM.
- Oentaryo, R. J.; Lim, E.-P.; Low, J.-W.; Lo, D.; and Finegold, M. 2014. Predicting response in mobile advertising with hierarchical importance-aware factorization machine. In *Proceedings of the 7th ACM international conference on Web search and data mining*, 123–132. ACM.
- Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.
- Rosales, R.; Cheng, H.; and Manavoglu, E. 2012. Post-click conversion modeling and analysis for non-guaranteed delivery display advertising. In *Proceedings of the fifth ACM international conference on Web search and data mining*, 293–302. ACM.
- Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Wang, H.; Wang, N.; and Yeung, D.-Y. 2015. Collaborative deep learning for recommender systems. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1235–1244. ACM.
- Yang, C.; Bai, L.; Zhang, C.; Yuan, Q.; and Han, J. 2017. Bridging collaborative filtering and semi-supervised learning: a neural approach for poi recommendation. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1245–1254. ACM.
- Zhang, S.; Yao, L.; and Sun, A. 2017. Deep learning based recommender system: A survey and new perspectives. *arXiv preprint arXiv:1707.07435*.
- Zhou, Z.-H., and Feng, J. 2017. Deep forest: towards an alternative to deep neural networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 3553–3559. AAAI Press.