



(51) International Patent Classification:

G06Q 30/02 (2012.01) G06N 20/00 (2019.01)
G06F 16/2457 (2019.01)

(21) International Application Number:

PCT/US2020/017861

(22) International Filing Date:

12 February 2020 (12.02.2020)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/807,258 19 February 2019 (19.02.2019) US
16/787,657 11 February 2020 (11.02.2020) US

(71) Applicant: **NEC LABORATORIES AMERICA, INC.**
[US/US]; 4 Independence Way, Suite 200, Princeton, NJ
08540 (US).

(72) Inventors: **CHENG, Wei**; 11 Blackhawk Ct., West Wind-
sor, NJ 08550 (US). **CHEN, Haifeng**; 5 Arnold Dr., Prince-
ton Junction, NJ 08550 (US).

(74) Agent: **BITETTO, James, J.**; TUTUNJIAN & BITET-
TO, P.C., 401 Broadhollow Road, Suite 402, Melville, NY
11747 (US).

(81) Designated States (unless otherwise indicated, for every
kind of national protection available): AE, AG, AL, AM,
AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ,
CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO,
DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN,
HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP,
KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME,
MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ,
OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA,
SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR,
TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every
kind of regional protection available): ARIPO (BW, GH,
GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ,
UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ,
TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK,
EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV,
MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,
TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW,
KM, ML, MR, NE, SN, TD, TG).

(54) Title: INTERPRETABLE CLICK-THROUGH RATE PREDICTION THROUGH HIERARCHICAL ATTENTION

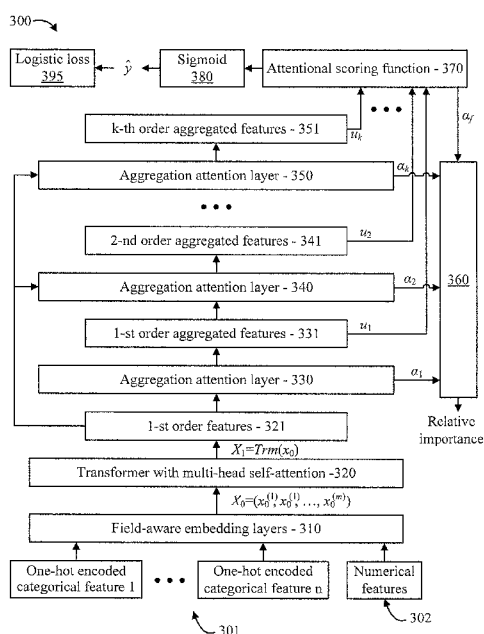


FIG. 3

(57) Abstract: A system is provided for interpretable viewing interest. A transformer with multi-head self-attention derives different hierarchical orders of input features. A plurality of hierarchical attention layers aggregate the different hierarchical orders to obtain aggregated single-order feature representations and derive aggregation attention weights for the different hierarchical orders based on an applied order of the plurality of hierarchical attention layers. A hierarchical interpretation layer determines a respective importance of each of the input features in various input feature combinations from which various CTR predictions are derived based on the aggregation attention weights and the significance of each order. A display device configured to display each of the various input feature combinations for the various CTR predictions along with the respective importance of each of the constituent one of the input features in the various input feature combinations.

Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

INTERPRETABLE CLICK-THROUGH RATE PREDICTION THROUGH HIERARCHICAL ATTENTION

RELATED APPLICATION INFORMATION

[0001] This application claims priority to U.S. Patent Application No. 16/787,657 filed on February 11, 2020, which claims priority to U.S. Provisional Patent Application No. 62/807,258, filed on February 19, 2019, the contents of which are incorporated herein by reference herein its entirety.

BACKGROUND

Technical Field

[0002] The present invention relates to recommender systems and more particularly to interpretable click-through rate prediction through hierarchical attention.

Description of the Related Art

[0003] Click-Through Rate (CTR) refers to accurately estimating the probability of a user clicking on a particular recommender item or advertisement. CTR plays a significant role in recommender systems such as online advertisements since it directly affects the revenue of advertising agencies. In order to attain a high CTR prediction accuracy, the complexity of proposed models has been ever increasing. Along with the growth of CTR prediction model complexity, the limitation of interpretability is gaining seriousness. The weights and activations of neurons in those models are deemed non-explainable so that reasonable interpretations of their decision making processes are hard to find. The lack of persuasive rationales to explain the outcomes of the models casts shadow on the reliability and security of the predictions. In online marketing and advertising scenarios, unexplainable recommendations provided by deep recommender systems may not have severe effect. However,

untrustworthy recommendations can lead to serious consequences such as the loss of health and wealth in scenarios of medication recommendation and financial services or products recommendation. Since CTR prediction is a critical component of recommender systems, interpretability of CTR prediction is also a necessity in aforementioned cases. Thus, there is a need for interpretable CTR prediction that indicates why a particular recommendation was made.

SUMMARY

[0004] According to an aspect of the present invention, a system is provided for interpretable viewing interest. The system includes a transformer with multi-head self-attention configured to derive different hierarchical orders of input features by learning various semantic interactions of the input features in different semantic subspaces. The system further includes a plurality of hierarchical attention layers configured to aggregate the different hierarchical orders of the input features to obtain aggregated single-order feature representations starting from a first order and continuing to a k-th order and derive aggregation attention weights for the different hierarchical orders of the input features based on an applied order of the plurality of hierarchical attention layers. The system also includes an attentional scoring layer configured to evaluate the aggregated single-order feature representations to output a significance of each of the orders with respect to various Click-Through Rate (CTR) predictions. The system additionally includes a hierarchical interpretation layer configured to determine a respective importance of each of the input features in various input feature combinations from which the various CTR predictions are derived based on the aggregation attention weights and the significance of each of the orders with respect to the CTR predictions. The system further includes a display

device configured to display each of the various input feature combinations for the various CTR predictions along with the respective importance of each of the constituent one of the input features in the various input feature combinations.

[0005] According to another aspect of the present invention, a method is provided for interpretable click viewing interest. The method includes deriving, by a transformer with multi-head self-attention, different hierarchical orders of input features by learning various semantic interactions of the input features in different semantic subspaces. The method further includes aggregating, by a plurality of hierarchical attention layers, the different hierarchical orders of the input features to obtain aggregated single-order feature representations starting from a first order and continuing to a k-th order. The method also includes deriving, by the plurality of hierarchical attention layers, aggregation attention weights for the different hierarchical orders of the input features based on an applied order of the plurality of hierarchical attention layers. The method additionally includes evaluating, by an attentional scoring layer, the aggregated single-order feature representations to output a significance of each of the orders with respect to various Click-Through Rate (CTR) predictions. The method also includes determining, by a hierarchical interpretation layer, a respective importance of each of the input features in various input feature combinations from which the various CTR predictions are derived based on the aggregation attention weights and the significance of each of the orders with respect to the CTR predictions. The method additionally includes displaying, by a display device, each of the various input feature combinations along with the respective importance of each of the constituent one of the input features in the various input feature combinations.

[0006] These and other features and advantages will become apparent from the following detailed description of illustrative embodiments thereof, which is to be read in connection with the accompanying drawings.

BRIEF DESCRIPTION OF DRAWINGS

[0007] The disclosure will provide details in the following description of preferred embodiments with reference to the following figures wherein:

[0008] FIG. 1 is a block diagram showing an exemplary processing system, in accordance with an embodiment of the present invention;

[0009] FIG. 2 is a diagram showing an exemplary output of an interpretable recommendation system, in accordance with an embodiment of the present invention;

[0010] FIG. 3 is a block diagram showing an exemplary machine learning pipeline, in accordance with an embodiment of the present invention; and

[0011] FIGs. 4-5 are flow diagrams showing an exemplary method for interpretable click-through prediction through hierarchical attention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

[0012] Embodiments of the present invention are directed to interpretable Click-Through Rate (CTR) prediction through hierarchical attention.

[0013] In an embodiment, the present invention provides an Interpretable CTR prediction with Hierarchical Attention (InterHAt) that can provide reliable explanations together with accurate recommendations. The present invention quantifies the effects of feature combinations of arbitrary orders by a hierarchical attention mechanism and explains the recommending decision according to learned feature salience.

[0014] However, one impediment that degenerates the performance of important cross-feature detection is that factors may have conflicting influences on the CTR. For example, a record of $\{f_1:\text{movie.type}=\text{horror}, f_2:\text{user.type}=\text{young}, f_3:\text{event.time}=\text{8am}\}$ has conflicting factors: $\{f_1; f_2\}$ encourages the click-through whereas $\{f_2; f_3\}$ inhibits it since movie watching usually happens at night. Therefore, the present invention is equipped with a Transformer with multi-head self-attention to comprehensively study different possible feature-wise interactions since multiple attention heads can capture manifold semantic information from different latent semantic sub-spaces. Self-attention, sometimes called intra-attention, is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence.

[0015] In one or more embodiments, the present invention uses hierarchical attention to explain the CTR prediction by pinpointing the significant single features or multiple feature combinations that have great contributions to making the click-through. The present invention is able to efficiently handle high-order feature combinations within a linear complexity. As used herein, the term “high-order” refers to feature orders above the 2nd order. The term “low-order” refers to feature orders below the 3rd order.

[0016] In one or more embodiments, the present invention utilizes a Transformer with multi-head self-attention to thoroughly analysis possible interactive relations between features in different latent semantic subspaces.

[0017] FIG. 1 is a block diagram showing an exemplary processing system 100, in accordance with an embodiment of the present invention. The processing system 100 includes a set of processing units (e.g., CPUs) 101, a set of GPUs 102, a set of memory devices 103, a set of communication devices 104, and set of peripherals 105.

The CPUs 101 can be single or multi-core CPUs. The GPUs 102 can be single or multi-core GPUs. The one or more memory devices 103 can include caches, RAMs, ROMs, and other memories (flash, optical, magnetic, etc.). The communication devices 104 can include wireless and/or wired communication devices (e.g., network (e.g., WIFI, etc.) adapters, etc.). The peripherals 105 can include a display device, a user input device, a printer, an imaging device, and so forth. Elements of processing system 100 are connected by one or more buses or networks (collectively denoted by the figure reference numeral 110).

[0018] In an embodiment, memory devices 103 can store specially programmed software modules to transform the computer processing system into a special purpose computer configured to implement various aspects of the present invention. In an embodiment, special purpose hardware (e.g., Application Specific Integrated Circuits, Field Programmable Gate Arrays (FPGAs), and so forth) can be used to implement various aspects of the present invention.

[0019] In an embodiment, memory devices 103 can include a machine learning pipeline 103A. In an embodiment, machine learning pipeline 103A can be implemented by the machine learning pipeline 300 of FIG. 3 described herein below.

[0020] Of course, the processing system 100 may also include other elements (not shown), as readily contemplated by one of skill in the art, as well as omit certain elements. For example, various other input devices and/or output devices can be included in processing system 100, depending upon the particular implementation of the same, as readily understood by one of ordinary skill in the art. For example, various types of wireless and/or wired input and/or output devices can be used. Moreover, additional processors, controllers, memories, and so forth, in various configurations can also be utilized. These and other variations of the processing

system 100 are readily contemplated by one of ordinary skill in the art given the teachings of the present invention provided herein.

[0021] Moreover, it is to be appreciated that various figures as described below with respect to various elements and steps relating to the present invention that may be implemented, in whole or in part, by one or more of the elements of system 100.

[0022] As employed herein, the term “hardware processor subsystem” or “hardware processor” can refer to a processor, memory, software or combinations thereof that cooperate to perform one or more specific tasks. In useful embodiments, the hardware processor subsystem can include one or more data processing elements (e.g., logic circuits, processing circuits, instruction execution devices, etc.). The one or more data processing elements can be included in a central processing unit, a graphics processing unit, and/or a separate processor- or computing element-based controller (e.g., logic gates, etc.). The hardware processor subsystem can include one or more on-board memories (e.g., caches, dedicated memory arrays, read only memory, etc.). In some embodiments, the hardware processor subsystem can include one or more memories that can be on or off board or that can be dedicated for use by the hardware processor subsystem (e.g., ROM, RAM, basic input/output system (BIOS), etc.).

[0023] In some embodiments, the hardware processor subsystem can include and execute one or more software elements. The one or more software elements can include an operating system and/or one or more applications and/or specific code to achieve a specified result.

[0024] In other embodiments, the hardware processor subsystem can include dedicated, specialized circuitry that performs one or more electronic processing

functions to achieve a specified result. Such circuitry can include one or more application-specific integrated circuits (ASICs), FPGAs, and/or PLAs.

[0025] These and other variations of a hardware processor subsystem are also contemplated in accordance with embodiments of the present invention.

[0026] FIG. 2 is a diagram showing an exemplary output 200 of an interpretable recommendation system, in accordance with an embodiment of the present invention.

[0027] The objective in providing the output is to explicitly explain factors (1-order features or high-order feature combinations) that contribute to the recommendation. As shown, the present invention may report the first order feature (“wes Anderson”) and 2-order feature (“witty” and “action”) for the film that play important parts in the recommendation.

[0028] FIG. 3 is a block diagram showing an exemplary machine learning pipeline 300, in accordance with an embodiment of the present invention.

[0029] Basically, the pipeline 300 makes use of multi-head self-attention on embeddings of different 1-order fields. The, the pipeline 300 uses multi-layer high-order attention aggregation to capture cross-features and the final feature vectors are the concatenation of different orders of features.

[0030] The pipeline 300 includes one-hot encoded categorical features 1 through n, collectively denoted by the figure reference numeral 301, and further includes numerical features 302. The numerical data 302 can be in buckets. These features 301 and 302 represent click-through data. Each record of the click-through data includes userID, features of the user, itemID, feature of the item, and if the user clicked the item or the score of the rating the user gave to the item.

[0031] The features 301 and 302 are input to field-aware embedding layers 310, whose output is fed to a transformer 320 with multi-head self-attention. The

transformer 320 outputs 1-st order features 321 to a first aggregation attention layer 330 that outputs 1-st order aggregated features 331. The 1-st order aggregated features 331 are input to a second aggregation attention layer 340 that outputs 2-nd order aggregated features 341. The 2-nd order aggregated features 341 are input to further aggregation layers (collectively designated by the figure reference numeral 345) up to a k-th aggregation layer 350 that outputs k-th order aggregation features 351.

[0032] Each of the aggregation attention layers 330, 340, and 345 is operatively connected to a hierarchical interpretation layer 360 which uses attention weights to interpret the relative importance of different features (or cross-features).

[0033] All of the aggregated attention features are input to an attentional scoring function 370 whose output is fed to a sigmoid function 380. The output of the sigmoid function 380 is a prediction $\hat{y}$. Also output from pipeline 300 is a logistic loss 395 that denotes how far away between the predicted values (1 means clicked, 0 means non-clicked) and the ground-truth values in the training set.

[0034] Feature embedding (mapping to a vector a real numbers) is a prerequisite for CTR prediction since the raw features are mostly categorical and not directly applicable to the numeric computations of deep learning models. For categorical features, we apply multi-field one-hot encoding to field-aware embedding layers for low-dimensional continuous-valued feature representations. A hierarchical attention layer includes two parts as follows: (i) a transformer 320 with multi-head self-attention; and (ii) a series of hierarchical attention layers 330, 340, 345, 350.

Transformer 320 is powerful in capturing the rich pair-wise feature correlations by a multi-head self-attention module regardless of the orders of fields. The multiple attention heads learn the various feature interactions in different subspaces which

represents manifold factors that affect the click-through. The second-order of attention is multi-head self-attention and the higher-order attention is the cross-layer attention.

[0035] Further regarding the field-aware embedding layers 310, the same map different features (fields) to low-dimensional embeddings (i.e., embedding having a dimension number of less than three). For fields with non-value multi-categories, using one-hot embedding first, then map to a low-dimensional space.

[0036] Further regarding the transformer with multi-head self-attention, for categorical features, the pipeline 300 applies multi-field one-hot encoding to field-aware embedding layers for low-dimensional continuous-valued feature representations. Given its the value of field f , x_f , and the embedding layer of f , $emb_f(.)$, the embedding of field f , $v_{f,cat}$ in $\mathbb{R}^n$, is derived by $v_{f,cat} = emb_f(one - hot(x_f))$. The multiple attention heads learn the various feature interactions in different (semantic) subspaces which represents manifold factors that affect the click-through. Self-attention is computed not once but multiple times in the transformer's architecture, in parallel and independently.

[0037] Further regarding the sigmoid function, the same is applied to make the output prediction value to be between 0 and 1.

[0038] The elements of the machine learning pipeline 300 are described in further detail hereinbelow.

[0039] FIGs. 4-5 are flow diagrams showing an exemplary method 400 for interpretable click-through prediction through hierarchical attention.

[0040] At block 405, receive, by field-aware embedding layers, numerical features and one-hot encoded categorical features, by the field-aware embedding layers.

[0041] At block 410, map, by the field-aware embedding layers, the features to low-dimensional embeddings (i.e., embeddings having 2 or less dimensions) to output single order features as 1-st order features through k-th order features.

[0042] At block 415, derive, by a transformer with multi-head self-attention, different hierarchical orders of input features by learning various semantic interactions of the input features in different semantic subspaces.

[0043] At block 415, aggregate, by each of multiple hierarchical aggregation attention layers, a respective one of the different hierarchical orders of the input features to obtain aggregated single-order feature representations starting from a first order and continuing to a k-th order. Thus, for example, aggregate the 1-st order features by a first aggregation attention layer to obtain aggregated 1-st order features, aggregate the 2-nd order features by a second aggregation attention layer to obtain aggregated 2-nd order features, and so on up to aggregating the k-th order features by a k-th aggregation layer to obtain aggregated k-th order features.

[0044] At block 420, derive by each of multiple hierarchical aggregation attention layers, aggregation attention weights for the different hierarchical orders of the input features based on an applied order of the multiple hierarchical attention layers.

[0045] At block 425, evaluate, by an attentional scoring layer, the aggregated single-order feature representations to output a significance of each of the orders with respect to various Click-Through Rate (CTR) predictions

[0046] At block 430, determine, by a hierarchical interpretation layer, a respective importance of each of the input features in various input feature combinations from which the various CTR predictions are derived based on the aggregation attention weights and the significance of each of the orders with respect to the CTR predictions.

[0047] At block 435, display, by a display device, each of the various input feature combinations along with the respective importance of each of the constituent one of the input features in the various input feature combinations. FIG. 2 shows one example screen shot from a display device that can correspond to block 435.

[0048] At block 440, control the arrangement (order of presentation) of web pages to a user such that the web pages go from highest likelihood to obtain a click through to lowest likelihood to obtain the click through. In this way, the change for a click through is maximized. Likelihoods can be derived from the output of the sigmoid function for each of the CTR predictions.

[0049] A further description will now be given regarding the embedding layers, in accordance with an embodiment of the present invention.

[0050] Feature embedding is a prerequisite for CTR prediction since the click-through records include discrete categorical terms that are not directly applicable to numerical computations. A click-through record includes a set of fields F and a binary label y as the ground truth representing whether a click-through is actually made. Each field $f \in F$ has either a categorical or a numerical value. Distinct values are defined as different features. For categorical fields, we apply multi-field one-hot encoding to field-aware embedding layers for low-dimensional real-valued feature representations. Specifically, each distinct feature value v of a field is assigned a trainable d -dimensional continuous vector as its representation. If a particular feature appears in a click-through as the field representation. For numerical fields, we assign one vector to each field as its embedding. Given v_f as the normalized value of a numerical field f and $x_{num,0}^{(f)} \in \mathcal{R}^d$ as the trainable representation associated with this field, the representation of the feature, $x_{num}^{(f)} \in \mathcal{R}^d$, is derived by $x_{num}^{(f)} = v_f \cdot x_{num,0}^{(f)}$. The initial input representation matrix is $X_0 = (x_0^{(1)}, x_0^{(2)}, \dots, x_0^{(m)})$ where $m = |F|$.

[0051] A further description will now be given regarding the multi-head transformer, in accordance with an embodiment of the present invention.

[0052] The transformer is prevalent in NLP thanks to the outstanding power to learn the co-effects to the text semantics of word pairs within a sentence or across sentences regardless of the orders and distances of the words. In the context of CTR prediction, we define the coeffects of the features, i.e., feature interactions, towards different polarity as the “polysemy”. Therefore, we equip the present invention with a multi-head self-attention based Transformer to capture the rich pair-wise feature interactions and learn the diversified polysemy of feature interactions in different semantic subspaces, i.e., diversified implications towards the CTR in different click-through contexts.

[0053] Given the input matrix X_0 that includes the learnable embeddings of features of a training CTR record, the latent representation H_i of Transformer head i is obtained by a scaled dot-product attention, $H_i = \text{softmax}_i \left(\frac{QK^T}{\sqrt{d_K}} \right) V$, $Q = QW_i^{(Q)}X_0$, $K = W_i^{(K)}X_0$, $V = W_i^{(V)}X_0$.

[0054] Matrices $W_i^{(Q)} \in \mathbb{R}^{d_K \times d}$, $W_i^{(K)} \in \mathbb{R}^{d_K \times d}$, and $W_i^{(V)} \in \mathbb{R}^{d_K \times d}$ are weight parameters to learn for head i and d_K denotes the dimension of K and $H_i \in \mathbb{R}^{d_K \times m}$.

[0055] A combination of hidden features H_i forms an augmented representation matrix X_1 that preserves both the intrinsic and polysemic information of each feature. Computationally, we use concatenation followed by a feed-forward layer and a ReLU for the combination to learn the non-linearity of the combined information as $X_1 = \text{ReLU}(\text{FeedForward}(W_m[H_1; H_2; \dots; H_h]))$, where $W_m \in \mathbb{R}^{d \times h d_K}$ includes the weights and h is the number of attention heads and “;” denotes the concatenation of

matrices. The $X_1 \in \mathcal{R}^{d \times m}$ is the matrix with polysemy-augmented features and ready to be sent to the hierarchical attention layer for explainable CTR prediction.

[0056] A description will now be given regarding hierarchical attention, in accordance with an embodiment of the present invention.

[0057] The augmented feature matrix X_1 is served as the input of the hierarchical attention layers which learn the feature interaction and generate interpretations simultaneously. However, computing the high-order multi-feature interactions by enumerating all possible combinations is expensive due to the combinatorial explosion. Such potential expense motivates the aggregation of the current order before proceeding to the computation of the higher order. That is, in order to generate the $(i+1)$ -th order cross-features X_{i+1} , we first aggregate the i -th layer hidden features to u_i as a summarization of X_i . The interaction between X_i and X_1 , from which we derive X_{i+1} , is computed by the proxy of X_i , i.e., the attentional aggregation u_i from Equation (1), and X_1 . Mathematically, given the i -th feature matrix $X_i = (x_i^{(1)}, \dots, x_i^{(m)})$, its attentional aggregation representation u_i is as follows:

$$u_i = \text{AttentionalAgg}(X_i) = \sum_{j=1}^m \alpha_i^{(f)} x_i^{(f)} \quad (1)$$

where $\alpha_i^{(j)} \in \mathcal{R}$ denotes the attention on the j -th field in the i -th attentional aggregation layer. $\alpha_i^{(j)}$ is computed by the following:

$$\alpha_i^{(j)} = \frac{\exp(c_i^T \text{ReLU}(W_i x_i^{(j)}))}{\sum_{f \in f} \exp(c_i^T \text{ReLU}(W_i x_i^{(f)}))} \quad (2)$$

where $W_i \in \mathcal{R}^{s \times d}$ is the weight of layer i , $c_i \in \mathcal{R}^s$ is the context vector of layer i , and s denotes the attention space size. Note that other attention mechanisms can also be adopted here, such as the gated attention mechanism. Using u_i and X_i , we derive $x(j)_{i+1}$ in X_{i+1} by a cross-product transformation as follows:

$$x_{i+1}^{(j)} = u_i \circ x_1^{(j)} + x_i^{(j)}, j \in \{1, \dots, m\} \quad (3)$$

where $\circ$ denotes the Hadamard product of two vectors.

[0058] Recurrently applying Equation (1) and Equation (3) produces u_i and X_i for feature orders from the 1st order to the k -th, the highest cross-feature order to analyze, by a series of attentional aggregation layers. These layers composite a hierarchy that extracts features from low order to higher ones and the lower ones contribute to the construction of one-order higher features using the proposed attentional aggregation and cross-product transformation.

[0059] As the last step, we combine attentional aggregations $U = (u_1, u_2, \dots, u_k)$ to predict the probability of click-through. U gathers all combinatorial feature semantics of k orders. By modifying k , the present invention is able to capture arbitrary order of feature interactions, and yet avoids the exponential cardinality of high-order feature combinations.

[0060] A further description will now be given regarding the objective function and optimization, in accordance with an embodiment of the present invention.

[0061] The final CTR prediction function $g(U) = \hat{y} \in [0,1]$ maps U to a probability that quantifies the CTR. $g(U)$ is implemented as the following. It first computes the attentional aggregation of U by Equation (4) and Equation (5) to obtain its aggregation $u_f \in \mathcal{R}^d$ and attention $\alpha_f \in \mathcal{R}^k$ as follows:

$$u_f = \text{Attentional Agg}(U) = \sum_{j=1}^k \alpha_f^{(j)} u_j \quad (4)$$

$$\alpha_f^{(j)} = \frac{\exp(c_f^T \text{ReLU}(W_f u_j))}{\sum_{f \in \{1, \dots, k\}} \exp(c_f^T \text{ReLU}(W_f u_f))} \quad (5)$$

where α_f is the importance distribution across k feature orders, c_f and W_f are learnable parameters. Finally, the prediction $\hat{y}$ is then made by $\hat{y} = \text{sigmoid}(\text{MLP}(u_f))$ where $\text{MLP}(\cdot)$ refers to a shallow Multi-layer Perceptron that reduces the output dimension

from d to 1. The objective function, Equation (6), of the present invention is a cross entropy loss of binary classification as follows:

$$\mathcal{L}(\Theta) = \sum_{t \in D} [-y_t \log(\hat{y}_t) - (1 - y_t) \log(1 - \hat{y}_t)] + \lambda \|\Theta\|_2 \quad (6)$$

where D denotes the training set and Θ includes all trainable parameters, namely feature embeddings and the parameters of Transformer and hierarchical layers. An L2 regularization weighted by λ is applied to Θ to prevent overfitting. In an embodiment, we optimize Equation (6) by the Adam or other gradient descent optimizer.

[0062] A further description will now be given regarding interpretation, in accordance with an embodiment of the present invention.

[0063] The description elaborates on how to “understand” the attentions in the hierarchy as important factors that trigger the prediction of CTR. Note that the attention mechanism only highlights the salience of features so it is not expected to generate completely human readable interpretations.

[0064] Here is a walk-through of the interpretation using the salience distribution $(\alpha_1, \alpha_2, \dots, \alpha_k)$ and $\alpha_f \cdot \alpha_f$ includes the significance of all k orders of features and signifies the feature orders that are influential to the ultimate CTR prediction.

Dominant weights in $\alpha_f \in \mathcal{R}^k$ pinpoint the X_i 's that include significant i -th order features. According to α_f , we learn the numbers of orders, i.e., the numbers of interacting features, that have the strongest impact to encourage the user to click through the recommended ads. The attention weights in corresponding α_i identify the candidate individual features that participate in the contributory i -th order features. For example, if the attention weights of features of fields f_1 and f_2 , i.e., $\alpha_i[f_1]$ and $\alpha_i[f_2]$, outweigh the rest of the features in α_i , we learn that features of field f_1 and f_2 both contribute to an i -th order feature since they actively interact with the $i - 1$ order aggregation features. Finally, following the above steps, we can identify all features in

different orders. The actual click-through is interpreted by identifying salient features layer by layer and order by order.

[0065] Embodiments described herein may be entirely hardware, entirely software or including both hardware and software elements. In a preferred embodiment, the present invention is implemented in software, which includes but is not limited to firmware, resident software, microcode, etc.

[0066] Embodiments may include a computer program product accessible from a computer-usable or computer-readable medium providing program code for use by or in connection with a computer or any instruction execution system. A computer-usable or computer readable medium may include any apparatus that stores, communicates, propagates, or transports the program for use by or in connection with the instruction execution system, apparatus, or device. The medium can be magnetic, optical, electronic, electromagnetic, infrared, or semiconductor system (or apparatus or device) or a propagation medium. The medium may include a computer-readable storage medium such as a semiconductor or solid state memory, magnetic tape, a removable computer diskette, a random access memory (RAM), a read-only memory (ROM), a rigid magnetic disk and an optical disk, etc.

[0067] Each computer program may be tangibly stored in a machine-readable storage media or device (e.g., program memory or magnetic disk) readable by a general or special purpose programmable computer, for configuring and controlling operation of a computer when the storage media or device is read by the computer to perform the procedures described herein. The inventive system may also be considered to be embodied in a computer-readable storage medium, configured with a computer program, where the storage medium so configured causes a computer to operate in a specific and predefined manner to perform the functions described herein.

[0068] A data processing system suitable for storing and/or executing program code may include at least one processor coupled directly or indirectly to memory elements through a system bus. The memory elements can include local memory employed during actual execution of the program code, bulk storage, and cache memories which provide temporary storage of at least some program code to reduce the number of times code is retrieved from bulk storage during execution.

Input/output or I/O devices (including but not limited to keyboards, displays, pointing devices, etc.) may be coupled to the system either directly or through intervening I/O controllers.

[0069] Network adapters may also be coupled to the system to enable the data processing system to become coupled to other data processing systems or remote printers or storage devices through intervening private or public networks. Modems, cable modem and Ethernet cards are just a few of the currently available types of network adapters.

[0070] Reference in the specification to “one embodiment” or “an embodiment” of the present invention, as well as other variations thereof, means that a particular feature, structure, characteristic, and so forth described in connection with the embodiment is included in at least one embodiment of the present invention. Thus, the appearances of the phrase “in one embodiment” or “in an embodiment”, as well as any other variations, appearing in various places throughout the specification are not necessarily all referring to the same embodiment. However, it is to be appreciated that features of one or more embodiments can be combined given the teachings of the present invention provided herein.

[0071] It is to be appreciated that the use of any of the following “/”, “and/or”, and “at least one of”, for example, in the cases of “A/B”, “A and/or B” and “at least one of

A and B”, is intended to encompass the selection of the first listed option (A) only, or the selection of the second listed option (B) only, or the selection of both options (A and B). As a further example, in the cases of “A, B, and/or C” and “at least one of A, B, and C”, such phrasing is intended to encompass the selection of the first listed option (A) only, or the selection of the second listed option (B) only, or the selection of the third listed option (C) only, or the selection of the first and the second listed options (A and B) only, or the selection of the first and third listed options (A and C) only, or the selection of the second and third listed options (B and C) only, or the selection of all three options (A and B and C). This may be extended for as many items listed.

[0072] The foregoing is to be understood as being in every respect illustrative and exemplary, but not restrictive, and the scope of the invention disclosed herein is not to be determined from the Detailed Description, but rather from the claims as interpreted according to the full breadth permitted by the patent laws. It is to be understood that the embodiments shown and described herein are only illustrative of the present invention and that those skilled in the art may implement various modifications without departing from the scope and spirit of the invention. Those skilled in the art could implement various other feature combinations without departing from the scope and spirit of the invention. Having thus described aspects of the invention, with the details and particularity required by the patent laws, what is claimed and desired protected by Letters Patent is set forth in the appended claims.

WHAT IS CLAIMED IS:

1. A system for interpretable viewing interest, comprising:
 - a transformer with multi-head self-attention (320) configured to derive different hierarchical orders of input features by learning various semantic interactions of the input features in different semantic subspaces;
 - a plurality of hierarchical attention layers (330, 340, 350) configured to aggregate the different hierarchical orders of the input features to obtain aggregated single-order feature representations starting from a first order and continuing to a k-th order and derive aggregation attention weights for the different hierarchical orders of the input features based on an applied order of the plurality of hierarchical attention layers;
 - an attentional scoring layer (370) configured to evaluate the aggregated single-order feature representations to output a significance of each of the orders with respect to various Click-Through Rate (CTR) predictions;
 - a hierarchical interpretation layer (360) configured to determine a respective importance of each of the input features in various input feature combinations from which the various CTR predictions are derived based on the aggregation attention weights and the significance of each of the orders with respect to the CTR predictions;
 - and
 - a display device configured to display each of the various input feature combinations for the various CTR predictions along with the respective importance of each of the constituent one of the input features in the various input feature combinations to evaluate viewing interest of a displayed panel.

2. The system of claim 1, further comprising a plurality of field-aware embedding layers for generating feature embeddings from sets of categorical and numerical fields.
3. The system of claim 1, further comprising a plurality of feature-aware embedding layers configured to map different ones of the input features to low-dimensional feature embeddings having less than three dimensions.
4. The system of claim 1, wherein the transformer with multi-head self-attention processes categorical features and numerical as the input features.
5. The system of claim 4, wherein the categorical features are one-hot embedded prior to being processed by the transformer with multi-head self-attention.
6. The system of claim 1, further comprising a sigmoid function configured to generate each of the various predictions jointly based on the concatenated 1-order features and attention pooling vectors.
7. The system of claim 1, wherein the respective importance relates to a relevance degree of a particular one of the input features to a particular one of the various predictions.
8. The system of claim 1, wherein the feature interactions represent manifold factors that affect different ones of the various CTRs.

9. The system of claim 1, wherein the input feature lack temporal correlations.
10. The system of claim 1, wherein the aggregation attention weights are layer based such that different layers have different values for the aggregation attention weights.
11. The system of claim 1, wherein the significance indicates the orders of interacting ones of the input features having a highest impact to encourage a user to follow through on a given one of the various CTR predictions.
12. The system of claim 1, wherein the hierarchical interpretation layer interprets the various CTR predictions layer-by-layer and order-by-order.
13. The system of claim 1, wherein the transformer with multi-head self-attention learns the various semantic interactions of the input features in the different semantic subspaces by learning co-effects to text semantics of word pairs one of within a sentence or across multiple sentences regardless of the orders and distances of words in the word pairs, the word pairs being input feature values.
14. The system of claim 1, wherein the transformer with multi-head self-attention derives the different hierarchical orders of the input features based on pair-wise interactions of the input features.

15. The system of claim 1, wherein the transformer with multi-head self-attention, the plurality of hierarchical attention layers, the attentional scoring layer, and the hierarchical interpretation layer are comprised in a machine learning pipeline.

16. The system of claim 1, further comprising a hardware processor configured to control an order of presentation of web pages to a user such that the web pages go from highest likelihood to obtain a click through to lowest likelihood to obtain the click through responsive to each of the various input feature combinations for the various CTR predictions along with the respective importance of each of the constituent one of the input features in the various input feature combinations.

17. A method for interpretable viewing interest, comprising:
deriving (415), by a transformer with multi-head self-attention, different hierarchical orders of input features by learning various semantic interactions of the input features in different semantic subspaces;
aggregating (420), by a plurality of hierarchical attention layers, the different hierarchical orders of the input features to obtain aggregated single-order feature representations starting from a first order and continuing to a k-th order;
deriving (425), by the plurality of hierarchical attention layers, aggregation attention weights for the different hierarchical orders of the input features based on an applied order of the plurality of hierarchical attention layers;
evaluating (430), by an attentional scoring layer, the aggregated single-order feature representations to output a significance of each of the orders with respect to various Click-Through Rate (CTR) predictions;

determining (430), by a hierarchical interpretation layer, a respective importance of each of the input features in various input feature combinations from which the various CTR predictions are derived based on the aggregation attention weights and the significance of each of the orders with respect to the CTR predictions; and

displaying (435), by a display device, each of the various input feature combinations along with the respective importance of each of the constituent one of the input features in the various input feature combinations to evaluate viewing interest of a displayed panel.

18. The computer-implemented method of claim 17, wherein the significance indicates the orders of interacting ones of the input features having a highest impact to encourage a user to follow through on a given one of the various CTR predictions.

19. The computer-implemented method of claim 17, wherein the various CTR predictions are interpreted by the hierarchical interpretation layer in layer-by-layer and order-by-order.

20. The computer-implemented method of claim 17, wherein the various semantic interactions of the input features in the different semantic subspaces are learned by the transformer with multi-head self-attention by learning co-effects to text semantics of word pairs one of within a sentence or across multiple sentences regardless of the orders and distances of words in the word pairs, the word pairs being input feature values.

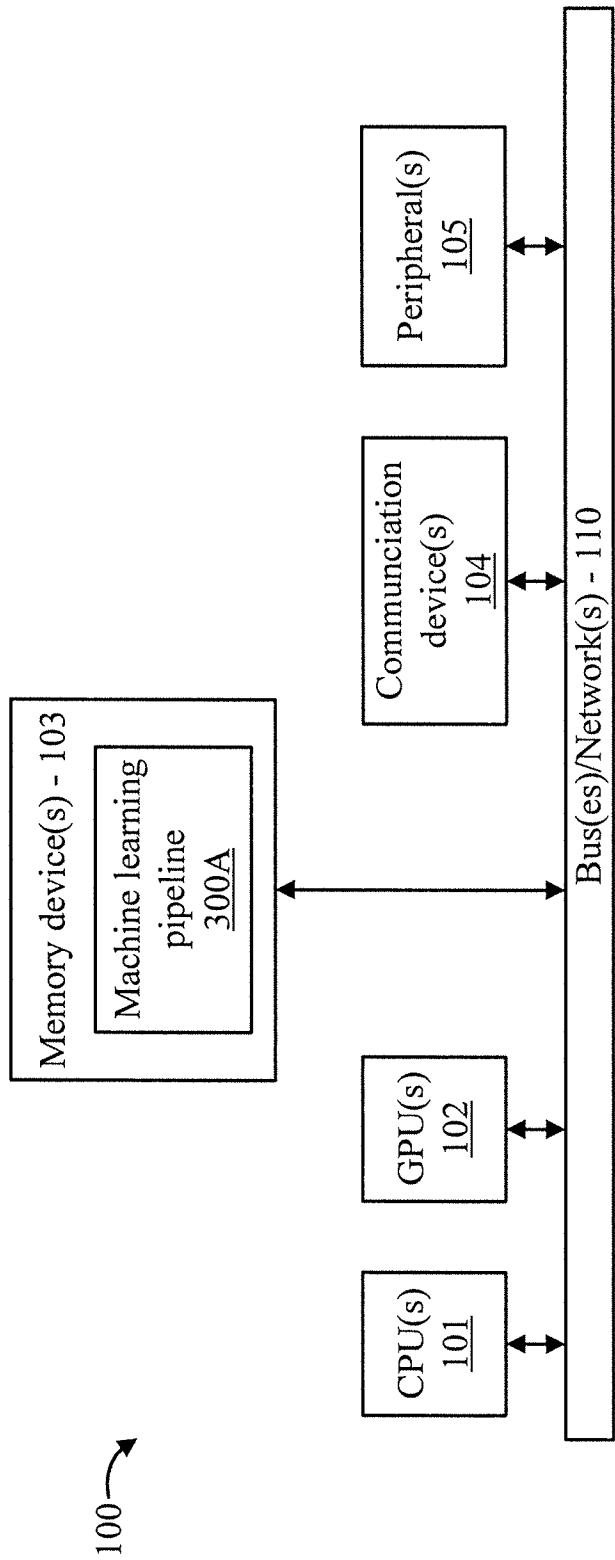


FIG. 1

200 →

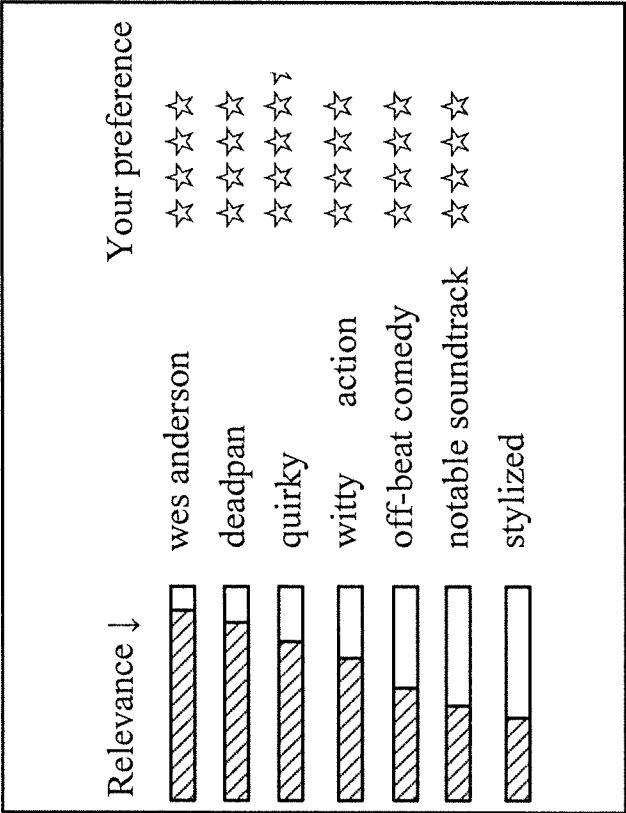


FIG. 2

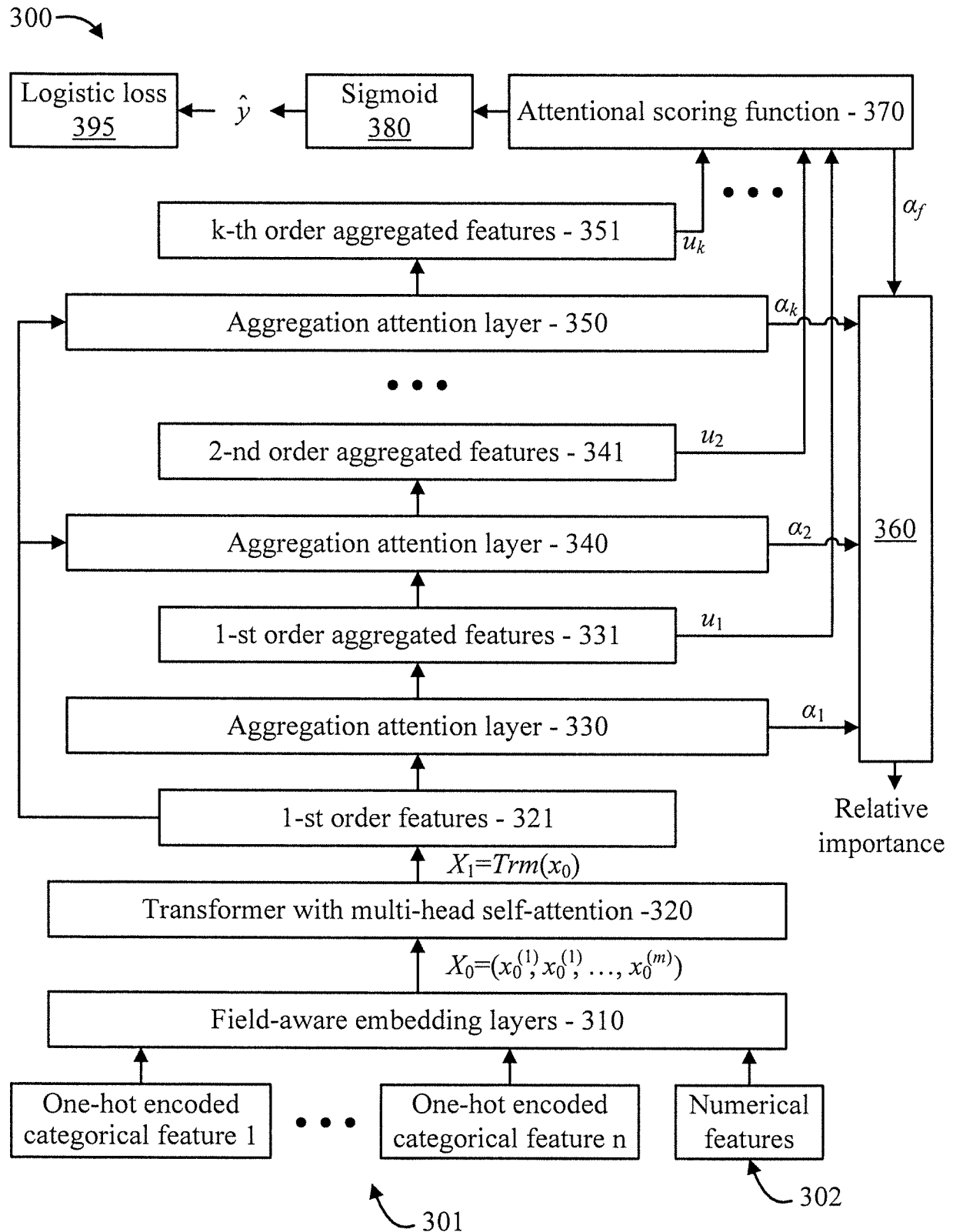


FIG. 3

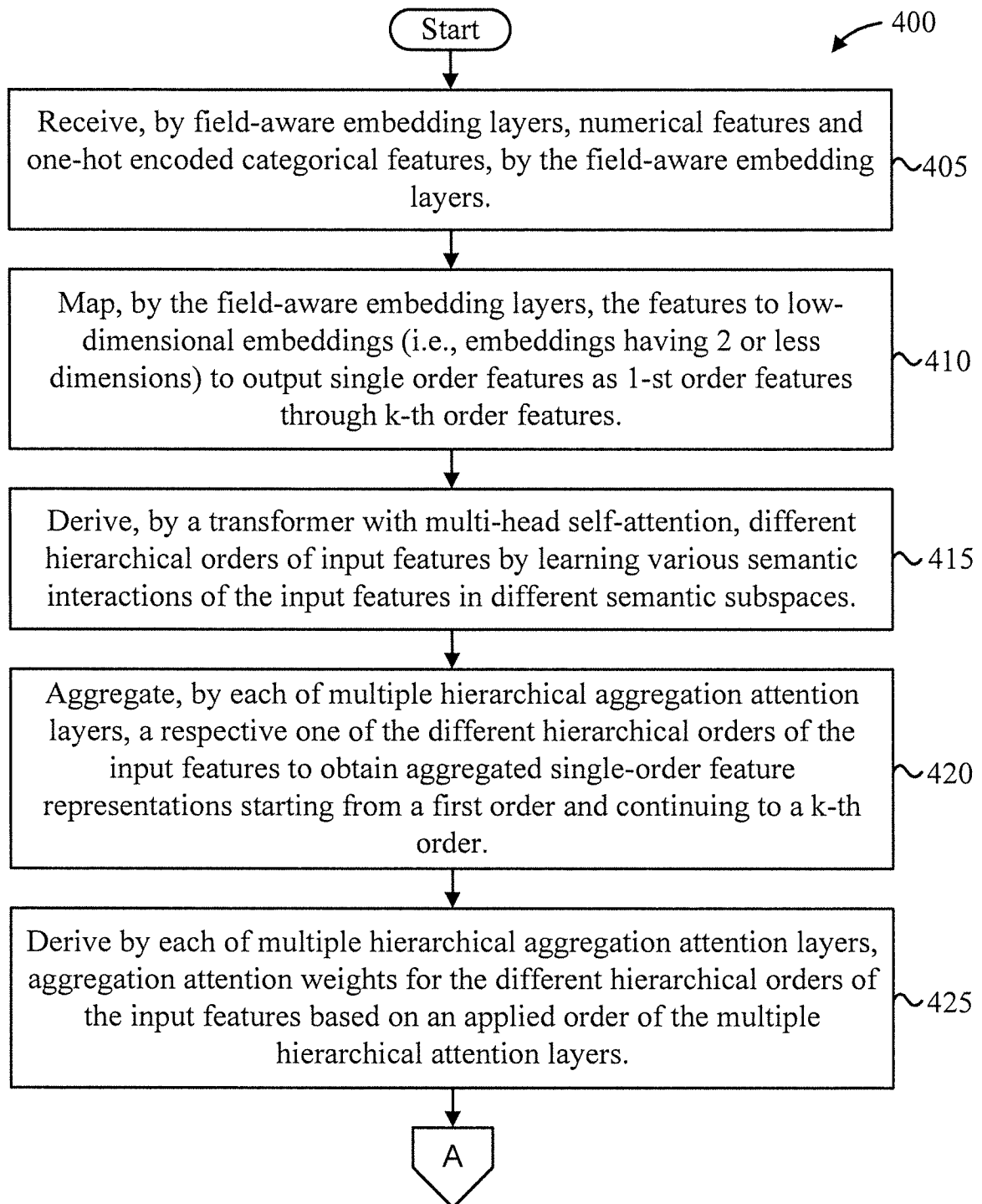


FIG. 4

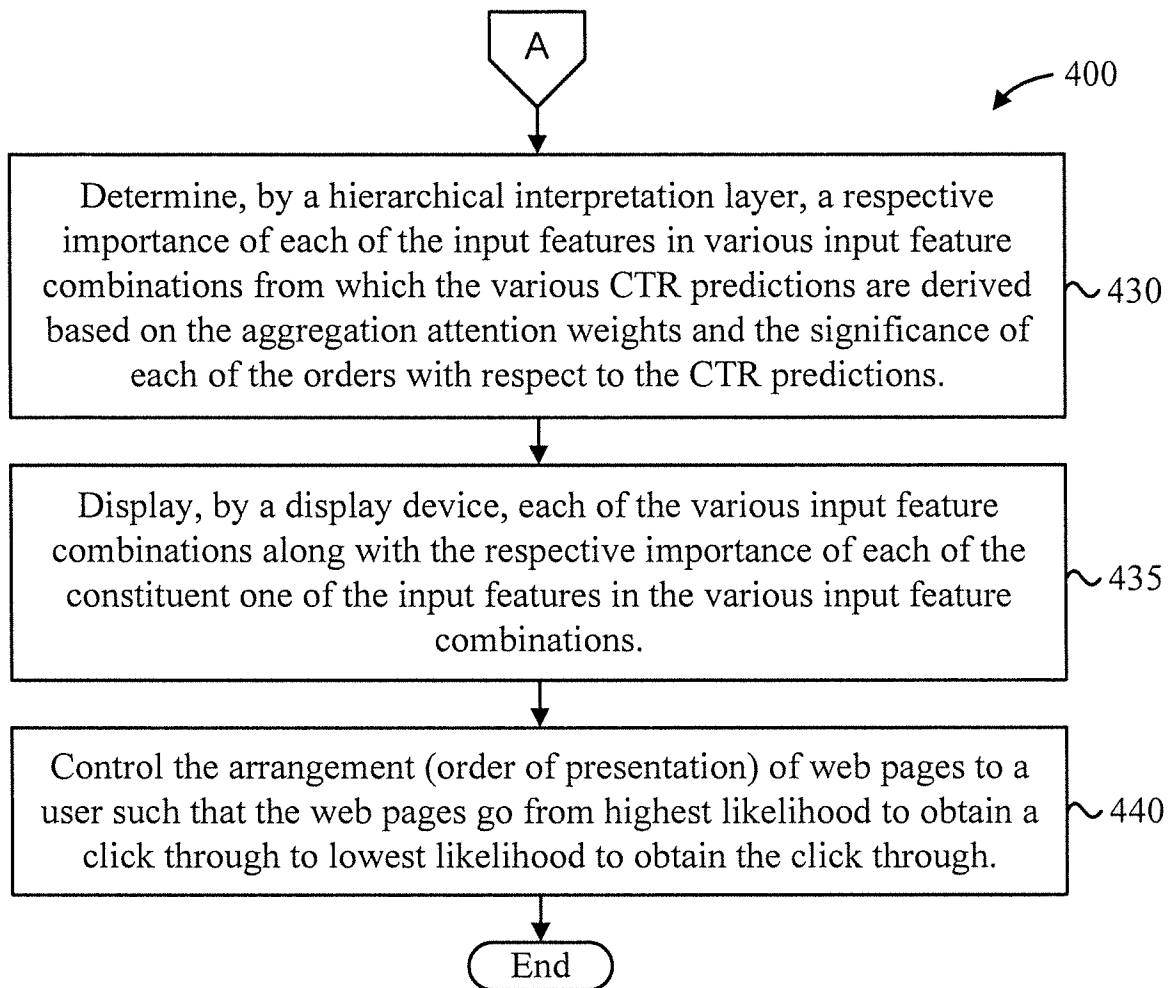


FIG. 5

A. CLASSIFICATION OF SUBJECT MATTER**G06Q 30/02(2012.01)i, G06F 16/2457(2019.01)i, G06N 20/00(2019.01)i**

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06Q 30/02; G06F 15/173; G06Q 10/00; G06Q 30/00; G06F 16/2457; G06N 20/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models

Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS(KIPO internal) & Keywords: interpretable viewing interest, aggregate, order, hierarchical attention layer, Click-Through Rate (CTR)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	QIANQIAN WANG et al.. A New Approach for advertising CTR Prediction Based on Deep Neural Network via Attention Mechanism. Computational and Mathematical Methods in Medicine, Volume 2018, Article ID 8056541, 13 September 2018, pages 1-11. Pages 1-8; and figure 3	1-20
A	HUIFENG GUO et al.. DeepFM: A Factorization-Machine based Neural Network for CTR Prediction. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), 13 March 2017, pages 1725-1731. Pages 1725-1730; and figures 1-5	1-20
A	GUORUI ZHOU et al.. Deep Interest Network for Click-Through Rate Prediction. 13 September 2018, pages 1-9. Pages 1-5; and figure 2	1-20
A	US 2011-0099059 A1 (DEEPAK K. AGARWAL et al.) 28 April 2011 Paragraphs [0040]-[0050]; and figures 3-4	1-20
A	US 2007-0156887 A1 (DANIEL WRIGHT et al.) 05 July 2007 Paragraphs [0026]-[0035]; and figures 1-2	1-20

☒ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"D" document cited by the applicant in the international application

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

12 June 2020 (12.06.2020)

Date of mailing of the international search report

12 June 2020 (12.06.2020)

Name and mailing address of the ISA/KR

International Application Division

Korean Intellectual Property Office

189 Cheongsa-ro, Seo-gu, Daejeon, 35208, Republic of Korea

Facsimile No. +82-42-481-8578

Authorized officer

HEO, Joo Hyung

Telephone No. +82-42-481-8150



INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2020/017861

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	ZEYU LI et al.. Interpretable Click-Through Rate Prediction through Hierarchical Attention. Technical Presentation, WSDM ' 20, 07 February 2020, pages 313-321. Pages 315-320; and figure 1	1-20

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2020/017861

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2011-0099059 A1	28/04/2011	US 8380570 B2	19/02/2013
US 2007-0156887 A1	05/07/2007	AU 2006-332534 A1	12/07/2007
		AU 2006-332534 B2	03/02/2011
		AU 2011-201512 A1	21/04/2011
		AU 2011-201512 B2	29/11/2012
		CA 2635040 A1	12/07/2007
		CN 101390118 A	18/03/2009
		EP 1969550 A2	17/09/2008
		JP 2009-522668 A	11/06/2009
		JP 4747200 B2	17/08/2011
		KR 10-1044683 B1	28/06/2011
		KR 10-2008-0086923 A	26/09/2008
		WO 2007-079405 A2	12/07/2007
		WO 2007-079405 A3	10/01/2008