Florida State University

TRANSFORMERS FOR RECSYS

Pramesh Regmi - **pr23n**
Pratima Sapkota - **ps23bc**
Sudheer Bommichetty – **sb23m**

PAPERS

PAPER 1

Deep Multifaceted Transformers for Multi-objective Ranking in Large-Scale E-commerce Recommender Systems

PAPER 2

Transformers4Rec: Bridging the Gap Between NLP and Sequential/Session-Based Recommendation

Paper 2

MOTIVATION

Why Transformer for Recommender Systems

- Current advancements in NLP is due to the ability of Transformer to understand long sequential relevance using attention.

- Natural language modeling is similar in a way to recommendation system modeling because of the sequential nature of both application.

NLP ⟶ RECSYS

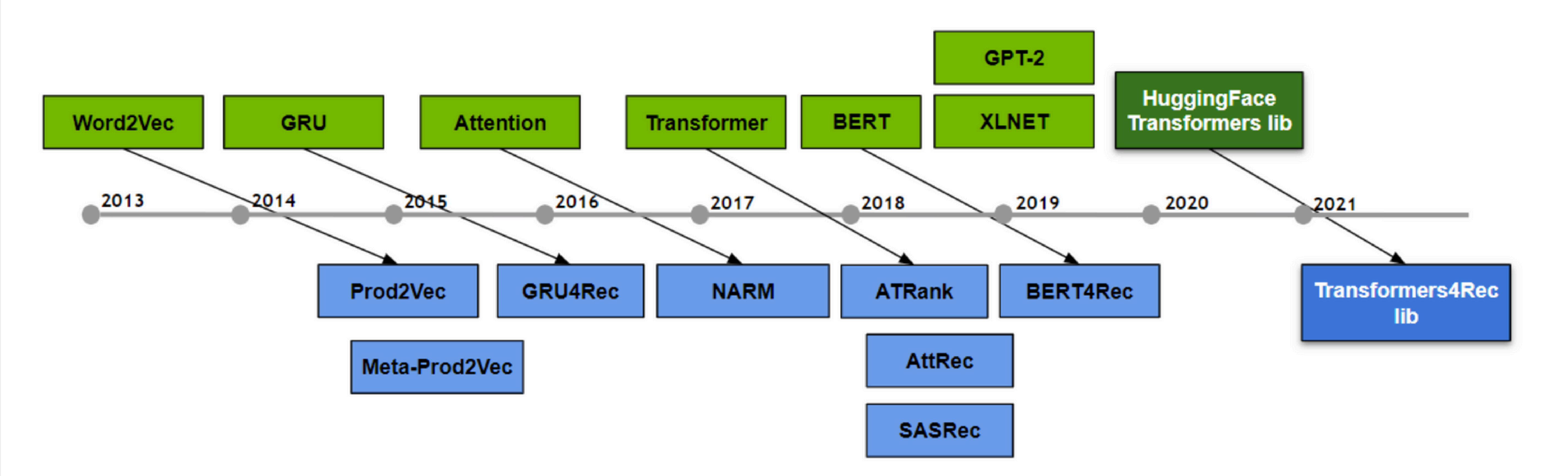


fig. The advancements in NLP have always been extended to RecSys within a few years of its inception.

Paper 1

INTRODUCTION

Key Role of Recommender Systems

- Enable personalized recommendations, primarily used for product suggestion (e-commerce) and content suggestions (social media)

Paper’s (and Our) Focus

- Study the ranking stage in e-commerce, crucial for determining what the user sees at the top
- Capture diverse user interests from behavior sequences over multiple time scales for robust modelling

CHALLENGES

Simultaneous Multi-objective Optimization:
How to Optimize CTR (likelihood of a click) and CVR (likelihood of a purchase after a click)

Joint Modeling of Diverse User Behaviors:
How to Integrate behaviors like clicks, adding items to the cart, and purchases into a unified framework

Reduction of Bias:
How to address selection bias (e.g., items at the top are more likely to be clicked) using novel techniques.

HYPOTHESIS

MULTI-OBJECTIVE LEARNING HYPOTHESIS

- Modeling and jointly optimizing multiple objectives (e.g., CTR and CVR) using shared representations can improve the overall performance.

MULTIFACETED INTEREST HYPOTHESIS

- Users' diverse behaviors (e.g., clicks, adds-to-cart, and orders) reflect distinct aspects of preferences and should be modeled independently.

THEORETICAL FRAMEWORK

Input Representation

- **Categorical Features**
    - Represent user-item interactions, such as product ID, category, and brand.
    - Each item in a user's behavior sequence is represented by embeddings for its associated attributes.

- **Dense features**
    - Includes user profile (e.g., purchase power, preferences), item profile (e.g., CTR, CVR), and user-item interaction features.
    - Normalized to ensure compatibility with neural network models.

IMPLEMENTATION FRAMEWORK

Deep Multifaceted Transformers for Multi-objective Ranking in Large-Scale E-commerce Recommender Systems

Yulong Gu[1], Zhuoye Ding[1], Shuaiqiang Wang[2], Lixin Zou[3], Yiding Liu[2], Dawei Yin[2]
[1]JD.com, Beijing, China [2]Baidu Inc., Beijing, China [3]Tsinghua University, Beijing, China
{guyulongcs,zoulixin15,liuyiding.tanh}@gmail.com
dingzhuoye@jd.com,wangshuaiqiang@baidu.com,yindawei@acm.org

ABSTRACT

Recommender Systems have been playing essential roles in e-commerce portals. Existing recommendation algorithms usually learn the ranking scores of items by optimizing a single task (*e.g.*, Click-through rate prediction) based on users' historical click sequences, but they generally pay few attention to simultaneously modeling users' multiple types of behaviors or jointly optimize multiple objectives (*e.g.*, both Click-through rate and Conversion rate), which are both vital for e-commerce sites. In this paper, we argue that it is crucial to formulate users' different interests based on multiple types of behaviors and perform multi-task learning for significant improvement in multiple objectives simultaneously. We propose Deep Multifaceted Transformers (DMT), a novel framework that can model users' multiple types of behavior sequences simultaneously with multiple Transformers. It utilizes Multi-gate Mixture-of-Experts to optimize multiple objectives. Besides, it exploits unbiased learning to reduce the selection bias in the training data. Experiments on JD real production dataset demonstrate the effectiveness of DMT, which significantly outperforms state-of-art methods. DMT has been successfully deployed to serve the main traffic in the commercial Recommender System in JD.com. To facilitate future research, we release the codes and datasets at https://github.com/guyulongcs/CIKM2020_DMT.

CCS CONCEPTS

• **Information systems → Personalization; Recommender systems;**

Figure 1: A real example in the Recommender Systems in JD.

- **Deep Multifaceted Transformers (DMT)**
  - Leverage multiple transformers to model user behavior sequences effectively.
- **Multi-gate Mixture-of-Experts (MMoE)**
  - Enable the system to manage complex relationships and conflicts between CTR and CVR.
- **Bias Deep Neural Network (BDNN)**
  - Use additional features to model and mitigate biases in training data.

Deep Multifaceted Transformers (DMT) Layer

- **Architecture**
  - Uses three distinct Transformers for each behavior type (clicks, adds-to-cart, orders).
  - These Transformers generate interest vectors for each behavior type.

- **Self-Attention Mechanism**
  - Learns relationships between items in a sequence by attending to all items simultaneously.
  - Encodes dependencies between items to capture the evolution of user preferences.

- **Positional Encoding**
  - Adds sequence information to embeddings.

Multi-Gate Mixture-of-Experts (MMoE) Layer

- **Purpose**
  - Model task-specific relationships and conflicts between CTR and CVR.
- **Architecture**
  - It uses N expert networks (MLPs) with ReLu, to model shared input and get the outputs of each expert.
  - For each task k, it exploits a gating network $NNGk$ to learn the weights of each expert, and get the weighted sum of expert outputs.
- **Theoretical Benefit:**
  - Allows tasks to share useful features while maintaining task-specific specialization.

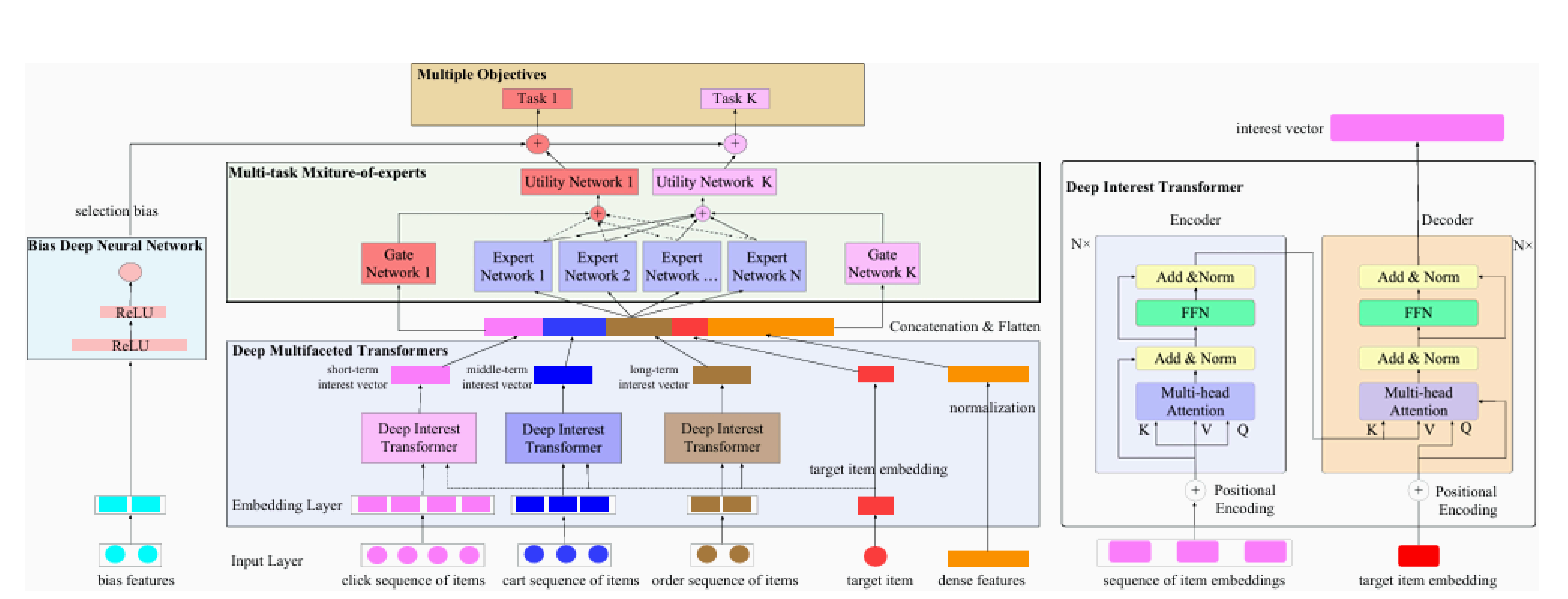
Bias Deep Neural Network (Bias DNN)

- **Purpose**
  - Correct biases inherent in implicit feedback data.

- **Modeled Bias Types**
  - **Position Bias**: Items in higher-ranked positions are more likely to be clicked.
  - **Neighboring Bias:** Interaction probabilities are influenced by surrounding items.

- **Implementation**:
  - Bias features are embedded and processed through MLPs to estimate bias correction.
  - Position bias is modeled by using "Position_index" and "Position_page", derived from item's rank within the recommendation list and its page position.
  - Neighboring bias is corrected using item category and its six nearest neighbors which are embedded into low-dimensional vectors and processed through MLPs.

OVERALL ARCHITECTURE



Deep Multifaceted Transformers (bottom), is consisted of multiple Deep Interest Transformers (right), to extract users' multifaceted interests from their diverse behavior sequences, exploits Multi-gate Mixture-of-Experts (MMoE) (top) to simultaneously optimize multiple objectives, and uses a Bias Deep Neural Network (left) to reduce the bias in training data.

MATHEMATICAL FOUNDATIONS

01

SELF-ATTENTION MECHANISM

- Q, K, V: Query, Key, and Value matrices.
- d_k = Dimensionality of keys.

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

02

MULTI-GATE MIXTURE OF EXPERTS

For task *k*: w_i^k=Task-specific gating weight

For expert *i*, $e_i(x)$= Output of *i*

$$f^k(x) = \sum_{i=1}^{N} w_i^k e_i(x)$$

03

BIAS-CORRECTION

Bias-adjusted utility score for task

$$u_k = \sigma(u_k^{MMoE} + y_b)$$

Training Details

- Loss Function
  - Both objective (CTR and CVR) use cross-entropy loss.
  - Total loss is a weighted sum of individual task losses
- Bias Correction
  - During training, utility scores from the MMoE layer are adjusted with bias terms estimated by the Bias DNN.

Prediction Details

- For each task k, a sigmoid activation is applied to the task-specific utility score to compute probabilities.
- Final ranking scores are computed as a weighted sum of task-specific scores

RESULT

Significantly outperforms state-of-the-art baselines (DIN, DIEN, GBDT) on JD.com's dataset, achieving substantial improvements in both click and order prediction metrics.

Metric	Baseline (GBDT)	DIEN	DMT (Without Bias)	DMT (With Bias)
CTR Improvement	0%	+14.3%	+18.2%	**+18.8%**
CVR Improvement	0%	+14.6%	+16.9%	**+19.2%**
GMV Improvement	0%	+11.9%	+16.2%	**+17.9%**

Paper 2

HUGGINGFACE

PREPROCESS

Integrated with NVTabular for large-scale, GPU-accelerated feature engineering.

TRAIN

Provides modular, configurable pipeline for training with ranking-based metrics.

EVALUATE

Supports session-based recommendation-specific evaluation metrics (like NDCG@20, Recall@20) and incremental evaluation for production-like scenarios.

CONCRETE VALIDATION

Dataset	Metric	Best Transformer (Method)	Performance	Best Baseline	Baseline Performance	Improvement (%)
REES46 (eCommerce)	NDCG@20	XLNet (RTD)	0.2546	GRU4Rec (FT)	0.2231	+14.15
	HR@20	XLNet (RTD)	0.4886	VSTAN	0.4857	+0.60
YOOCHOOSE (eCommerce)	NDCG@20	XLNet (RTD)	0.3776	GRU4Rec (FT)	0.3442	+9.75
	HR@20	BERT (MLM)	0.6349	GRU4Rec (FT)	0.5891	+7.78
G1 (News)	NDCG@20	ELECTRA (RTD)	0.3588	GRU	0.3549	+1.10
	HR@20	XLNet (PLM)	0.6634	GRU	0.6632	+0.03
ADRESSA (News)	NDCG@20	XLNet (MLM)	0.3822	GRU	0.3799	+0.61
	HR@20	XLNet (CLM)	0.7378	GRU	0.7413	-0.47

CONCLUSION

- Transformers architectures have higher performance for recommendation systems for e-commerce than any other baselines.
- Modeling diverse behaviors distinctly provides a strong modeling however, needs a module to confirm to the task specific conflicts.
- Modeling RecSys by incorporating biases strengthens the recommendation performance.
- With integration with every auto-diff library and availability of transformers specifically for RecSys, the future seems transforming.

FUTURE DIRECTIONS

IMPLEMENTATION

An example trial can be done with the pre-existing transformers library and available datasets.

VALIDATION

The results of the primary paper can be verified by simply training a transformer using HuggingFace’s transformer library and testing against existing metrics.

Florida State University

THANK YOU

fsu.edu