Joint Attention Neural Model for Demand Prediction in Online Marketplaces

Ashish Gupta[1], Rishabh Mehrotra[2]

[1]Walmart Labs, India, [2]Spotify Research, London, UK
ashish.gupta@walmartlabs.com, rishabhm@spotify.com

Abstract

As an increasing number of consumers rely on online marketplaces to purchase goods from, demand prediction becomes an important problem for suppliers to inform their pricing and inventory management decisions. Business volatility and the complexity of factors influence demand, which makes it a harder quantity to predict. In this paper, we consider the case of an online classified marketplace and propose a joint multi-modal neural model for demand prediction. The proposed neural model incorporates a number of factors including product description information (title, description, images), contextual information (geography, similar products) and historic interest to predict demand. Large-scale experiments on real-world data demonstrate superior performance over established baselines. Our experiments highlight the importance of considering, quantifying and leveraging the textual content of products and image quality for enhanced demand prediction. Finally, we quantify the impact of the different factors in predicting demand.

1 Introduction

In recent years, two sided marketplaces have steadily emerged as leading business models for many real world scenarios, including accommodation (Airbnb, Booking.com), app stores (Apple and Google App Store) and online shopping (Etsy, Ebay). These marketplaces act as intermediaries between suppliers and consumers, and facilitate transactions between the buyer and the seller. In order to efficiently manage inventory and price commodities, suppliers rely on accurately predicting demand for their products.

When selling such goods online, a combination of tiny, nuanced details in a product description can sway consumer interest and consequently, impact sale and revenue. The product description shabbily written or written in an ambiguous way, the image of the product being either too dull or too glossy, etc. - all these factors negatively impact sales and hence, demand for a particular product. This leads the supplier to think that the product didn't meet the expectations of people. Additionally, even with an optimized product listing, demand for a product may simply not exist - frustrating sellers who may have over-invested in marketing. This emphasizes the need to understand the interplay between different factors affecting consumer behavior, in order to give suppliers a good approximation of the demand for their product.

In this paper, we tackle the problem of demand prediction in classified marketplace, and leverage and quantify the impact of a number of factors including product description information (title, description, images), contextual information (geography, similar products) and historic interest in the product to predict demand. We propose a joint multi-view neural model for demand prediction, which jointly encodes textual information with image specific features, along with contextual and historic signals to accurately predict demand. We jointly model images via deep residual network [5] over image data with shortcut connections, with a bidirectional LSTM with 1D convolution over text, and train the model end-to-end for demand prediction task.

We perform large scale experimentation on real world data comprising of over 2 million products and 20M purchases, and compare the performance

https://doi.org/10.7557/18.5170

of the proposed neural model and demonstrate superior performance with over 18% improvement in prediction results over established baselines. Further, we quantify the impact of different factors contributing to the prediction, and observe that good textual description and images are most predictive of demand among all factors considered. Our results inform the design of inventory management and pricing strategies based on demand in online marketplaces.

2 Related Work

Problems of demand prediction could include predicting many different signals, such as CTR(Click Through Rate), traffic related data, traffic flow, money made from a particular campaign and utilizing these in online marketplaces.

Demand Prediction is heavily based on pricing of a product as well as traffic coming from a particular location [11]. Traditional approach have centered around using autoregressive integrated moving average (ARIMA) and its variants [10]. Furthermore, past work has begin exploring the use of additional meta-data for demand prediction, including image features and visual appearance [3]. [2] use Bayesian hierarchical structure to disseminate the cross-series information across a set of time series.

As the volume of data increase, more complex methods based on neural network are proposed to predict CTR [8]. Prior work has also leveraged how the user behaved along with the past time, and presented RNN based [14] and CNN based models [8] to predict demand.Interested readers are referred to a recent tutorial [9] on recommendations in a marketplace for a broader description of recent research on the topic.

In this paper we include image, text, location, date and pricing to predict demand of any particular product. We propose a technique of amalgamating these features so as to get holistic view of this problem and get better insights.

3 Multimodal Neural Model

Demand of a product in an online marketplace refers to how many consumers are inclined and willing to purchase the product. A number of factors contribute towards predicting demand of items in e-commerce marketplaces. While aspects like region, city, price and product category are direct indicators, often the quality of product description and images describing the product play a major role in garnering consumer interest in the product. Our goal is to jointly model such varied information in predicting demand.

Data	Description	Example
item_id	Ad id	b912c3c6a6ad
region	Ad region	Rostov Oblast
city	Ad city	Rostov-on-Don
parent_category	Top level category	Consumer electronics
category_name	ad category	Audio and video
title	Ad title	Philips bluray
description	Ad description	home theater with bluray
param_1	Optional parameters for Ad	Video, DVD and Blu-ray players
param_2	Optional parameters for Ad	NaN
price	Ad price	4000.0 ruble
activation_date	Date ad was placed	2017-03-20
image	Id code	b7f250e
demand_probability	selling likelihood	0.43177

Table 1: Dataset description

Figure 1: High vs Low demand products: the image quality plays a major role in gathering consumer interest in a product, and is a good predictor of demand.

3.1 Data Context

Table 1 describes about the dataset which we used to perform experiments mentioned in the paper. It consists of over 20M advertisments ranging from products used for personal belongings, consumer electronics, home and garden, real estate, hobbies & leisure, transport, business, etc. The users who purchased these products are categorized into private, company or even retail shops. Total data was divided into 15M train and 5M test sets.

3.2 Metadata Representations

To leverage textual meta-data of products, we extract a number of features.

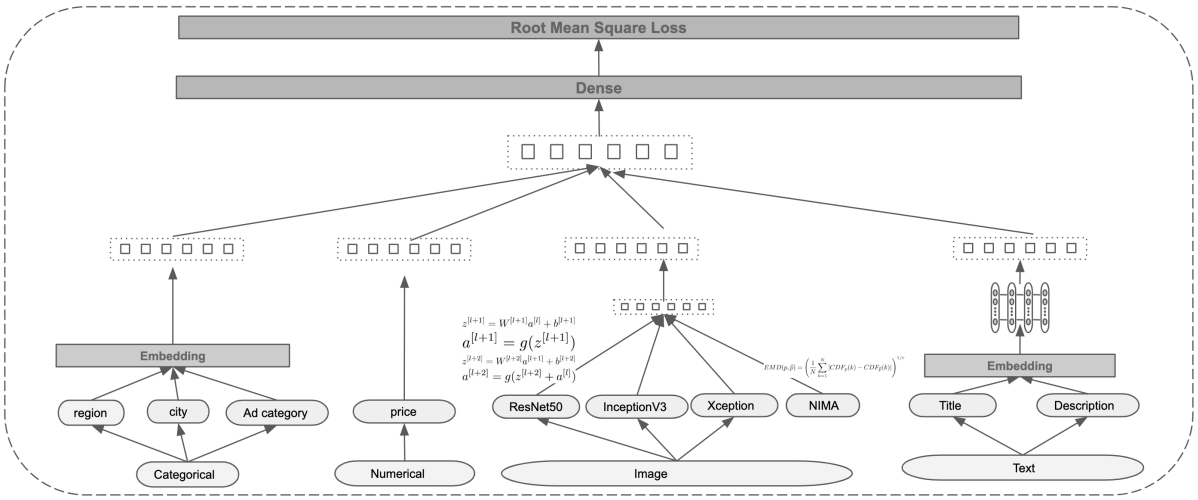

Figure 2: Overview of the proposed neural model for the advertised products data.

Text features: We extract tf-idf on product title, description and other textual metadata(param_1, param_2) attached with the product.

Bi-directional LSTMS for Textual Embeddings: In addition to tf-idf features, we train a 300 dimension vector embedding for each word in the metadata by training a fastText embedding model [1]. Vectors from words in title and description are concatenated tf-idf based features extracted from unigrams and bigrams.

We apply a Spatial Dropout on the title, description and meta data combination to improve generalization performance since some of the features were strongly correlated. We feed the concatenated representation for each word to a bi-directional LSTM model to learn sequential representation of the description, title and meta-data. Then, we perform global max pooling.

$$u_t^W = BiLSTM(u_{t-1}^W, w_t^W, c_t^W) \quad (1)$$

where $u_1^W, ..., u_n^W$ is concatenated representation for words present in title and description which are obtained after applying Spatial Dropout on the embeddings of those words. Words in title and description are converted to their word-level and character level ($\{w_t^W\}_{t=1}^n$ and $\{c_t^W\}_{t=1}^n$) embedded representation.

Embedding Categorical Features: We learn an embedding for numerical features like price and categorical features obtained from region, city, category_name and parent_category_name attributes. The parent category for the items were personal belongings, real estate, consumer electronics, hobbies & leisure, transport, services. The categories for the items were clothes, shoes, accessories, etc.

We concatenate the output of the global max pooling from bi-directional LSTM over textual features and the flattened representation of embeddings of region, city, parent category name and category name.

Finally, we applied attention over the hidden representation obtained from BiLSTM to extract specific words like:

1. car cradle in a good condition
2. Wedding shoes sell new shoes of the firm Louisa Peeress 37r

The 1^{st} item shows a high demand due to words like good condition of car cradle compared to the 2^{nd} item like selling new wedding shoes which is very obvious for wedding shoes to be new. Hence, the 2^{nd} item description is ambiguous hence, has low demand due to these words.

3.3 Learning Image Representations

Image Features: Image quality significantly impacts demand of the product. If an image is too dull, white, blurry, large, small, etc this hampers the demand for a product as it does not

make an appropriate impact on customer and they prefer another similar product over it. To extract base representation of images, we employ InceptionV3[12] and Xception[4] networks, both trained on ImageNet. We classified the image into their respective categories using these networks. We noticed an interesting result while classification of images i.e. when these networks are not able to clearly predict what the image is then the demand probability of the product also went down. The correlation between demand_probability(a.k.a demand_probability) and image_confidence was stronger than with title and description of product. Also, this shows a clear correlation between human's perception of a product with the correct image classification. Hence, if the image is ambiguous to these image networks then it will be ambiguous to the human beings too and thus unattractive as a product.

Further, we used NIMA(Neural Image Assessment)[13] where a deep CNN is trained on the images to predict which images a typical user would rate as good(technically) or attractive (aesthetically). In NIMA, the last layer of the baseline CNN with a fully-connected layer with 10 neurons followed by softmax activations. Baseline CNN weights are initialized by training on the ImageNet dataset[7], and then an end-to-end training on quality assessment is performed. Our goal while using NIMA is to predict the distribution of ratings for a given image. Ground truth distribution of human ratings of a given image can be expressed as an empirical probability mass function $\mathbf{p}$=$[p_{s_1}....p_{s_N}]$ where $s_1 \leq s_i \leq s_N$, where s_i denotes the i^{th} score bucket, and N denotes the total number of score buckets.

Loss function used in NIMA is minimizing Earth Mover's distance(EMD)

$$EMD(p,\widehat{p}) = \left(\frac{1}{N}\sum_{k=1}^{N}|CDF_p(k) - CDF_{\widehat{p}}(k)|\right)^{1/r} \tag{2}$$

where $CDF_p(k)$(Cumulative Distribution Function) is $\sum_{i=1}^{k} p_{s_i}$. Also this closed-form solution requires both distributions to have equal mass as $\sum_{i=1}^{k} p_{s_i} = \sum_{i=1}^{k} \widehat{p}_{s_i}$. Here, r is set as 2 to penalize the Euclidean distance between the CDFs. r = 2 allows easier optimization when working with gradient descent.

Deep Residual Network for Image Representation: In addition to NIMA and Inception model based representations, we extract image features using a pre-trained deep residual model (ResNet34[5]). ResNet used because it solves vanishing/exploding gradient problems. ResNet34 did not gave better results so we switched to ResNet with 50 layers. Our ResNet50 model is pre-trained on ImageNet data and consists of 50 convolution layers of 3*3 filters and has advantage of using skip connection (adding shortcut connections). Image is fed as an input at layer $l^{[0]}$. Activation of a particular layer is calculated by using non-linearity on previous layer's output as well as current activation from previous to previous layer's activation.

$$z^{[l+1]} = W^{[l+1]}a^{[l]} + b^{[l+1]} \tag{3}$$

$$a^{[l+1]} = g(z^{[l+1]}) \tag{4}$$

$$z^{[l+2]} = W^{[l+2]}a^{[l+1]} + b^{[l+2]} \tag{5}$$

$$a^{[l+2]} = g(z^{[l+2]} + a^{[l]}) \tag{6}$$

where $W^{[l+1]}$ are weights which needs to be tuned in layer l+1, $a^{[l]}$ is the activation obtained by using ReLU non-linearity over the output from previous layer, $b^{[l+1]}$ is the bias term in layer l+1, $z^{[l+1]}$ is the output obtained by taking dot product of activation from previous layer and the weights to be tuned in the current layer. In Eqn.(4) there is a skip connection added from layer l and thus the activation in layer l+2 is obtained by using non-linearity over output from previous layer as well as activation from previous to previous layer.

3.4 Joint Multi-modal Neural Model

Our joint neural model combines the different representations learnt by applying dropout of 0.1, followed by a layer ReLU non-linearity and then passed through a fully connected layer with sigmoid activation function to get the final demand of the product. Figure 2 shows the overall architecture in detail. We used a minibatch size of 64, *Adam* for optimization with hyperparameters such as learning rate(α) of 0.004, β_1 as 0.9 and β_2 as 0.996. The

Model	Text	Categorical	Image	RMSE
vLSTM	Y	N	N	0.251
catEmd	N	Y	N	0.259
vCNN	N	N	Y	0.263
LogReg	Y	Y	Y	0.2461
GBDT	Y	Y	Y	0.2343
BiLSTM + CNN1D	Y	Y	N	0.230
2 BiLSTM + ResNet50	Y	Y	Y	0.224
BiLSTM + CNN1D + ResNet50	Y	Y	Y	**0.215***&
BiLSTM + Attention + CNN1D + ResNet50	Y	Y	Y	**0.201***&

Table 2: Performance of different models on RMSE error (lower is better) with information on the use of different types of features. * and & signify statistically significant difference between the method and top 2 best performing baseline using χ^2 test with $p \leq 0.05$

loss function for the model is Root Mean Squared Error $\sqrt[2]{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}$ where $\hat{y}$ is the predicted demand and y is the actual demand.

4 Experimental Evaluation

We evaluate the efficacy of the proposed neural model on a large scale real world dataset of classified ads from a major commercial online ads platform. Our training data consists of over 15M ads samples comprising of various attributes like item_id, user_id, region, city, parent_category_name, category_name, title, description, price, user_type, image and demand_probability telling demand for an online ad. We train the neural model to predict demand_probability, i.e., the probability of the consumer buying the product.

4.1 Baselines

We compare the proposed neural model with various baselines, both non-neural and neural:

1. LogReg: logistic regression over feature combination (text, image and categorical features)
2. CatEmd: categorical embedding
3. GBDT: gradient boosting decision trees
4. v-LSTM: vanilla LSTM over textual features
5. v-CNN: vanilla CNN over images

The different models are evaluated on RMSE (Root Mean Squared Error) error.

4.2 Benefit of neural model

Table 2 presents the RMSE results. We observe significant gains offered by the neural over non-neural baselines, with the same set of features. The classical Gradient boosting method gave an RMSE of about 0.2343, LightGBM model[6] with features based on date like weekday, week of year, day of month gave an RMSE of 0.232 while the best Neural model gave an RMSE of 0.201. This highlights that the neural model is better able to model the interplay between various features in order to predict demand probability.

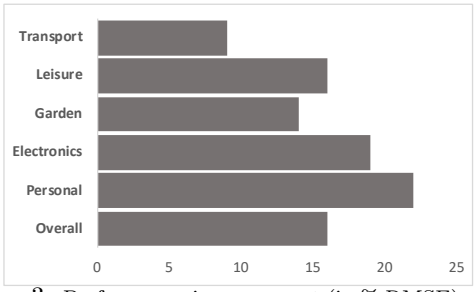

Figure 3: Performance improvement (in % RMSE) against BiLSTM model offered by adding image metadata across different categories.

4.3 Importance of Joint Modeling

We find our model with categorical, numerical and textual features gave RMSE of about 0.23 when image features were not included while it gave 0.201 as RMSE when image features were used. Baselines using partial information (text vs image vs categorical) perform worse than models using all information, there by highlighting the importance of jointly leveraging all information.

Among models which use all information, we observe that the proposed model outperforms logistic regression and boosted trees, which use similar information but lack the understanding of hierarchical structure in data (presence of images and text). We observe an improvement of over 12% in RMSE errors with the proposed model. Finally, we also observe that leveraging attention mechanism across different metadata helps the model improve performance.

4.4 Impact of Different Features

Upon measuring the impact of different features in predicting demand, we observe that textual features are more predictive than categorical features. Further, the proposed model observed a boost of over 4% when images were included. Hence, we find that the images are necessary to make better

predictions.

4.5 Performance across Categories

Comparing the importance of image features in predicting demand across different product categories (Fig 3), we observe that personal belongings and electronics benefit most by addition of image features over a textual metadata only model. Indeed, images play a major role in eliciting response from buyers, which places a strong emphasis on modeling image features into demand prediction module.

5 Conclusion

We proposed a joint multi-model neural architecture to predict demand in an online classified ads marketplace. Experimental results demonstrate the benefit of leveraging multimodal data (textual description, along with product features and image features). Our findings highlight the need for including good descriptive text and images in the ads, for attracting consumers.

References

[1] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[2] N. Chapados. Effective bayesian modeling of groups of related count time series. *arXiv preprint arXiv:1405.3738*, 2014.

[3] H. Cheng, R. v. Zwol, J. Azimi, E. Manavoglu, R. Zhang, Y. Zhou, and V. Navalpakkam. Multimedia features for click prediction of new ads in display advertising. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 777–785. ACM, 2012.

[4] F. Chollet. Xception: Deep learning with depthwise separable convolutions. *arXiv preprint*, pages 1610–02357, 2017.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pages 3146–3154, 2017.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[8] Q. Liu, F. Yu, S. Wu, and L. Wang. A convolutional click prediction model. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1743–1746. ACM, 2015.

[9] R. Mehrotra and B. Carterette. Recommendations in a marketplace. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 580–581, 2019.

[10] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas. Predicting taxi–passenger demand using streaming data. *IEEE Transactions on Intelligent Transportation Systems*, 14(3):1393–1402, 2013.

[11] Ö. Özer, O. Ozer, and R. Phillips. *The Oxford handbook of pricing management*. Oxford University Press, 2012.

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[13] H. Talebi and P. Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.

[14] Y. Zhang, H. Dai, C. Xu, J. Feng, T. Wang, J. Bian, B. Wang, and T.-Y. Liu. Sequential click prediction for sponsored search with recurrent neural networks. In *AAAI*, volume 14, pages 1369–1375, 2014.