PAN: PERSONALIZED ATTENTION NETWORK FOR OUTFIT RECOMMENDATION

Huijing Zhan[1], *Jie Lin*[1]

[1]Institute for Infocomm Research, A*STAR, Singapore

2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506344

ABSTRACT

Recent years have witnessed the dramatic development of e-fashion industry, it becomes essential to build an intelligent fashion recommender system. Most of existing works on fashion recommendation focus on modeling the general compatibility while ignoring the user preferences. In this paper, we present a Personalized Attention Network (PAN) for fashion recommendation. The key component of PAN includes a user encoder, an item encoder and a preference predictor. To modeling users' diverse interests, we develop an attention network to incorporate the learnt user representation into the item encoder component. More specifically, the attention module consists of a sequential user-aware channel-level and a spatial-level sub-module. Moreover, a novel ranking an user-specific loss, is proposed to capture the interest of different users on the same outfit. To make the training more effective and efficient, a novel user-aware online hard negative mining strategy is proposed. Extensive experiments on Polyvore-U dataset demonstrate the excellence of the proposed system and the effectiveness of different modules.

***Index Terms*—** Fashion Outfit Recommendation, Personalized, Attention-based, Ranking Loss, Online Hard Negative

1. INTRODUCTION

With the rapid expansion of online fashion market, massive items are emerged and posted online everyday [1]. This poses great difficulties for users to decide their desired ones from tons of products. Therefore, building an automatic fashion recommender system is in great needs. Moreover, it can bring huge benefits in both industries and research communities.

The majority of the research works on fashion recommendation [2, 3, 4, 5, 6] are dedicated to predicting the generic compatibility between different items within a fashion outfit. Veit *et al.* [3] leveraged the Siamese network for the relationship between co-purchased products. Han *et al.* [4] proposed to predict the outfit compatibility by extracting the feature representation with LSTM model. The representation power has proven to be significant in improving the matching quality. Variants of attention mechanisms have been proposed [7, 8] to extract informative features by learning where to focus. Taking the complexity of the attention blocks into account, Convolutional Block Attention Module (CBAM) is utilized as the starting backbone, flexible to be plugged into any networks.

Fig. 1. Examples of user-created outfits according to their fashion tastes.

Our framework is built upon the pioneering work of [9], in which Lu designed an efficient personalized fashion outfit recommendation system via hashing techniques. However, it fails to capture the different users' interests on different aspects of the outfit, *e.g.,* logos on the clothes, patterns of the T-shirt or category preference bias. For example, as shown in Fig. 1, both user 1 and 2 have added outfit 1 into their "like" lists. From their respective clicking histories, it can be seen that user 1 may prefer "black and white pointy shoes" for its pointy structure and the black color probably plays an important role in user 2's decision.

To address the issue of distinctive attention of fashion items for different users, in this paper, we present a personalized attention network (PAN) for fashion recommender system. More specifically, the user embedding is integrated into the item representation model to compute the user-aware channel-level and spatial-level attention maps. Besides optimizing the network in the Beyesian Personalized Ranking (BPR) [10] criterion with a triplet of one user and two items [11], we develop a novel ranking loss with a triplet of two users and one item. This can efficiently capture the diverse interests on the same items for different users. To differentiate the preferences of different users and train the network more effectively, we propose a simple but effective online user-aware hard negative mining approach, which evaluates the difficult level of the outfits based on the loss and the similarity of learnt user embeddings. Extensive experiments on the Polyvore-U dataset (Polyvore-630 and Polyvore-519) validate the excellence of our proposed system and also the effectiveness of each module.

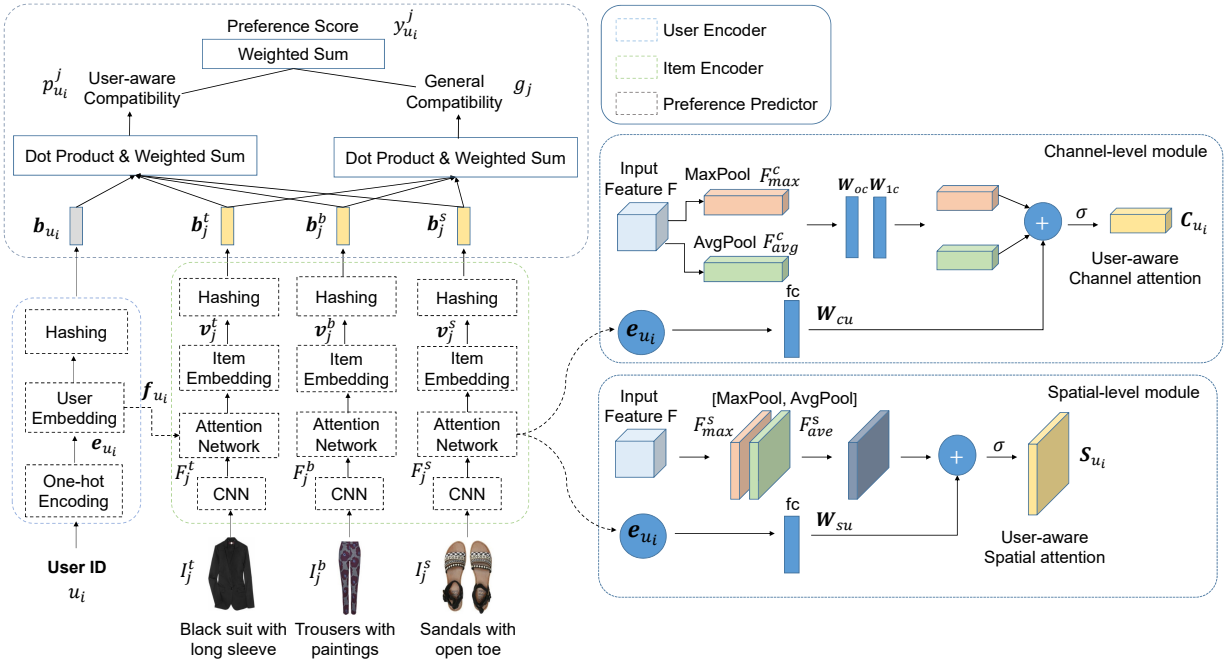

Fig. 2. The pipeline of the proposed PAN for fashion outfit recommendation.

2. PROPOSED METHOD AND PROBLEM FORMULATION

As shown in Fig. 2, there are three essential blocks in our framework, corresponding to: 1) an item encoder, which aims to learn the item embedding; 2) a user encoder, which aims to learn the embedding of user; 3) a preference predictor, which is utilized to predict the preference score of the candidate outfit. The channel- and spatial-level attention mechanism is incorporated in learning the feature map of the item encoder to select discriminative features according to individual preferences. Within the preference predictor module, the preference score is a weighted combination of general compatibility and user-aware compatibility score.

2.1. User Encoder

Given the user ID u_i, it is firstly encoded into an one-hot vector $\mathbf{e}_{u_i} \in \mathbb{R}^M$ by setting all values to 0 except the u-th entry. Here M is the total number of users. Then one fully-connected (fc) layer is utilized to further encode $\mathbf{e}_{u_i}$ into the user embedding f_{u_i}, as shown below:

$$\mathbf{f}_{u_i} = \mathbf{F}\mathbf{e}_{u_i}, \tag{1}$$

where $\mathbf{F} \in \mathbb{R}^{D \times M}$ denotes the parameter of the weight matrix and D is the user embedding size. The hashing sub-module is then utilized to replace the continuous embedding $\mathbf{f}_{u_i}$ with discrete binary representation $\mathbf{b}_{u_i}$.

2.2. Item Encoder

The image sequence (I_j^t, I_j^b, I_j^s) of the outfit are forwarded into a deep CNN to obtain the item's feature map $(\mathbf{F}_j^t, \mathbf{F}_j^b, \mathbf{F}_j^s)$. Given the intermediate feature map $\mathbf{F}$, the proposed user-aware attention mechanism infers the channel map $\mathbf{C}_{u_i} \in \mathbb{R}^{C\times1\times1}$ and the spatial map $\mathbf{S}_{u_i} \in \mathbb{R}^{1\times H\times W}$. The channel-level attention and spatial-level attention submodules are sequentially appended to the network for achieving the item embedding $(\mathbf{v}_j^t, \mathbf{v}_j^b, \mathbf{v}_j^s)$. Note that for illustration simplicity, the text embedding branch is not presented in Fig. 2. 基本和block attention保持一致，多加了用户的embedding

Channel-level Attention Motivated by [12], we employ the max-pooling and average-pooling operations, generating two spatial context descriptors, which are further forwarded into a multi-layer perceptron. The goal of user-aware channel-level attention is select the channel that effectively characterizes the user preferences. Given the user embedding u_i, the average-pooled descriptors $\mathbf{F}_{ave}^c$ and the max-pooled descriptors $\mathbf{F}_{max}^c$, we use a one-layer network to compute the attention score map, denoted as below:

$$\mathbf{C}_{u_i} = \sigma(\mathbf{W}_{1c}(\mathbf{W}_{0c}(\mathbf{F}_{avg}^c)) + \mathbf{W}_{1c}(\mathbf{W}_{0c}(F_{max}^c)) + \mathbf{W}_{cu}u_i), \tag{2}$$

where $\mathbf{W}_{1c}$ and $\mathbf{W}_{0c}$ indicate the first and second layer parameters. Here $\mathbf{W}_{cu}$ denotes the weight matrix of the user embedding, and σ denotes the sigmoid function.

Spatial-level Attention Compared to the channel-level attention, the spatial-level attention focuses on which region of the item the user pays attention to. Given an input fea-

ture map F, the average-pooled features across the channel $\mathbf{F}^s_{avg} \in \mathbb{R}^{1\times H\times W}$, the max-pooled features across the channel $F^s_{max} \in \mathbb{R}^{1\times H\times W}$, the user-aware spatial attention map is calculated as below:

$$\mathbf{S}_{u_i} = \sigma(\mathbf{f}^{s_k\times s_k}([\mathbf{F}^s_{avg}; \mathbf{F}^s_{max}]) + \mathbf{W}_{su}\mathbf{u}_i), \tag{3}$$

where $\mathbf{f}^{s_k\times s_k}$ represents a convolution filter with the size equal to $s_k \times s_k$ and σ denotes the sigmoid function. $\mathbf{W}_{su}$ denotes the weighting matrix of the user embedding branch. Then the multiplied attention feature map is forwarded into the item embedding sub-module to obtain the item vectors $(\mathbf{v}^t_j, \mathbf{v}^b_j, \mathbf{v}^s_j)$ and the hashing sub-module to obtain the binary codes, denoted as $(\mathbf{b}^t_j, \mathbf{b}^b_j, \mathbf{b}^s_j)$.

2.3. Preference Predictor

The preference predictor module aims to predict the preference score of a user on the candidate outfit. The score $y^j_{u_i}$ of the candidate outfit $O_j = (I^t_j, I^b_j, I^s_j)$ is calculated by a weighted sum of the general compatibility score g_j and the user-aware preference score $p^j_{u_i}$, which is denoted as below:

$$\begin{aligned} p^j_{u_i} &= \frac{1}{z_1}\sum_{m\in(t,b,s)} \mathbf{b}_{u_i}\mathbf{W}^{(\mathbf{u})}\mathbf{b}^m_j, \quad \text{用户对每个item的偏好} \\ g_j &= \frac{1}{z_2}\sum_{m\in(t,b,s)}\sum_{n\in(t,b,s),m\neq n} \mathbf{b}^m_j\mathbf{W}^{(\mathbf{i})}\mathbf{b}^n_j, \quad \text{每个item两两兼容分数} \\ y^j_{u_i} &= g_j + \alpha p^j_{u_i}, \end{aligned} \tag{4}$$

where $\mathbf{W}^{(\mathbf{u})}$ and $\mathbf{W}^{(\mathbf{i})}$ are the weighting matrixes for computing the general and personalized compatibility score, z_1 and z_2 are normalization parameters. Here α is utilized to balance the general and user-aware compatibility score.

The pairwise ranking loss with BPR [10] optimization criterion is employed to learn the parameters of the proposed PAN. The training set contains two kinds of triplets: 1) the triplet set S_1 with one user u_i and a pair of outfits O_j, O_k; 2) the triplet set S_2 with one outfit O_k and a pair of users u_m, u_n. S_1 and S_2 are denoted as below:

$$\begin{aligned} S_1 &= \{(u_i, O_j, O_k)|y^j_{u_i} > y^k_{u_i}\}, \\ S_2 &= \{(u_m, u_n, O_k)|y^k_{u_m} > y^k_{u_n}\}, \end{aligned} \tag{5}$$

where S_1 denotes the user u_i prefers O_j than O_k, and S_2 denotes the user u_m prefers outfit O_k than user u_n. Therefore, the ranking loss over set S_1 is denoted as below:

$$\ell_{BPR} = \sum_{S_1} \log\left(1 + \exp(-(y^j_{u_i} - y^k_{u_j}))\right), \tag{6}$$

and the proposed user-specific ranking loss over set S_2 is denoted as below:

$$\ell_u = \sum_{S_2} \log\left(1 + \exp(-(y^k_{u_m} - y^k_{u_n}))\right). \tag{7}$$

Moreover, the visual-semantic embedding loss [13] is utilized to enforce the constraint on the item's visual and textual representation. Here the item's textual embedding follows the same pipeline as that of the visual embedding. It aims to enlarge the similarity of matching visual-text pairs and pull away the distance of the non-matching text-visual pairs. The loss ℓ_{vse} is denoted as below:

$$\begin{aligned} \ell_{vse} = &\sum_{v,k} max\{0, d(\mathbf{v}, \mathbf{f_k}) - d(\mathbf{v}, \mathbf{f}) + m\} \\ &+ \sum_{f,k} max\{0, d(\mathbf{v}_k, \mathbf{f}) - d(\mathbf{v}, \mathbf{f}) + m\}, \end{aligned} \tag{8}$$

where $d(\mathbf{v}, \mathbf{f})$ indicates the distance between the matching visual-text pairs while $d(\mathbf{v}, \mathbf{f}_k)$ and $d(\mathbf{v}_k, \mathbf{f})$ represent the distances for the non-matching pairs. Here m is the margin parameter.

The overall objective loss L_{total} to optimize is described as below:

$$L_{total} = \ell_{BPR} + \lambda_1\ell_u + \lambda_2\ell_{vse}, \tag{9}$$

where λ_1 and λ_2 denote the weighting parameters for different losses.

User-aware Hard Negative Mining The hard negative mining scheme plays an important role in training a stable recommender system and the fast convergence [14] of the network. In the prior work of [9], the negative examples of outfits is randomly selected. To better differentiate the similar users, we develop an online user-aware hard negative mining strategy by calculating the Euclidean distance between the user embedding. Within a mini-batch, given a user u_i with the latent user embedding $\mathbf{b}_{u_i}$, the distance between u_i and the other users are computed and ranked in a ascending order. The items with the top K-ratio close distance are selected as the hard negatives.

3. EXPERIMENTS

3.1. Dataset and Evaluation Protocol

Most of the existing fashion datasets such as DeepFashion [15], Street2Shop [16], WoW [17], DeepShoe [18] cannot be directly used for our task for the lack of user profile. Polyvore-U is the most appropriate benchmark dataset for personalized fashion recommendation. It contains two sets of subsets: 1) Polyvore-630, which consists of about 150,000 outfits and 200,000 items; 2) Polyvore-519, which consists of about 98,000 outfits and 185,400 items.

Follwing Lu *et al.* following [9], the effectiveness of our proposed approach is evaluated on two tasks: 1) Normalized Discounted cumulative Gain (NDCG); 2) Area Under the ROC curve (AUC) and 3) fill-in-the-blank (FITB). It aims to choose an item among a list of candidate items to form the compatible outfit with the given items. The ground truth item

Table 1. Comparisons with state-of-the-art method on the Polyvore-630 and Polyvore-519 dataset.

Dataset	Approach	FITB	AUC	NDCG
Polyvore-630	SiameseNet [3]	0.5103	0.7703	0.6109
	Bi-LSTM [4]	0.5515	0.8102	0.6629
	CSN [2]	0.5536	0.8187	0.6744
	FHN [9]	0.6461	0.9176	0.8541
	Proposed	**0.6608**	**0.9311**	**0.8694**
Polyvore-519	SiameseNet [3]	0.5304	0.8026	0.6648
	Bi-LSTM [4]	0.5232	0.7746	0.6210
	CSN [2]	0.5617	0.8215	0.6703
	FHN [9]	0.6386	0.9137	0.8448
	Proposed	**0.6503**	**0.9242**	**0.8562**

Table 2. Evaluating the effect of user-specific loss (UL), user-aware hard negative mining (HN) and the attention network (AN). Note that the pre-trained resnet 50 feature map is utilized instead of updating the backbone network parameters as shown in Table 1.

Dataset	Approach	FITB	AUC	NDCG
Polyvore-630	None	0.6283	0.9109	0.8422
	UL	0.6399	0.9189	0.8528
	UL + HN	0.6429	0.9235	0.8578
	UL + HN + AN	**0.6458**	**0.9259**	**0.8602**
Polyvore-519	None	0.6188	0.9027	0.8253
	UL	0.6274	0.9103	0.8362
	UL + HN	0.6301	0.9139	0.8392
	UL + HN + AN	**0.6329**	**0.9153**	**0.8414**

is the correct answer and the performance is measured by accuracy of the answers.

3.2. Implementation details

In our experiments, AlexNet [19] is simply utilized as the backbone. The dimension for textual feature is 2400 and that for the visual feature is 4096. The type-aware embedding modules for items consist of two fc layers. The loss weight λ_1 and λ_2 are experimentally set as 0.1. The learning rate for the channel-wise and spatial-wise attention module is set to be 0.1. The ratio between the hard negative and randomly-chosen negative examples is set to be 0.5. Experimentally, the kernel size s_k is set as 7.

3.3. Experimental Results

We compare the performance of the proposed approach with the following baseline methods: 1) SiameseNet [3]; 2) Bi-LSTM [4]; 3) CSN [2]; 4) FHN [9]. From the experimental results conducted on Polyvore-630 and Polyvore-519 as shown in Table 1, it can be seen that our approach outperforms the other competitors by clear margins. Fig. 3 demonstrates the attentive feature maps for four exemplar users. For the shoes, user A puts more emphasis on the heel and the platform while user B focuses more on the toe. For the top, user C are merely interested in the collar design while user D demonstrates preference over both the collar and sleeve length. It experimentally verifies that different users demonstrate diverse outfit interests.

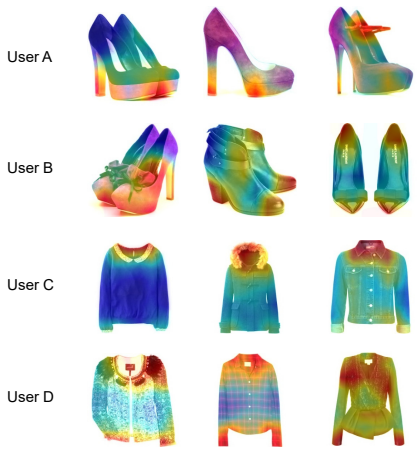

Fig. 3. Visualization of heatmaps for four exemplar users. Best view in color.

3.4. Effectiveness of Different Modules

We evaluated the effectiveness of the attention module, user-specific loss and hard negative mining during the model learning, as shown in Table 2. To greatly reduce the training time, we utilize the pre-trained resnet 50 feature map without updating the backbone network parameters. The proposed module is appended to the baseline framework step by step for evaluation. It can be seen that each module is essential in improving the overall performance of the network.

4. CONCLUSION

In this paper, we propose a novel attention-based personalized fashion recommendation system based on the user preferences. The attention map is incorporated at both the channel- and spatial-level. We also develop a novel user-specific ranking loss that is capable of capturing the different aspects of respective users. Moreover, we propose a simple but effective online user-aware hard negative mining approach. The experimental results on the personalized fashion recommendation dataset demonstrates the superiority of our proposed approach.

5. REFERENCES

[1] Elaine M Bettaney, Stephen R Hardwick, Odysseas Zisimopoulos, and Benjamin Paul Chamberlain, "Fashion outfit generation for e-commerce," *arXiv preprint arXiv:1904.00741*, 2019.

[2] Mariya I Vasileva, Bryan A Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth, "Learning type-aware embeddings for fashion compatibility," *arXiv preprint arXiv:1803.09196*, 2018.

[3] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie, "Learning visual clothing style with heterogeneous dyadic co-occurrences," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4642–4650.

[4] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis, "Learning fashion compatibility with bidirectional lstms," in *Proceedings of the ACM on Multimedia Conference*, 2017.

[5] Yuncheng Li, Liangliang Cao, Jiang Zhu, and Jiebo Luo, "Mining fashion outfit composition using an end-to-end deep learning approach on set data," *IEEE Transactions on Multimedia*, pp. 1–1.

[6] Zeyu Cui, Zekun Li, Shu Wu, Xiao-Yu Zhang, and Liang Wang, "Dressing as a whole: Outfit compatibility learning based on node-wise graph neural networks," in *The World Wide Web Conference*, 2019, pp. 307–317.

[7] Yuxin Peng, Jinwei Qi, and Yunkan Zhuo, "Mava: Multi-level adaptive visual-textual alignment by cross-media bi-attention mechanism," *IEEE Transactions on Image Processing*, vol. 29, pp. 2728–2741, 2019.

[8] Yuxin Peng, Jinwei Qi, and Yuxin Yuan, "Modality-specific cross-modal similarity measurement with recurrent attention network," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5585–5599, 2018.

[9] Zhi Lu, Yang Hu, Yunchao Jiang, Yan Chen, and Bing Zeng, "Learning binary code for personalized fashion recommendation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[10] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme, "Bpr: Bayesian personalized ranking from implicit feedback," *arXiv preprint arXiv:1205.2618*, 2012.

[11] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua, "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*, 2017, pp. 335–344.

[12] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[13] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis, "Automatic spatially-aware fashion concept discovery," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1463–1471.

[14] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[15] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[16] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg, "Where to buy it: Matching street clothing photos in online shops," in *Proceedings of International Conference on Computer Vision*, 2015, pp. 3343–3351.

[17] Si Liu, Jiashi Feng, Zheng Song, Tianzhu Zhang, Hanqing Lu, Changsheng Xu, and Shuicheng Yan, "Hi, magic closet, tell me what to wear!," in *Proceedings of the 20th ACM international conference on Multimedia*, 2012, pp. 619–628.

[18] Huijing Zhan, Boxin Shi, and Alex C Kot, "Cross-domain shoe retrieval with a semantic hierarchy of attribute classification network," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5867–5881, 2017.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.