A JOINT MULTI-TASK CNN FOR CROSS-AGE FACE RECOGNITION

Jinbiao Yu and Liping Jing

Beijing Key Lab of Traffic Data Analysis and Mining
Beijing Jiaotong University
{16120447, lpjing}@bjtu.edu.cn

ABSTRACT

Cross-age face recognition (CAFR) has received more and more attention in real applications, but it is a challenging task due to complex facial aging process. One popular way is modeling CAFR as a traditional face classification problem. However, most of them suffer from one main difficulty: how to effectively extract identity sensitive features that are age insensitive. In this paper, we propose a joint multi-task convolutional neural network (JMCNN) framework. JMCNN consists of two tasks: one for face recognition to learn identity sensitive features (i.e., age-invariant features), the other for age classification to learn age sensitive features, meanwhile, two tasks enhance each other by enforcing a regularization term on two kinds of features. The experimental results on two well-known cross-age datasets (Morph Album 2, CACD) have shown JMCNN is superior to the existing methods.

Index Terms— Cross-age face recognition, age classification, multi-task learning, age-invariant feature, convolutional neural network

1. INTRODUCTION

Cross-age face recognition as an emerging research field obtained more and more attention in academic and industry areas. For instance, it can be applied to finding missing children and identifying escaped criminals. However, it is a challenging task because aging process over time can substantially change facial appearance, as shown in Figure 1. The main difficulty is how to effectively extract identity sensitive features that are age insensitive.

In order to solve this problem, a surge of methods have been proposed and can be roughly divided to three categories. The first category aims to build generative model for synthesizing face images in different age ranges [1, 2, 3, 4, 5]. Even though such approaches, to some extent, compensate for huge intra-personal changes caused by aging, they have to depend on several parameters and cost a lot to train the model, which usually results in unstable performance.

An alternative is discriminative learning approach that aims to design face feature descriptor and use supervised learning algorithm to settle CAFR problem. Ling *et al.* [7] proposed a robust feature descriptor, gradient orientation pyramids (GOP) to represent cross-age faces, and adopted support vector machine (SVM) as recognition classifier. A densely sampled local feature description approach is given in [8] by integrating scale invariant feature transform and local binary pattern. Gong *et al.* [9] proposed a hidden factor analysis (HFA) model to represent each facial image as a linear combination of identity component and changing age component. Later, a robust local feature descriptor based on maximum entropy theory is given [10]. Du and Ling [11] use boosting idea to deemphasize age sensitive features and emphasize identity sensitive features during feature learning process for CAFR.

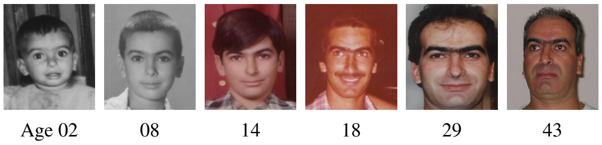

Fig. 1: Cross-age faces of one subject from FG-NET [6] dataset. The figure shows one subject's substantially change of facial appearance over time.

However, all above methods are based on the hand-crafted features, which limits their discrimination ability due to the separation of feature learning and recognition process. Recently, the neural network (NN)-based deep learning [12, 13, 14, 15] have been applied to solving CAFR problem. Among them, Wen *et al.* [14] combined base Convolutional NN (CNN) and the latent identity analysis (LIA) model. Xu *et al.* [12] integrated auto-encoder (AE) network and non-linear latent factor model to fit the complex aging process. Lin *et al.* [13] introduced the generalized similarity model in CNN architecture for CAFR. Zheng *et al.* [15] proposed a AE-CNN framework to combine face recognition and age classification tasks. Although these methods make use of deep NN, they usually focus on one learning task, face recognition or aging process, while ignore the correlation between these two tasks.

Therefore, in this paper, we propose a joint multi-task convolutional neural network (JMCNN) framework. It simultaneously models face recognition and age classification tasks by sharing a same CNN model and a regularization term, so

978-1-4799-7061-2/18/$31.00 ©2018 IEEE

that the interaction between identity sensitive features and age sensitive features are encouraged via the regularization loss.

The joint neural network can be efficiently trained via stochastic gradient descent and backpropagation algorithm. For testing phase, only identity sensitive features are used for face recognition. To the best of our knowledge, this is the first attempt to use such a regularization to enhance both tasks in a joint multi-task CNN architecture for CAFR.

The rest of the paper is organized as follows. The JMCNN model is proposed in Section 2. A series of experiments have been conducted and their results are given and discussed in Section 3. Finally, we briefly conclude this work and give the future work in Section 4.

2. JOINT MULTI-TASK CNN FRAMEWORK

In cross-age face recognition, the key issue is to extract the identity sensitive features that are age insensitive. In this section, we propose a joint multi-task convolutional neural network (JMCNN) framework, as shown in Figure 2. JMCNN consists of two tasks, one for identity recognition and the other for age classification. Both tasks share the CNN model, which accounts for basic face feature learning to obtain sharing feature pool $F1$ as indicated in the dashed box. Additional two fully connected layers follow the feature pool $F1$ for each task respectively. A regularization loss constraints both tasks in feature space to enhance each other.

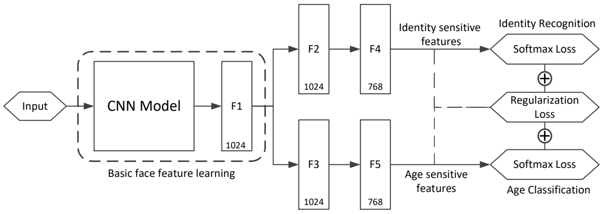

Fig. 2: The proposed joint multi-task convolutional neural network framework.

2.1. Model Formulation

To extract basic facial features, convolutional neural network is adopted due to it's powerful feature learning ability in computer vision field [16, 17, 18]. In order to avoid gradient vanishing and speed up training process, residual unit [19] and inception module [20] are introduced into CNN structure as shown in Figure 3, where the detailed information about parameters are listed in the box. More specifically, 3×3 / 2 denotes the filter size is 3×3 for convolutional or max pooling layers, where the stride is 2. (32) represents there are 32 filters. Given 3-channels RGB face images, three convolutional layers and one max pooling layer follow the inputs. Each max pooling layer is followed by a residual unit, then there are three residual units total. As suggested by [20], the inception module is adopted in the second unit (1×7 kernels followed by 7×1 kernels) and third residual unit (1×3 kernels followed by 3×1 kernels) rather than the first unit. The ReLU activation function is adopted in all convolutional layers. Finally, the basic facial features (denoted as $F1$) are obtained with size 1024.

For the proposed joint multi-task model, the dimension of output layer for each task is set to 768, and the hidden layer $F2$ and $F3$ also have same dimension 1024. In the identity recognition task, following the basic facial feature $F1$, the proposed JMCNN uses extra two fully connected layers ($F2$ and $F4$) to extract identity sensitive features, and evaluates output layer F_4 with a softmax loss. Given cross-age training dataset $\{x_i, y_i^I, y_i^A\}_{i=1}^M = \{X, Y^I, Y^A\}$, the loss function for identity recognition task is defined as:

$$L_I(X, Y^I) = -\frac{1}{M}\sum_{i=1}^{M} log \frac{e^{(W^I)^{\mathrm{T}}_{y_i^I} f_I(x_i) + b^I_{y_i^I}}}{\sum_{j=1}^{C} e^{(W^I)^{\mathrm{T}}_j f_I(x_i) + b^I_j}} \tag{1}$$

where x_i represents i-th face image, and y_i^I, y_i^A indicate the corresponding identity label and age label respectively. X, Y^I and Y^A represent set for x_i, y_i^I and y_i^A respectively. $f_I(x_i)$ is output embedding of network for identity recognition task. W^I and b^I are weights and bias of last layer for classification of this task. C is the number of subjects, and M is the number of training samples.

Similarly, for age classification task, JMCNN uses extra two layers ($F3$ and $F5$) for age sensitive features learning. The loss function of this task is defined as:

$$L_A(X, Y^A) = -\frac{1}{M}\sum_{i=1}^{M} log \frac{e^{(W^A)^{\mathrm{T}}_{y_i^A} f_A(x_i) + b^A_{y_i^A}}}{\sum_{j=1}^{G} e^{(W^A)^{\mathrm{T}}_j f_A(x_i) + b^A_j}} \tag{2}$$

Where G is the number of age groups which can be obtained by dividing the age range according grow processing of human beings. The superscript A of W, b and subscript of f donate symbol for age classification task.

It can be seen that identity recognition and age classification are two traditional classification tasks. Both tasks share the same feature pool ($F1$) obtained by the CNN model, while identity recognition task pursues identity sensitive features and age classification task purses age sensitive features. For feature competition, the features sensitive to identity recognition task that should be insensitive to age classification task, and the features sensitive to age classification task should be insensitive to identity recognition task. Such correlations suggest that the relationship of conflict features for both tasks should be negative correlation. We define $f_I(x_i)$ and $f_A(x_i)$ are the features of i-th face image for two tasks respectively, then a cosine similarity regularizer can be introduced over $f_I(x_i)$ and $f_A(x_i)$ (as shown in Eq.(3)) to encode feature conflict between these two tasks.

$$L_R(f_I, f_A) = \frac{f_I(x_i)^{\mathrm{T}} f_A(x_i)}{\|f_I(x_i)\|_2 \|f_A(x_i)\|_2} \tag{3}$$

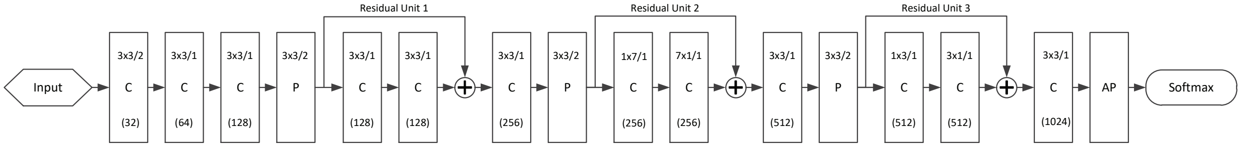

Fig. 3: The CNN architecture for basic face feature learning. C donates convolution layer with a ReLU activation function, P donates max pooling layer, while AP donates average pooling layer. Three residual units are used after each max pooling layer. Then through the final convolution layer followed by a average pooling layer the output embedding of network is fed to softmax classifier.

By combining (1), (2) and (3), we can define the regularized joint loss for the proposed JMCNN model as follows:

$$L(X, Y^I, Y^A) = L_I(X, Y^I) + L_A(X, Y^A) + \lambda L_R(f_I, f_A) \quad (4)$$

where λ is the regularization parameter to balance the softmax loss and the regularizer.

2.2. Optimization

The optimization for the model is implemented by stochastic gradient descent (SDG) and standard backpropagation algorithm. For the backward propagation, the derivative of L with respect to f_I and f_A need to be calculated. We use δ_I and δ_A indicate backpropagation errors of final loss layer for each task [21]. Different with other layers, the gradient of L with respect to f_I can be calculated by the chain rule as follows:

$$\begin{aligned}\frac{\partial L}{\partial f_I} =& \frac{\partial L_I}{\partial f_I} + \lambda \frac{\partial L_R}{\partial f_I} \\ =& (W^I)^T \delta_I + \lambda \Big(\frac{f_A}{\|f_I\|_2 \|f_A\|_2} - \frac{(f_I^{\mathrm{T}} f_A) f_I}{\|f_I\|_2^3 \|f_A\|_2} \Big) \quad (5)\end{aligned}$$

Similarly, the gradient of L with respect to f_A can be calculated via:

$$\begin{aligned}\frac{\partial L}{\partial f_A} =& \frac{\partial L_A}{\partial f_A} + \lambda \frac{\partial L_R}{\partial f_A} \\ =& (W^A)^T \delta_A + \lambda \Big(\frac{f_I}{\|f_I\|_2 \|f_A\|_2} - \frac{(f_I^{\mathrm{T}} f_A) f_A}{\|f_I\|_2 \|f_A\|_2^3} \Big) \quad (6)\end{aligned}$$

Adam optimization algorithm is adopted for the parameters updating [22].

3. EXPERIMENT

3.1. Experiments Details

Dataset. We conducted experiments on two widely used face aging datasets: Morph Album 2 [23] and CACD [24]. The dataset CASIA-WebFace [25] is used to train the CNN model. Two experimental datasets fine-tune the CNN model and learn the subsequent fully connected layers with the proposed model (4). For age classification, according to the grow processing of human beings, the ages are partitioned into 9 groups $\{15\sim18,\ 19\sim21,\ 22\sim26,\ 27\sim31,\ 32\sim36,\ 37\sim41,\ 42\sim47,\ 48\sim54,\ 55\sim77\}$, i.e., $G = 9$ in (2).

Preprocessing. All face images are detected by the algorithm [26], and five landmarks (tow eyes, nose and mouth corners) are used for face alignment. The faces are cropped to 128x128 RGB images. Each pixel of the RGB image is normalized via subtracted by 127.5 and divided by 128.

Detailed settings. The proposed JMCNN model is implemented by TensorFlow [27]. In mini-batch gradient descent processing, the batch size is set to 128, and the learning rate begins with 0.1 and adaptively updated by adam optimizer.

3.2. Experiments Results and Discussion

Experiments on Morph. Morph Album 2 contains more than 55,000 images of 13000 individuals with age ranging from 17 to 77, which has large inter-personal age gap while small intra-personal age gap. Follow the testing scheme in [13], the dataset is divided into two parts. One part incudes 10000 individuals which are used to fine-tune the proposed JMCNN model. The remaining 3000 individuals are used for evaluating. Note that there is no overlapping subject between these two parts. For each subject when evaluating, the youngest age face image is selected as training point while the oldest age face image as testing point. The recognition result is evaluated with Rank-1 identification rate.

Our first experiment is conducted to analysis the effect of parameter λ on JMCNN model (4). The results are shown in Figure 4 where λ is tuned in {0, 1e-3, 0.01, 0.1, 1}. It can be seen that JMCNN performs better and better with the increasing of λ, until it achieves best result when $\lambda = 0.1$, and

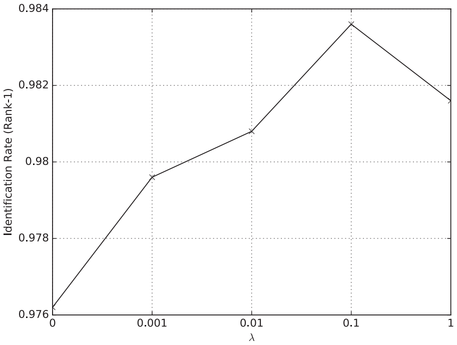

Fig. 4: Recognition rates of different λ on Morph 2.

then its performance is degraded. The result is reasonable because the regularization term does not work if λ is too small, which may lead to a big overlapping between identity sensitive features and age sensitive features. Meanwhile, a large λ makes learning process focus on the regularization term and weakens the performance of separate learning task.

Furthermore, we compare the proposed method (JMCNN) with (1) the CNN baseline model as Figure 3 describes (2) the CNN baseline fine-tuned by Morph dataset (3) several state-of-the-art approaches on CAFR problem. From the results shown in Table 1, we have following conclusions.

Method	Rank-1 Identification Rate
HFA (2013) [9]	91.14%
CARC (2014) [24]	92.80%
MEFA (2015) [10]	93.80%
MEFA+SIFT+MLBP (2015) [10]	94.59%
GMS (2017) [13]	94.40%
LF-CNN (2016) [14]	97.51%
AE-CNN (2017) [15]	98.13%
CNN baseline (trained by CASIA data)	91.96%
CNN baseline (fine-tuned by Morph)	97.23%
JMCNN (fine-tuned by Morph)	**98.36%**

Table 1: Recognition rates of different methods on Morph 2 dataset.

Firstly, the CNN baseline model only has the result of 91.96%, which is inferior to most of other results in the table. It shows that it's not appropriate to utilize basic CNN model directly on CAFR problem, and the basic facial features obtained by the CNN model still contain the age information. This result confirms that it is desirable to separate age information from the basic facial features. Secondly, the CNN baseline fine-tuned by Morph dataset can achieve a high performance (97.23%), which outperforms other traditional methods with a clear margin, but it's still inferior to the existing LF-CNN and the proposed method. It indicates that fine-tune operation fits the data distribution well but there is still a promotion room. Finally, the proposed method can achieve new state-of-the art (98.36%). This results further show the effectiveness of the proposed method. And the regularization term is effective for JMCNN to learn much robust age-invariant features.

Experiments on CACD. CACD dataset is a recently largest cross-age face dataset, which contains 2000 celebrities with age ranging from 16 to 62. Following the experimental setting in [24], 1880 celebrities are used to fine-tune the JMCNN model. The left 120 celebrities are used for testing. Among them, images taken at 2013 are as query images, the remaining images taken at 2004-2006, 2007-2009 and 2010-2012 are separated into three groups as database images respectively.

Mean average precision (MAP) is used for evaluation metric. To be more specific, let $q_i \in Q$ be the query image in query set Q, and the positive image set corresponding to q_i in the database are denoted as $\{I_1, I_2, ..., I_{m_i}\}$. We define R_{ik} as retrieval results of q_i in descending order from the top image to the k-th image. The whole MAP of Q can be calculated as follows:

$$MAP(Q) = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{m_i}\sum_{k=1}^{m_i} Precision(R_{ik}) \qquad (7)$$

Where $Precision(R_{ik})$ is the radio of positive images in R_{ik}.

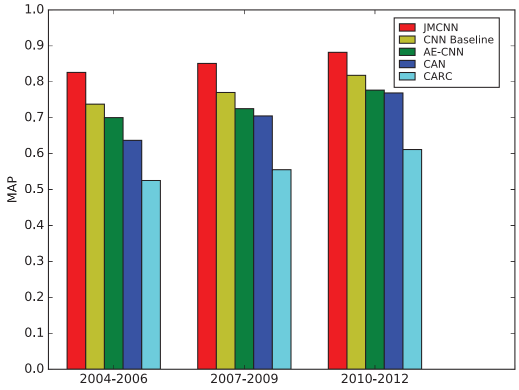

Fig. 5: The retrieval results on CACD of different methods.

We compare JMCNN with several existing methods including CARC [24], CAN [12], AE-CNN [15] and the CNN baseline model. Figure 5 reports the comparison results. It can be seen that the CNN baseline outperforms the existing methods, while our method still has an obviously performance boosting. The main reason is that both inception module and residual unit are adopted in the CNN model. Meanwhile, age-invariant features benefit from the regularization term by considering the conflict between identity sensitive features and age sensitive features.

4. CONCLUSIONS

In this paper, we proposed a joint multi-task convolutional neural network for cross-age face recognition. Unlike many existing deep learning methods, the proposed method simultaneously learn identity sensitive features and age sensitive features to obtain robust age-invariant features. Experiments on two public cross-age datasets have shown the proposed JMCNN is superior to the state-of-the-art. In the feature, we will consider facial pose or emotion information to design much robust cross-age face recognition system.

5. REFERENCES

[1] U. Park, Y. Tong, and A. K. Jain, "Age-invariant face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 947–954, 2010.

[2] J. Du, C. Zhai, and Y. Ye, "Face aging simulation based on nmf algorithm with sparseness constraints," in *International Conference on Intelligent Computing*. Springer, 2011, pp. 516–522.

[3] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 442–455, 2002.

[4] C. N. Duong, K. G. Quach, K. Luu, T. H. N. Le, and M. Savvides, "Temporal non-volume preserving approach to facial age-progression and age-invariant face recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3755–3763.

[5] X. Shu, J. Tang, H. Lai, L. Liu, and S. Yan, "Personalized age progression with aging dictionary," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3970–3978.

[6] FG-NET, "Fg-net aging database," (`http://www-prima.inrialpes.fr/FGnet/html/home.html`).

[7] H. Ling, S. Soatto, N. Ramanathan, and D. W. Jacobs, "Face verification across age progression using discriminative methods," *IEEE Transactions on Information Forensics and Security*, vol. 5, pp. 82–91, 2010.

[8] Z. Li, U. Park, and A. K. Jain, "A discriminative model for age invariant face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 6, pp. 1028–1037, 2011.

[9] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang, "Hidden factor analysis for age invariant face recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2872–2879.

[10] D. Gong, Z. Li, D. Tao, J. Liu, and X. Li, "A maximum entropy feature descriptor for age invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5289–5297.

[11] L. Du and H. Ling, "Cross-age face verification by coordinating with cross-face age verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2329–2338.

[12] C. Xu, Q. Liu, and M. Ye, "Age invariant face recognition and retrieval by coupled auto-encoder networks," *Neurocomputing*, vol. 222, pp. 62–71, 2017.

[13] L. Lin, G. Wang, W. Zuo, X. Feng, and L. Zhang, "Cross-domain visual matching via generalized similarity measure and feature learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1089–1102, 2017.

[14] Y. Wen, Z. Li, and Y. Qiao, "Latent factor guided convolutional neural networks for age-invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4893–4901.

[15] T. Zheng, W. Deng, and J. Hu, "Age estimation guided convolutional neural network for age-invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1–9.

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

[17] G. Chen, T. X. Han, Z. He, R. Kays, and T. Forrester, "Deep convolutional neural network based species recognition for wild animal monitoring," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 858–862.

[18] T. Zhi, L. Duan, Y. Wang, and T. Huang, "Two-stage pooling of deep convolutional features for image retrieval," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2465–2469.

[19] 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*, 2016, pp. 770–778.

[20] 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*, 2016, pp. 2818–2826.

[21] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533, 1986.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*. IEEE, 2006, pp. 341–345.

[24] B. Chen, C. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *European Conference on Computer Vision*. Springer, 2014, pp. 768–783.

[25] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[26] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, pp. 1499–1503, 2016.

[27] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.