УДК 004.855.5:004.93'1

PRACTICAL ASPECTS OF DEEP LEARNING FOR PERSON IDENTIFICATION

***N.E. Rogachev*[1]**
Systems analyst at HiQo Solutions, PhD student at the BSU

***Y.Y. Ramanouski* [2]**
Systems analyst at HiQo Solutions

***N.V. Guletskii*[3]**
Systems analyst at HiQo Solutions, master's student at the BSU

[1]*HiQo Solutions, Ltd., Republic of Belarus*
E-mail: nikolai.rogachev@hiqo-solutions.com
[2]*HiQo Solutions, Ltd., Republic of Belarus*
E-mail: yury.romanovskiy@hiqo-solutions.com
[3]*HiQo Solutions, Ltd., Republic of Belarus*
E-mail: nikolay.guletskiy@hiqo-solutions.com

N.E. Rogachev
Graduated from the Belarusian State University. Master of science. PhD student at the BSU. Works as a systems analyst at HiQo Solutions, Ltd, Republic of Belarus, Minsk. Research interests are artificial neural networks, computer vision and ethics of machine learning.

Y.Y. Ramanouski
Graduated from the Belarusian State University. Master of science. Works as a systems analyst at HiQo Solutions, Ltd, Republic of Belarus, Minsk. Fields of research are data analysis, artificial neural networks, time series forecasting.

N.V. Guletskii
Graduated from the Belarusian State University. Master's student at the BSU. Works as a systems analyst at HiQo Solutions, Ltd, Republic of Belarus, Minsk. Fields of research include functional analysis, abstract harmonic analysis, optimization, artificial neural networks and computer vision.

Abstract. This paper focuses on the peculiarities of implementing and training fast and accurate models for person identification. During our research, we have evaluated numerous datasets, deep neural network architectures, loss functions and training techniques both for face- and body-based identification.

Keywords: deep learning, supervised learning, person identification.

Person identification in images and video (e.g., in surveillance footage) has applications in public security, access control and human resource management, among others.

First attempts to formalize the problem date back to the notion of soft biometrics. This concept, along with the term itself, appeared in the early 2000s to describe an approach when biometrical features (ethnicity, gender, age, height, body complexity, hair color, eye color, etc.) are used as an input for such methods as fuzzy logic, machine learning and evolutionary computing [1]. The main idea was that these features, although not informative enough when used one at a time, become suitable for identification and similar tasks when used in groups.

Biometric traits may be characterized by different parameters, such as nature of value and distinctiveness (see Table 1). One may notice that facial traits are considered the most effective for human identification. However, in cases of poor filming conditions or face occlusion other soft biometric traits have also proven to be useful [2], so the first experiments were using low-resolution footage for efficiency and were body-based.

Table 1

Characteristics of soft biometric traits [3, Table 1]

Soft biometric trait	**Face/body/ accessory**	**Nature of value**	**Permanence**	**Distinctiveness**	**Subjective perception**
Skin color	Face	Continuous	Medium	Low	Medium
Hair color	Face	Continuous	Medium	Medium	Medium
Eye color	Face	Continuous	High	Medium	Medium
Beard	Face	Binary	Low/medium	Low	Medium
Moustache	Face	Binary	Low/medium	Low	Medium
Facial measurements	Face	Continuous	High	Medium	Medium/high
Facial shapes	Face	Discrete	High	High	High
Facial feature measurements	Face	Continuous	High	High	Medium/high
Facial feature shapes	Face	Discrete	High	High	High
Make-up	Face	Discrete	Low	Low	Medium
Ethnicity	Face	Discrete	High	Medium	Medium
Marks	Face/body	Discrete	High	Medium/high	Low
Gender	Face/body	Binary	High	Low	Low
Age	Face/body	Continuous	Low/medium	Medium	Medium
Height	Body	Continuous	Medium/high	Medium	Medium
Weight	Body	Continuous	Medium	Medium	Medium
Gait	Body	Continuous	Medium	Medium	High
Body measurements	Body	Continuous	Medium/high	Medium/high	Medium
Body shapes	Body	Discrete	Medium	Medium	Medium
Clothes color	Accessory	Discrete	Low	Medium	Medium

Many approaches have been proposed to establish precise identification. The ones considered state-of-the-art make use of convolutional neural networks (CNNs) and so-called ID-Discriminative Embeddings (IDE) [4]. The main idea is to learn such mapping of the images that will help distinguish between the images of different people.

One of the examples how such mapping may be learned is to use Triplet Loss training, introduced in [5]. In this framework, the features of the image are extracted using a deep convolutional neural network, producing a multidimensional representation of the input. This representation is then L_2-normalized to obtain an embedding of the image into the hypersphere (Figure 1).

Figure 1. Triplet Loss framework structure [5, Figure 2]

The desired property of the embedding subspace — that any image x_i^a of a specific person (*anchor*) should be closer to all other images x_i^p of the same person (*positive examples*) than it is to any image x_i^n of any different person (*negative example*) (Figure 2) — is expressed in a form of a Triplet Loss function

$$L = \sum_{\mathrm{T}} h\Big[D\big(f\left(x_i^a\right), f\left(x_i^p\right)\big) - D\big(f\left(x_i^a\right), f\left(x_i^n\right)\big)\Big],$$

where T is the set of all valid triplets in the training set, $f\left(x\right)$ is the embedding of x, $D\left(f_1, f_2\right)$ is a distance measure (squared Euclidean distance $\left\| f_1 - f_2 \right\|_2^2$ was used in [5]), $h(d) = \left[d + \alpha \right]_+$ — a hinge function with classification margin α. This loss function is then minimized during training. It is a straightforward way of achieving desired behavior [6], and yields impressive results.

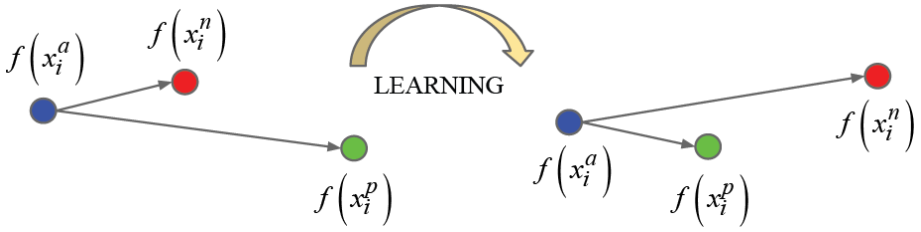

Figure 2. Rearrangement of triplets during Triplet Loss training [5, Figure 3]

The Triplet Loss framework has the following advantages:

- feature extractor is essentially treated as a black box [5] and may be represented by the widespread architectures for image classification, which enables transfer learning [7] and, therefore, significantly speeds up training – the reused architectures are frequently called backbones in this context;
- training is performed end-to-end, which simplifies engineering efforts and makes the loss itself a direct – although limited by generalization error – estimate of the model performance;
- similarity of images is represented by Euclidean distance of their embeddings, which enables usage of distance-based classification methods (e.g., the *neighbors* module of [8]);
- network does not learn the rules for closed-set classification, but performs metric learning for the embedding space, allowing new identities to be introduced later in development and production.

First evaluations of the model have shown that the performance in practice depends heavily on the consistent clothing and stable lighting conditions. Another factor of performance may be the stability of bounding boxes: manually labeled ones are consistent and accurate, while automatically labeled ones may either have wide margins around the body or leave out some parts of the body. One of the solutions is just to have more data. Indeed, collecting more training examples has proven to increase classification accuracy. To increase it further, we applied color, crop and noise augmentations.

The color augmentation is meant to help generalize for the cases when the images have unusual colors in them, i.e. too warm at sunrise or dark, cold and low-saturated in the evening with artificial lighting in place. It is performed by applying 'gradient maps' (provided in [9]) to a given

image, then performing soft light blending operation over the original image and the mapped image, after that the weighted sum of the blended image and the original image is calculated (Figure 3).

Figure 3. Color augmentation

Crop augmentation helps to deal with situations when detected bounding box is not accurate, resulting in body fragments being left outside or bodies taking only the part of the image. It is performed by cropping a fragment of an original image, with coordinates of the cropped drawn from the uniform distribution.

Noise augmentation is implemented using additive Gaussian noise (Figure 4). When distributed this way, the noise acts as a random corruption of the signal and thus is a natural choice for regularization. Parameters are the standard deviation of the noise and whether monochromatic noise (i.e. the same noise value for all three channels) should be used.

Figure 4. Noise augmentation

Another matter that has been addressed is low contrast. One of the simple yet effective approaches to correct that is intensity rescaling. It takes the intensity spectrum present in an image and rescales it so that it takes the whole range of intensity values. It helps to yield details from particularly dark, low-contrast images, which are common for the experimental environment (Figure 5). The implementation of intensity rescaling has been ported from [10].

Our experiments heavily employed results obtained in [11]. In particular, we switched to the non-squared Euclidean distance $D(f_1, f_2) = \|f_1 - f_2\|_2$ and the smooth approximation of the hinge function called softplus function: $h(d) = \ln(1 + \exp d)$.

We also solved the triplet mining problem as proposed in [11]. It is easy to see that the number of triplets grows cubically with the number of images, with the majority of them being trivial. However, if we sample P identities and K images for each identity and then find the hardest positive and the hardest negative samples for each of the PK anchors, then we have moderately hard triplets, as they are the hardest within a small subset of the data. This batch loss is also valid in a sense that it approximates L in the limit of infinite training.

Figure 5. Intensity rescaling

For the body-based identification, we initially evaluated ResNet-V1-50 [12] for backbone and later switched to MobileNet-V1-1.0-224 [13] for the sake of inference speed. Both backbones were initialized with pretrained weights from [14]. Training dataset was combined from the custom labeled data and Market-1501 dataset Net input size was 256×128 pixels. Our observations on hyperparameter tuning are the following:

- cutting the dimensionality of embeddings in half from the default value of 128 has shown no significant drops in accuracy, which should be kept in mind for the later step of classification in the high-dimensional space;
- image augmentations have a relatively low, but steady positive effect on the model performance;
- L_2 and Dropout [16] regularizations appeared to have no effect on the model accuracy (which was expected as per [13]);
- in accordance with the recent suggestions [17] that Adam optimization method [18] finds substantially different optima than Stochastic Gradient Descent with momentum [19] which arguably generalize worse, we tried switching from the default Adam method to SGD with momentum and obtained performance that is at least as good as when using Adam.

In general, hyperparameter tuning is far from having dramatic impact. In fact, the values suggested in [11] provide a good reference to where the search should start and yield the model that doesn't lag very much behind the best model we were able to train. Moreover, the effect that hyperparameter tuning has is comparable in magnitude to the variance that is caused by the mere randomness of the training process.

As a result, the model was able to learn semantically meaningful discriminative features such as clothes color and body complexity (Figure 6).

Face-based identification has been introduced by switching to high-resolution images and then following essentially the same path. The only certain differences are the square 128×128 shape of the network input for face images and the dataset – for faces MS-Celeb-1M [20] was chosen.

Figure 6. Barnes-Hut t-SNE [21] visualization of the test set

Resulting models may be used separately or in combination, with different strategies of classification in place. The accuracy of the combined model should be enough, for example, for matters of human resource management, but there are some concerns that one needs to address in order to achieve higher-grade performance for more critical tasks:

- faces regularly appear blurred, which complicates both detection and identification;
- lighting conditions adversely influence performance – especially at sunrise, at sunset and with artificial lighting in place (this is partly addressed by augmentation; experience with both this point and the previous one may vary with different camera equipment and environment characteristics);
- several recent papers [22–24] have suggested switching from Euclidean to angular margins of different forms in loss functions, which is justified both by geometrical intuition and by empirical performance reported by respective authors – one should probably evaluate these margins for his model.

References

[1]. Franke K., Ruiz-del-Solar J. Soft-Biometrics: Soft-Computing Technologies for Biometric-Applications // Advances in Soft Computing — AFSS 2002 / ed. Pal N.R., Sugeno M. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002. Vol. 2275. P. 171–177.

[2]. Rice A., Phillips P.J., O'Toole A. The Role of the Face and Body in Unfamiliar Person Identification: Face and body in person identification // Applied Cognitive Psychology. 2013. Vol. 27, № 6. P. 761–768.

[3]. Dantcheva A. et al. Bag of soft biometrics for person identification: New trends and challenges // Multimedia Tools and Applications. 2011. Vol. 51, № 2. P. 739–777.

[4]. Zheng L. et al. Person Re-identification in the Wild // arXiv:1604.02531 [cs]. 2016.

[5]. Schroff F., Kalenichenko D., Philbin J. FaceNet: A Unified Embedding for Face Recognition and Clustering // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015. P. 815–823.

[6]. Weinberger K.Q., Blitzer J., Saul L.K. Distance metric learning for large margin nearest neighbor classification // Advances in neural information processing systems. 2006. P. 1473–1480.

[7]. Pratt L.Y., Mostow J., Kamm C.A. Direct Transfer of Learned Information Among Neural Networks // Proceedings of the Ninth National Conference on Artificial Intelligence. 1991. P. 584–589.

[8]. Pedregosa F. et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2011. Vol. 12. P. 2825–2830.

[9]. Hunter J.D. Matplotlib: A 2D Graphics Environment // Computing in Science & Engineering. 2007. Vol. 9, № 3. P. 90–95.

[10]. Van der Walt S. et al. scikit-image: image processing in Python // PeerJ. 2014. Vol. 2. P. e453.

[11]. Hermans A., Beyer L., Leibe B. In Defense of the Triplet Loss for Person Re-Identification // arXiv:1703.07737 [cs]. 2017.

[12]. He K. et al. Deep Residual Learning for Image Recognition // arXiv:1512.03385 [cs]. 2015.

[13]. Howard A.G. et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications // arXiv:1704.04861 [cs]. 2017.

[14]. Silberman N., Guadarrama S. TensorFlow-Slim image classification model library [Electronic resource]. 2016. URL: https://github.com/tensorflow/models/tree/master/research/slim.

[15]. Zheng L. et al. Scalable Person Re-identification: A Benchmark // Computer Vision, IEEE International Conference on. 2015.

[16]. Srivastava N. et al. Dropout: a simple way to prevent neural networks from overfitting // The Journal of Machine Learning Research. 2014. Vol. 15, № 1. P. 1929–1958.

[17]. Wilson A.C. et al. The Marginal Value of Adaptive Gradient Methods in Machine Learning // arXiv:1705.08292 [cs, stat]. 2017.

[18]. Kingma D.P., Ba J. Adam: A Method for Stochastic Optimization // arXiv:1412.6980 [cs]. 2014.

[19]. Rumelhart D.E., Hinton G.E., Williams R.J. Learning representations by back-propagating errors // Nature. 1986. Vol. 323, № 6088. P. 533–536.

[20]. Guo Y. et al. MS-Celeb-1M: A Dataset and Benchmark for Large Scale Face Recognition // European Conference on Computer Vision. Springer, 2016.

[21]. Van Der Maaten L. Accelerating t-SNE using tree-based algorithms // The Journal of Machine Learning Research. 2014. Vol. 15, № 1. P. 3221–3245.

[22]. Liu W. et al. SphereFace: Deep Hypersphere Embedding for Face Recognition // arXiv:1704.08063 [cs]. 2017.

[23]. Wang H. et al. CosFace: Large Margin Cosine Loss for Deep Face Recognition // arXiv:1801.09414 [cs]. 2018.

[24]. Deng J. et al. ArcFace: Additive Angular Margin Loss for Deep Face Recognition // arXiv:1801.07698 [cs]. 2018.

ПРАКТИЧЕСКИЕ АСПЕКТЫ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ ИДЕНТИФИКАЦИИ ЛЮДЕЙ

Н.Е. РОГАЧЁВ
Системный аналитик, СООО «ХайКво Солюшенс»

Ю.Ю. РОМАНОВСКИЙ
Системный аналитик, СООО «ХайКво Солюшенс»

Н.В. ГУЛЕЦКИЙ
Системный аналитик, СООО «ХайКво Солюшенс»

СООО «ХайКво Солюшенс», Республика Беларусь, Минск
E-mail: nikolai.rogachev@hiqo-solutions.com
СООО «ХайКво Солюшенс», Республика Беларусь, Минск
E-mail: yury.romanovskiy@hiqo-solutions.com
СООО «ХайКво Солюшенс», Республика Беларусь, Минск
E-mail: nikolay.guletskiy@hiqo-solutions.com

Аннотация. В данной работе изучаются особенности разработки и обучения быстрых и точных моделей распознавания людей. В рамках исследования нами были рассмотрены многочисленные массивы данных, глубокие нейросетевые архитектуры, целевые функции и методики обучения для идентификации как по изображению лица, так и по изображению всего тела.

Ключевые слова: глубокое обучение, обучение с учителем, идентификация людей.