Image reconstruction from ResNet semantic feature vector

Vít LISTÍK[1]

[1]Dept. of Cybernetics, Faulty of Electrical Engineering,
Czech Technical University in Prague, Technická 2, 166 27 Praha, Czech Republic

listivit@fel.cvut.cz

Abstract. *Image classification tasks are successfully solved with convolutional neural networks. We want to prove that it is possible to use semantic feature extraction as anonymization mechanism. Semantic feature vectors may be extracted with the pre-trained convolutional neural network trained for image classification like ResNet. Those vectors may be successfully used for different classification tasks, therefore we propose that those vectors may be published as a dataset for tasks containing personal data.*

We want to verify if it is not possible to reconstruct the original image from the semantic feature vector. We have designed a convolutional neural network architecture based on DCGAN to generate the images. First, we trained the network to reconstruct training images. After completing this task we tried if the network is able to generalize the reconstruction to unseen images.

We came to the conclusion that the reconstruction of training data is possible. Therefore we think that the method is correct, but the generalization for unseen samples is computationally very difficult but possible. The reconstruction is not able to reconstruct personal information because it is only able to reconstruct main concepts of the image and previously seen details. Based on this conclusion we assume that it is possible to publish semantic feature vectors for datasets containing personal data.

Keywords

Image reconstruction, Generative adversarial networks (GAN), autoencoder, Convolutional Neural Network (CNN), Neural Network (NN)

1. Introduction

In this work, we want to prove that although it is possible to reconstruct the image from the semantic feature vector, it is not possible to reconstruct the details of the image. When we talk about semantic feature vector we have in mind the representation of the image extracted from the output of last convolutional layer of a deep convolutional neural network (CNN).

We are trying to prove that the exact reconstruction is not possible because we want to use the semantic feature vector as anonymization mechanism. For some tasks (like image spam classification) it is very hard to publish the dataset because of its private nature [2]. Therefore in some areas exists only a few publicly available datasets. Which leads to slower progress in those areas. We propose that the image data may be published after the anonymization based on semantic feature vector extraction, which would significantly help the research in those areas.

Convolutional Neural networks have been successfully used for image processing tasks lately [8]. CNNs are capable of compressing the information from multi-dimensional data like images. Thanks to this property and raising computational power the state of the art results for image classification and object detection changed [12].

One of the examples of those deep CNNs is ResNet [7]. This network architecture won the ImageNet classification challenge in 2015 [13]. The architecture is using residual connections which made training of network this deep possible. Architectures trained on huge datasets like this one are often used for a technique called transfer learning [10]. This technique is using the output of last convolutional layer of the pre-trained network and using it for a different task. The output is often called semantic feature vector. The feature vector contains some semantic information thanks to the task definition, the image classification, which forces the convolutional filters to extract image properties useful for the classification. This approach is useful for tasks with insufficient data for training the CNN from scratch.

Autoencoder is a neural network (NN) architecture using convolutions to compress the information contained in the image to much a smaller feature vector, which is more efficient for storage but mainly the dense representation is better for training [5] machine learning models. The architecture has the same input and output dimensions and bottleneck part in the middle. The architecture is trained via backpropagation with raw pixels used as input and output. The architecture should be able to reconstruct the image with minimal loss. The bottleneck layer may be used for other tasks similar to the ResNet feature vector.

Generative adversarial networks (GAN) may be used for generating images from latent vectors [6]. Those vec-

tors are often random but may also encode some information about the generated image. GAN architectures are composed of two NN. First NN is called generator and the second one is called discriminator. The Generator is supposed to generate realistic images and the discriminator is supposed to distinguish generated images from the real ones. The architecture is trained simultaneously.

2. Methods

The task is to generate an image from the semantic feature vector which will be very similar to the original image. Basically, the task is the same as for autoencoder with a difference of using pre-trained CNN. That means that the model is not trained as a whole like in autoencoder but only the decoder is trained in this work.

When the architecture will be able to reconstruct the images we will try to use it to generalize to unknown samples. This will simulate the situation with the published dataset. When personal data would be published in this form, the person trying to expose it would have to do the same process, train the generator network on the publicly available images and then use the trained generator network to generate private images. We want to show how successful will this process be.

The generation task may be divided into two parts: Feature extraction and image generation. The feature extraction is forward pass through the ResNet v1 pre-trained network. The network has defined image pre-processing based on the training, like the input size 224 x 224 px. The output of the last convolutional layer is a real-valued vector of size 2048.

The generation network is inspired by DCGAN [11]. The architecture is described in section 2.2.

2.1. Dataset

The dataset used for this task is called Imagenet [4]. We used only the images, not the labels. The original dataset consists of 14M labeled images. We are using random subsets of the dataset.

2.2. Model

The input ResNet vector is reshaped to 4x4x128 so the convolutions may be used. Then we apply the convolutions with the number of filters 2048, 1024, 512, 256, 128 and 64 followed by upsampling layers. The upsampling layers are doubling the size of the image by copying the rows and columns. This architecture scales the image up to 256 x 256 px. The convolutional filters have size 3 x 3. We have also tried 4 x 4, with no significant improvement in the generated image, but bigger filters are slower to train. Output layer consists of 3 filters 2 x 2 because we want to generate 3-channel (RGB) images. Padding for all the convolutions is set to same. We are using leaky ReLu activation [9] for the hidden convolutions and sigmoid activation for the output layer. We have used Adadelta optimizer for the training [14].

3. Experimental Results

Our solution is published as open-source code [1]. It is using Keras with Tensorflow implementation [1, 3].

Output image size corresponds to the network architecture. We wanted to reconstruct the image in the similar resolution which is used as the input of the network. It is straightforward to use powers of 2, therefore, we used 256 x 256 px. We have also tested the architecture on smaller images in first experiments to validate the idea, because of training speeds.

We tested the impact of the number of epochs to the resulting image, the result may be seen in Fig. 3.

After we were able to reconstruct the image we tested if the reconstruction is able to fit more images conditioned by the input semantic vector, the result may be seen in Fig. 2. We have also tested if the generator is not overfitting the data by feeding random vector as the input, the result may be seen in Fig. 3.

After we have proven that the generator is able to reconstruct known images, we tried if it is able to generalize for unknown images. On Fig. 4 training samples may be seen for 100, 1000 and 10000 images. Model with 100 samples was trained for 4000 epochs, 1000 and 10000 only for 1000 epochs because of very long training time. This explains why the first column is sharper. The last column is the original image.

On Fig. 5 generation of unseen images is shown. The same number of images and training epochs as for Fig. 4 apply. In the training dataset there was a lot of undersea images, therefore we may see that they are reconstructed more precisely. We can see that some simple concepts like the exclamation mark shape or black background or the object color is reconstructed correctly, but unseen and more complicated concepts are not reconstructed.

4. Conclusions and future work

In this work, we wanted to prove that it is not possible to extract personal information from semantic feature vectors extracted from images. We used pre-trained ResNet for the extraction [7]. We wanted to prove this because many ar-

[1] https://github.com/tivvit/resnet_gan

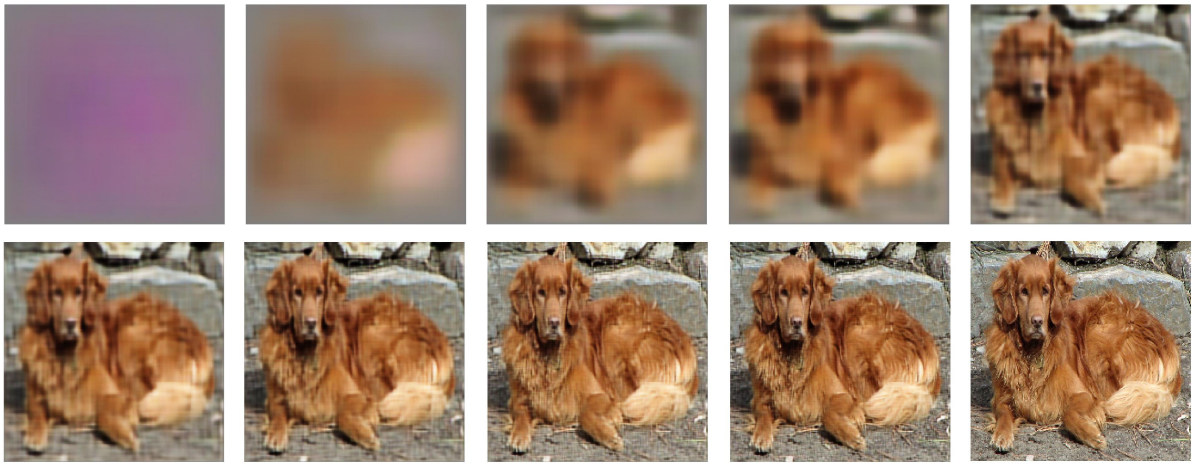

Fig.1. Reconstruction results for different number of training epochs - 1, 10, 50, 100, 500, 1000, 2000, 4000, 6000 and the original image.

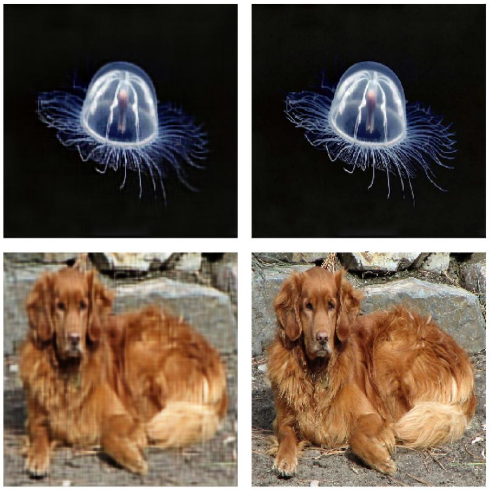

Fig. 2. Image reconstruction conditioned by semantic feature vector. Trained on two images for 4000 epoch.

Fig. 3. Image reconstruction from random vector. Same model as for 2.

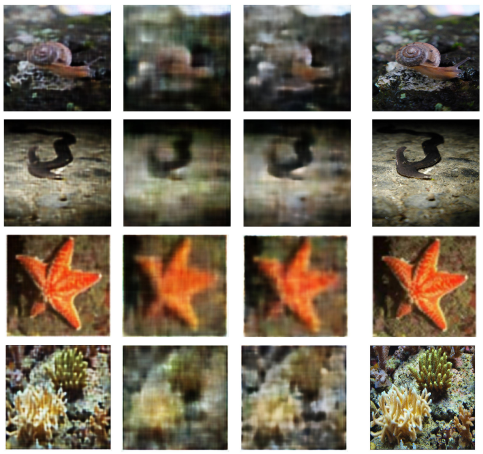

Fig. 4. Training examples, Trained on 100, 1000 and 10000 examples from the left, On the right there is an original image.

eas may benefit from the anonymization property of feature vectors.

We designed a convolution neural network architecture based on DCGAN and tested its properties for image reconstruction [11]. We discovered that it is able to capture the main aspects of the image after few hundreds iteration and the reconstruction almost unrecognizable after few thousand iterations.

The reconstruction is possible, therefore we wanted to examine the generalization properties of the network. We discovered that after feeding thousands of images the network is able to reconstruct main concepts of the image.

We assume that it is possible to train the neural network longer for better results and more images would make it possible to generate more precise reconstructions, but that would be very computationally expensive. Because the number of images is essential for good generalization and number of training epochs is essential for good reconstruction, but both those aspects are both slowing down the training.

Based on this results we concluded that is not possible to reconstruct the private information (like recognizable

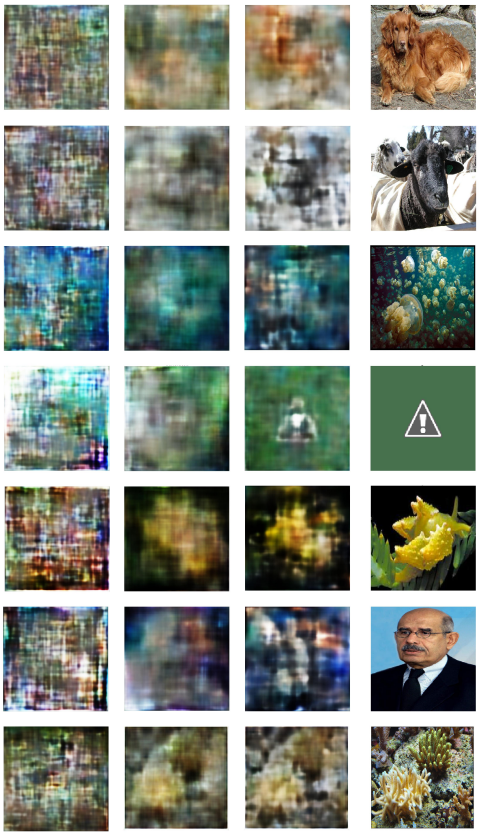

Fig. 5. Testing examples, Trained on 100, 1000 and 10000 examples from the left, On the right there is an original image.

faces) from the semantic feature vectors without a huge number of image samples and enormous computational power.

Acknowledgements

The research described in the paper was supervised by Prof. V. Hlaváč and J. Šedivý CSc. CIIRC in Prague and supported by the Seznam.cz company.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Battista Biggio, Giorgio Fumera, Ignazio Pillai, and Fabio Roli. A survey and experimental evaluation of image spam filtering techniques. *Pattern Recognition Letters*, 32(10):1436–1446, 2011.

[3] François Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[5] PN Druzhkov and VD Kustikova. A survey of deep learning methods and software tools for image classification and object detection. *Pattern Recognition and Image Analysis*, 26(1):9–15, 2016.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[9] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013.

[10] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[11] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[14] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.