

ImageNet pre-trained models with batch normalization

Marcel Simon, Erik Rodner, Joachim Denzler
Computer Vision Group
Friedrich-Schiller-Universität Jena, Germany

{marcel.simon, erik.rodner, joachim.denzler}@uni-jena.de

Abstract

Convolutional neural networks (CNN) pre-trained on ImageNet are the backbone of most state-of-the-art approaches. In this paper, we present a new set of pre-trained models with popular state-of-the-art architectures for the Caffe framework. The first release includes Residual Networks (ResNets) with generation script as well as the batch-normalization-variants of AlexNet and VGG19. All models outperform previous models with the same architecture. The models and training code are available at <http://www.inf-cv.uni-jena.de/Research/CNN+Models.html> and <https://github.com/cvjena/cnn-models>.

1. Introduction

The rediscovery of convolutional neural networks (CNN) [17] in the past years is a result of both the dramatically increased computational speed and the advent of large scale datasets as part of the big data trend. The computational speed was mainly boosted by the efficient use of GPUs for common computer vision functions like convolution and matrix multiplication. Large scale datasets [25, 19, 16, 5, 22, 6], on the other hand, provide the amount of data required for training large scale models with more than a hundred million parameters.

This combination allowed for huge advances in all fields of computer vision research ranging from traditional tasks like classification [11, 27, 28, 3, 8, 18], object detection [23, 26, 10], and segmentation [20, 4, 36], to new ones like image captioning [15, 24, 21, 35, 34], visual question answering [2, 9, 33] and 3D information prediction [7, 32]. Most of these works are based on models, which are pre-trained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset [25]. The classification task of the last year's ILSVRC contains 1.2 million training images categorized into one thousand categories, which represent a wide variety of everyday objects. Pre-training on this dataset proved to be a crucial step for obtaining highly

accurate models in most of the tasks mentioned above.

While computational speed was dramatically increased by the use of GPUs, training a large model like VGG19 [29] still takes several months on a high-end GPU. We hence release a continuously growing set of pre-trained models with popular architectures for the Caffe framework [14]. In contrast to most publicly available models for this framework, our release includes the batch normalization [13] variants of popular networks like AlexNet [17] and VGG19 [29]. In addition, we provide training code for reproducing the results of residual networks [11] in Caffe, which was not provided by the authors of the paper [12]. The release includes all files required for reproducing the model training as well as the log file of the training of the provided model.

2. Batch normalization for CNNs

Especially for larger models like VGG19, batch normalization [13] is crucial for successful training and convergence. In addition, architectures with batch normalization allow for using much higher learning rates and hence yield in models with better generalization ability. In our experiments, we found that higher learning rates show a slower initial convergence speed, but end up at a lower final error rate. This was the case for both AlexNet and VGG19.

The advantage of batch normalization is present even for fine-tuning in certain applications. For example, Amthor *et al.* [1] report that their multi-loss architectures only converged reliably if batch normalization was added to the networks. However, adding batch normalization afterwards to models trained without batch normalization yields in a severe increase in error rates due to mismatching output statistics. Instead, fine-tuning with our batch normalization models is directly possible, which allows for easy adaption to new tasks.

3. Implementation details

We modified AlexNet and VGG19 by adding a batch normalization layer [13] between each convolutional and activation unit layer as well as between each inner product and

activation unit layer. We followed the suggestions of [13] and removed the local response normalization and dropout layers. In addition, we also omitted the mean subtraction during training and replaced it by an batch normalization layer on the input data. This results in an adaptively calculated mean in training and relieves users from manually subtracting the mean during feature computation. In addition, this approach has the advantage that the mean adapts automatically during fine-tuning and no manual mean calculation and storage is required.

We train for 64 epochs on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 – 2016 dataset [25], which contains roughly 1.2 million images and one thousand object categories. A batch size of 256 and initial learning rate of 0.05 (AlexNet), 0.01 (VGG19) and 0.1 (ResNet) was used. The learning rate follows a linear decay over time. Due to batch normalization, it is important that the batch size is greater than sixteen to obtain robust statistics estimations in the batch normalization layers. In the Caffe framework, this means the batch size in the network definition needs to be sixteen or larger, the solver parameter `iter_size` does not compensate a too small batch size in the network definition. If you want to fine-tune a model but do not have enough GPU memory, you can enable the use of global statistics in training in order to lift this batch size requirement. This will disable the statistics estimation in each forward pass and global statistics will be used instead.

All images are resized such that the smaller side has length 256 pixel and the aspect ratio is preserved. During training, we randomly crop a 224×224 (ResNet, VGG19) or 227×227 (AlexNet) pixel square patch and feed it into the network. During validation, a single centered crop is used. We did not use any kind of color, scale or aspect ratio augmentation.

During training of residual networks, we also observe a sudden divergence at random time points in training as explained by Szegedy *et al.* [30]. In this case, we restarted the training using the last snapshot. Due to a different random seed, the order of the images is different and hence the training does not diverge at this time point anymore.

Please note, that the final models are not cherry picked based on the validation error. We provide the final model after the full training is completed. We did not intervene with training and especially did not manually changed the learning rate, as usually done if the step policy is used for the learning rate.

4. Results

The top-1 and top-5-error of the trained models are shown in Table 1. As observed in previous works [13], the error rates benefit from the added batch normalization layers. All provided models slightly improve the error rate achieved by previously trained models [31]. In case of AlexNet, for

Table 1. **Single-crop** top-1 and top-5 error of our models on the validation set of ILSVRC 2012.

Model	Top-1 error		Top-5 error	
	Ours	Original	Ours	Original
AlexNet	39.9%	42.6%	18.1%	19.6%
VGG19	26.9%	28.7%	8.8%	9.9%
ResNet-10	36.1%	–	14.8%	–
ResNet-50	24.6%	24.7%	7.6%	7.8%

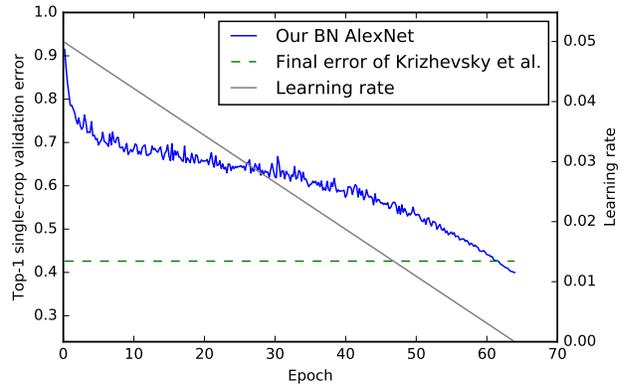


Figure 1. **Single crop** top-1 error of AlexNet on the validation set of ILSVRC 2012 with respect to the training time. We used a linear learning rate decay as shown by the gray curve, which explains the steep decrease in error towards the end of the training.

example, we even observe an error decrease of over 2.6%.

In addition to the final results, we also visualize the single-crop top-1 error on the validation set during the training of AlexNet in Fig. 1. As shown in the figure, the error decreases consistently and fairly quickly during training. Since we use linear learning rate decay, there is a steep error decrease towards the end of the training. While it might look like the error could decrease even further, this is not true. The reason is that the learning rate approaches 0 during the end of the training. Even if the learning rate is kept constant, no improvement can be observed. This is supported by several experiments we performed.

5. Conclusions

This paper presents a new set of pre-trained models for the ImageNet dataset using the Caffe framework. We focus on the batch-normalization-variants of AlexNet and VGG19 as well as residual networks. All models outperform previous pre-trained models. In particular, we were able to reproduce the ImageNet results of residual networks. All models, log files and training code are available at <http://www.inf-cv.uni-jena.de/Research/CNN+Models.html> and <https://github.com/cvjena/cnn-models>.

6. Acknowledgments

The authors thank Nvidia for GPU hardware donations. Part of this research was supported by grant RO 5093/1-1 of the German Research Foundation (DFG)

References

- [1] M. Amthor, E. Rodner, and J. Denzler. Impatient dnns - deep neural networks with dynamic time budgets. In *British Machine Vision Conference (BMVC)*, 2016. (accepted for publication). [1](#)
- [2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, December 2015. [1](#)
- [3] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson. From generic to specific deep representations for visual recognition. *CoRR*, abs/1406.5774, 2014. [1](#)
- [4] C.-A. Brust, S. Sickert, M. Simon, E. Rodner, and J. Denzler. Convolutional patch networks with spatial prior for road detection and urban scene understanding. In *International Conference on Computer Vision Theory and Applications (VIS-APP)*, pages 510–517, 2015. [1](#)
- [5] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. [1](#)
- [6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. *CoRR*, abs/1604.01685, 2016. [1](#)
- [7] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. [1](#)
- [8] A. Freytag, E. Rodner, M. Simon, A. Loos, H. Köhl, and J. Denzler. Chimpanzee faces in the wild: Log-euclidean cnns for predicting identities and attributes of primates. In *German Conference on Pattern Recognition (GCPR)*, 2016. [1](#)
- [9] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2296–2304. Curran Associates, Inc., 2015. [1](#)
- [10] S. Gidaris and N. Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. [1](#)
- [11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. [1](#)
- [12] K. He, X. Zhang, S. Ren, and J. Sun. Pretrained models for residual networks. <https://github.com/KaimingHe/deep-residual-networks>, 2015. [1](#)
- [13] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 448–456, 2015. [1](#), [2](#)
- [14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. [1](#)
- [15] J. Johnson, A. Karpathy, and F. Li. Denscap: Fully convolutional localization networks for dense captioning. *CoRR*, abs/1511.07571, 2015. [1](#)
- [16] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and F. Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *CoRR*, abs/1602.07332, 2016. [1](#)
- [17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 1106–1114, 2012. [1](#)
- [18] C. Lee, S. Xie, P. W. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, 2015. [1](#)
- [19] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. [1](#)
- [20] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3431–3440, 2015. [1](#)
- [21] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. *CoRR*, abs/1511.02283, 2015. [1](#)
- [22] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *International Conference on Computer Vision (ICCV)*, pages 2641–2649, 2015. [1](#)
- [23] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99, 2015. [1](#)
- [24] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating video content to natural language descriptions. In *International Conference on Computer Vision (ICCV)*, December 2013. [1](#)
- [25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li. Imagenet large scale visual recog-

- inition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1, 2
- [26] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013. 1
- [27] M. Simon and E. Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *International Conference on Computer Vision (ICCV)*, 2015. 1
- [28] M. Simon, E. Rodner, and J. Denzler. Part detector discovery in deep convolutional neural networks. In *Asian Conference on Computer Vision (ACCV)*, volume 2, pages 162–177, 2014. 1
- [29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 1
- [30] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016. 2
- [31] A. Vedaldi and K. Lenc. Pretrained models for matconvnet. <http://www.vlfeat.org/matconvnet/pretrained/>, 2015. 2
- [32] X. Wang, D. Fouhey, and A. Gupta. Designing deep networks for surface normal estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1
- [33] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. *CoRR*, abs/1603.01417, 2016. 1
- [34] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 2048–2057, 2015. 1
- [35] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision (ECCV)*, pages 69–85, 2016. 1
- [36] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional random fields as recurrent neural networks. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 1