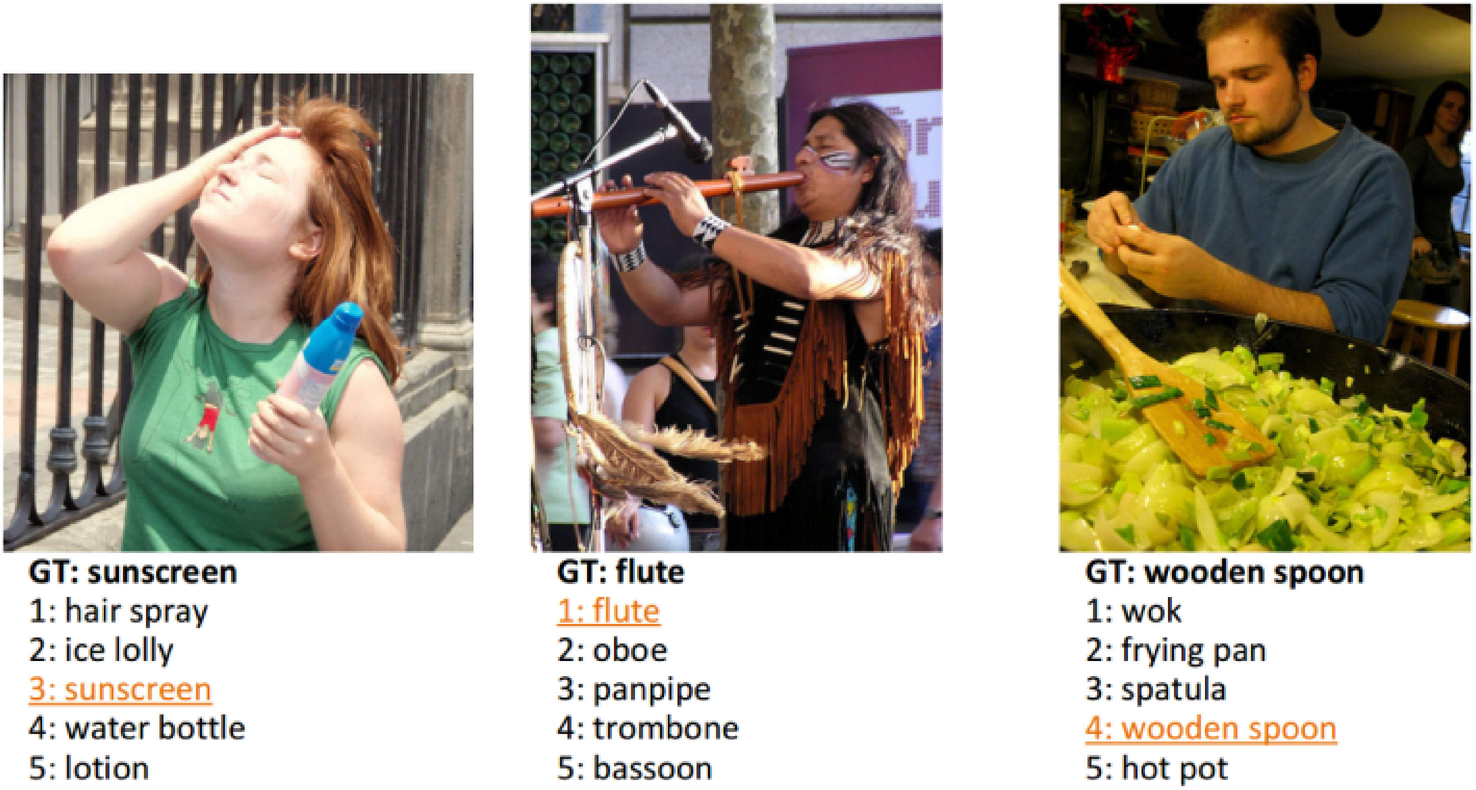
Deep Learning for Big Data



MARCH 25, 2015APRIL 3, 2015/**FANANYMI**

Review on The First Deep Learning that Surpasses Human-Level Performance

Posted by **Mohamad Ivan Fanany (http://www.researchgate.net/profile/Mohamad_Ivan_Fanany)**

Printed version (https://www.academia.edu/11616191/Review_on_The_First_Deep_Learning_that_Surpasses_Human-Level_Performance)

This writing summarizes and reviews on the first reported paper on ImageNet classification using deep learning that surpasses human-level performance: **Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification (http://arxiv.org/pdf/1502.01852v1.pdf)**.

Motivations:

- Convolutional neural networks (CNNs) [**17 (https://www.ics.uci.edu/~welling/teaching/273ASpring09/lecun-89e.pdf)**, **16 (http://www.cs.toronto.edu/~fritz/absps/imagenet.pdf)**] have demonstrated recognition accuracy better than or comparable to humans in several visual recognition tasks, including recognizing traffic signs [**3 (http://people.idsia.ch/~juergen/cvpr2012.pdf)**], faces [**30 (http://www.cv-foundation.org/openaccess/content_cvpr_2014/papers/Taigman_DeepFace_Closing_the_2014_CVPR_paper.pdf)**, **28 (http://arxiv.org/pdf/1406.4773)**], and handwritten digits [**3 (http://arxiv.org/pdf/1406.4773)**, **31 (http://cs.nyu.edu/~wanli/dropc/dropc.pdf)**].
- Tremendous improvements in neural networks recognition performance, mainly due to advances in two technical directions: building more powerful models and designing effective strategies against overfitting.
- Neural networks are becoming more capable of fitting training data due to:
  - Increased depth [**25 (http://arxiv.org/pdf/1409.1556v5.pdf)**, **29 (http://www.cs.unc.edu/~wliu/papers/GoogLeNet.pdf)**]
  - Enlarged width [**33 (http://ftp.cs.nyu.edu/~fergus/papers/zeilerECCV2014.pdf)**, **24 (http://arxiv.org/pdf/1312.6229.pdf)**]
  - The use of smaller strides [**33 (http://arxiv.org/pdf/1312.6229.pdf)**, **24 (http://arxiv.org/pdf/1312.6229.pdf)**, **2 (http://arxiv.org/pdf/1405.3531v4.pdf)**, **25 (http://arxiv.org/pdf/1409.1556v5.pdf)**]
  - New nonlinear activations [**21 (http://www.cs.toronto.edu/~fritz/absps/reluICML.pdf)**, **20 (http://web.stanford.edu/~awni/papers/relu_hybrid_icml2013_final.pdf)**, **34 (http://www.cs.toronto.edu/~fritz/absps/googlerectified.pdf)**, **19 (http://arxiv.org/pdf/1312.4400v3.pdf)**, **27 (http://people.idsia.ch/~rupesh/publications/NIPS2013_srivastava.pdf)**, **9 (http://arxiv.org/pdf/1302.4389.pdf)**]
  - Sophisticated layer designs [**29 (http://www.cs.unc.edu/~wliu/papers/GoogLeNet.pdf)**, **11 (http://arxiv.org/pdf/1406.4729)**].
- Better generalization in neural networks is achieved by
  - Effective regularization techniques [**12 (http://arxiv.org/pdf/1207.0580.pdf)**, **26 (http://www.cs.toronto.edu/~rsalakhu/papers/srivastava14a.pdf)**, **9 (http://arxiv.org/pdf/1207.0580.pdf)**, **31 (http:)**]
  - Aggressive data augmentation [**16 (http://www.cs.toronto.edu/~rsalakhu/papers/srivastava14a.pdf)**, **13 (http://arxiv.org/pdf/1312.5402)**, **25 (http://arxiv.org/pdf/1409.1556v5.pdf)**, **29 (http://www.cs.unc.edu/~wliu/papers/GoogLeNet.pdf)**]
  - Large-scale data [**4 (http://www.cs.toronto.edu/~fritz/absps/imagenet.pdf)**, **22 (http://arxiv.org/pdf/1409.0575v3.pdf)**].
- Among these advances, the rectifier neuron [**21 (http://www.cs.toronto.edu/~fritz/absps/reluICML.pdf)**, **8 (http://eprints.pascal-network.org/archive/00008596/01/glorot11a.pdf)**, **20 (http://web.stanford.edu/~awni/papers/relu_hybrid_icml2013_final.pdf)**, **34 (http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=AA894CF2CC7192E8DDEB8EE3400365AE?doi=10.1.1.308.5055&rep=rep1&type=pdf)**], e.g., Rectified Linear Unit (ReLU), is one

of several keys to the recent success of deep networks [**16 (http://www.cs.toronto.edu/~fritz/absps/imagenet.pdf)**].
- Rectifier neuron expedites convergence of the training procedure [**16 (http://www.cs.toronto.edu/~fritz/absps/imagenet.pdf)**] and leads to better solutions [**21 (http://www.cs.toronto.edu/~fritz/absps/reluICML.pdf)**, **8 (http://eprints.pascal-network.org/archive/00008596/01/glorot11a.pdf)**, **20 (http://web.stanford.edu/~awni/papers/relu_hybrid_icml2013_final.pdf)**, **34 (http://www.cs.toronto.edu/~fritz/absps/googlerectified.pdf)**] than conventional sigmoidlike units.
- Despite the prevalence of rectifier networks, recent improvements of models [**33 (http://ftp.cs.nyu.edu/~fergus/papers/zeilerECCV2014.pdf)**, **24 (http://arxiv.org/pdf/1312.6229)**, **11 (http://arxiv.org/pdf/1406.4729)**, **25 (http://arxiv.org/pdf/1409.1556v5.pdf)**,**29 (http://www.cs.unc.edu/~wliu/papers/GoogLeNet.pdf)**] and theoretical guidelines for training them [**7 (http://jmlr.org/proceedings/papers/v9/glorot10a/glorot10a.pdf)**, **23 (http://arxiv.org/pdf/1312.6120v3.pdf)**] have rarely focused on the properties of the rectifiers.

Key ideas:

- Investigate neural networks from two aspects particularly driven by these rectifiers. propose a new generalization of ReLU, which is called Parametric Rectified Linear Unit (PReLU).
- This activation function adaptively learns the parameters of the rectifiers, and improves accuracy at negligible extra computational cost.
- Study the difficulty of training rectified models that are very deep (e.g., 30 weights layers)
- Derive a theoretically sound initialization method, which helps with convergence of very deep models trained directly from the scratch by explicitly modeling the nonlinearity of rectifiers (ReLU/PReLU). This gives more flexibility to explore more powerful network architectures.

Datasets:

- 1000-class ImageNet 2012 dataset [**22 (http://arxiv.org/abs/1409.0575)**] which contains about 1.2 million training images, 50,000 validation images, and 100,000 test images (with no published labels).
- The results are measured by top-1/top-5 error rates [**22 (http://arxiv.org/abs/1409.0575)**].
- Only use the provided data for training. All results are evaluated on the validation set, except for the final results, which are evaluated on the test set.
- The top-5 error rate is the metric officially used to rank the methods in the classification challenge [**22 (http://arxiv.org/abs/1409.0575)**].

Results:

- On the 1000-class ImageNet 2012 dataset, the PReLU network (PReLU-net) leads to a single-model result of 5.71% top-5 error, which surpasses all existing multi-model results.
- The proposed multi-model result achieves 4.94% top-5 error on the test set, which is a 26% relative improvement over the ILSVRC 2014 winner (GoogLeNet, 6.66% [**29 (http://www.cs.unc.edu/~wliu/papers/GoogLeNet.pdf)**]).
- The result surpasses for the first time the reported human-level performance (5.1% in [**22**

(http://arxiv.org/abs/1409.0575)]) on this visual recognition challenge.

Parametric rectifiers:

- For PReLU, the coefficient of the negative part is not constant and is adaptively learned.
- Replacing the parameter-free ReLU (Rectified Linear Unit) activation by a learned parametric activation unit improves classification accuracy. Concurrently, Agostinelli et al. [**1 (http://arxiv-web3.library.cornell.edu/pdf/1412.6830v2)**] also investigated learning activation functions and showed improvement on other tasks.
- PReLU introduces a very small number of extra parameters. The number of extra parameters is equal to the total number of channels, which is negligible when considering the total number of weights. So we expect no extra risk of overfitting.
- The paper also considers a channel-shared variant where the coefficient is shared by all channels of one layer. This variant only introduces a single extra parameter into each layer.
- PReLU can be trained using backpropagation [**17 (https://www.ics.uci.edu/~welling/teaching/273ASpring09/lecun-89e.pdf)**] and optimized simultaneously with other layers.
- The time complexity due to PReLU is negligible for both forward and backward propagation.
- The paper adopts the momentum method when updating.
- It is worth noticing that the paper does not use weight decay (l2 regularization) when updating. A weight decay tends to push the coefficient controlling the slope of the negative part to zero, and thus biases PReLU toward ReLU. Even without regularization, the learned coefficients rarely have a magnitude larger than 1 in the experiments.
- The experiment in the paper does not constrain the range of the coefficient controling the slope of the negative part so that the activation function may be non-monotonic.

Baseline comparisons:

- As a baseline, the model is trained with ReLU applied in the convolutional (conv) layers and the first two fully connected (fc) layers. The training implementation follows [**10 (http://arxiv.org/abs/1412.1710)**]. The top-1 and top-5 errors are 33.82% and 13.34% on ImageNet 2012, using 10-view testing.
- The same architecture is then trained from scratch, with all ReLUs replaced by PReLUs. The top-1 error is reduced to 32.64%. This is a 1.2% gain over the ReLU baseline.
- The result also shows that channel-wise/channel-shared PReLUs perform comparably. For the channel-shared version, PReLU only introduces 13 extra free parameters compared with the ReLU counterpart. But this small number of free parameters play critical roles as evidenced by the 1.1% gain over the baseline. This implies the importance of adaptively learning the shapes of activation functions.

Initialization of filter weights for rectifiers:

- Rectifier networks are easier to train [**8 (http://eprints.pascal-network.org/archive/00008596/01/glorot11a.pdf)**, **16 (http://www.cs.toronto.edu/~fritz/absps/imagenet.pdf)**, **34 (http://www.cs.toronto.edu/~fritz/absps/googlerectified.pdf)**] compared with traditional sigmoid-like activation networks. But a bad initialization can still hamper the learning of a highly non-linear system. In the paper, the authors propose a robust initialization method that removes an obstacle of training extremely deep rectifier networks.

- Recent deep CNNs are mostly initialized by random weights drawn from Gaussian distributions [**16 (http://www.cs.toronto.edu/~fritz/absps/imagenet.pdf)**]. With fixed standard deviations (e.g., 0.01 in [**16 (http://www.cs.toronto.edu/~fritz/absps/imagenet.pdf)**]), very deep models (e.g., >8 conv layers) have difficulties to converge, as reported by the VGG team [**25 (http://arxiv.org/pdf/1409.1556v5.pdf)**] and also observed in the authors experiments. To address this issue, in [**25 (http://arxiv.org/pdf/1409.1556v5.pdf)**] they pre-train a model with 8 conv layers to initialize deeper models. But this strategy requires more training time, and may also lead to a poorer local optimum. In [29, 18], auxiliary classifiers are added to intermediate layers to help with convergence.
- Glorot and Bengio [**7 (http://jmlr.org/proceedings/papers/v9/glorot10a/glorot10a.pdf)**] proposed to adopt a properly scaled uniform distribution for initialization. This is called "Xavier" initialization in [**14 (http://ucb-icsi-vision-group.github.io/caffe-paper/caffe.pdf)**]. Its derivation is based on the assumption that the activations are linear. This assumption is invalid for ReLU and PReLU.
- In the paper, authors derive a theoretically more sound initialization by taking ReLU/PReLU into account. In their experiments, the proposed initialization method allows for extremely deep models (e.g., 30 conv/fc layers) to converge, while the "Xavier" method [**7 (http://jmlr.org/proceedings/papers/v9/glorot10a/glorot10a.pdf)**] cannot.
- The main difference between the proposed derivation and the "Xavier" initialization [**7 (http://jmlr.org/proceedings/papers/v9/glorot10a/glorot10a.pdf)**] is that the proposed derivation address the rectifier nonlinearities.
- The studies conducted in the paper show that the readiness to investigate extremely deep, rectified models by using a more principled initialization method. But in their current experiments on ImageNet, they have not observed the benefit from training extremely deep models.
- Accuracy saturation or degradation was also observed in the study of small models [**10 (http://arxiv.org/abs/1412.1710)**], VGG's large models [**25 (http://arxiv.org/pdf/1409.1556v5.pdf)**], and in speech recognition [**7 (http://jmlr.org/proceedings/papers/v9/glorot10a/glorot10a.pdf)**]. This is perhaps because the method of increasing depth is not appropriate, or the recognition task is not complex enough.
- Though the attempts of extremely deep models have not shown benefits, the proposed initialization method paves a foundation for further study on increasing depth.

Network architecture, hardware, and training time:

- The baseline architecture in the paper is the 19-layer model (A). For a better comparison, the paper also lists the VGG-19 model [**25 (http://arxiv.org/pdf/1409.1556v5.pdf)**]. The baseline model A has the following modifications on VGG-19:
  1. In the first layer, they use a filter size of 7×7 and a stride of 2;
  2. They move the other three conv layers on the two largest feature maps (224, 112) to the smaller feature maps (56, 28, 14). The time complexity is roughly unchanged because the deeper layers have more filters;
  3. They use spatial pyramid pooling (SPP) [**11 (http://arxiv.org/pdf/1406.4729)**] before the first fc layer. The pyramid has 4 levels – the numbers of bins are 7×7, 3×3, 2×2, and 1×1, for a

total of 63 bins.

- No evidence that the proposed model A is a better architecture than VGG-19, though the model A has better results than VGG-19's result reported by [**25 (http://arxiv.org/pdf/1409.1556v5.pdf)**].
- The model A and a reproduced VGG-19 (with SPP and the authors initialization) are comparable. The main purpose of using model A is for faster running speed. The actual running time of the conv layers on larger feature maps is slower than those on smaller feature maps, when their time complexity is the same.
- In four-GPU implementation, the model A takes 2.6s per mini-batch (128), and the reproduced VGG-19 takes 3.0s, evaluated on four Nvidia K20 GPUs.
- The proposed model B is a deeper version of A. It has three extra conv layers. The proposed model C is a wider (with more filters) version of B. The width substantially increases the complexity, and its time complexity is about 2.3× of B. Training A/B on four K20 GPUs, or training C on eight K40 GPUs, takes about 3-4 weeks.
- The authors choose to increase the model width instead of depth, because deeper models have only diminishing improvement or even degradation on accuracy.
- In recent experiments on small models [**10 (http://arxiv.org/abs/1412.1710)**], it has been found that aggressively increasing the depth leads to saturated or degraded accuracy.
- In the VGG paper [**25 (http://arxiv.org/pdf/1409.1556v5.pdf)**], the 16-layer and 19-layer models perform comparably. In the speech recognition research of [**7 (http://jmlr.org/proceedings/papers/v9/glorot10a/glorot10a.pdf)**, the deep models degrade when using more than 8 hidden layers (all being fc).
- The authors conjecture that similar degradation may also happen on larger models for ImageNet. After monitored the training procedures of some extremely deep models (with 3 to 9 layers added on B in Table 3), and found both training and testing error rates degraded in the first 20 epochs (but did not run to the end due to limited time budget, so there is not yet solid evidence that these large and overly deep models will ultimately degrade). Because of the possible degradation, the authors choose not to further increase the depth of these large models.
- On the other hand, the recent research [5] on small datasets suggests that the accuracy should improve from the increased number of parameters in conv layers. This number depends on the depth and width. So the authors choose to increase the width of the conv layers to obtain a higher capacity model.
- While all B models are very large, no severe overfitting are observed. The authors attribute this to the aggressive data augmentation used throughout the whole training procedure,

Training:

- The training algorithm mostly follows [**16 (http://www.cs.toronto.edu/~fritz/absps/imagenet.pdf)**, **13 (http://arxiv.org/pdf/1312.5402)**, **2 (http://arxiv.org/abs/1405.3531)**, **11 (http://arxiv.org/pdf/1406.4729)**, **25 (http://www.robots.ox.ac.uk/~vgg/research/very_deep/)**]. From a resized image whose shorter side is s, a 224×224 crop is randomly sampled, with the per-pixel mean subtracted. The scale is randomly jittered in the range of [256, 512], following **25 (http://arxiv.org/pdf/1409.1556v5.pdf)**]. One half of the random samples are flipped horizontally [**16 (http://www.cs.toronto.edu/~fritz/absps/imagenet.pdf)**]. Random color altering [**16 (http://www.cs.toronto.edu/~fritz/absps/imagenet.pdf)**] is also used.

- Unlike [**25 (http://arxiv.org/pdf/1409.1556v5.pdf)**] that applies scale jittering only during finetuning, the authors apply it from the beginning of training. Further, unlike [**25 (http://arxiv.org/pdf/1409.1556v5.pdf)**] that initializes a deeper model using a shallower one, the authors directly train the very deep model using their initialization. Their end-to-end training may help improve accuracy, because it may avoid poorer local optima.
- Other hyper-parameters that might be important are as follows.
  - The weight decay is 0.0005, and momentum is 0.9.
  - Dropout (50%) is used in the first two fc layers.
  - The minibatch size is fixed as 128. The learning rate is 1e-2, 1e-3.

Testing:

- The paper adopts the strategy of "multi-view testing on feature maps" used in the SPP-net paper [**11 (http://arxiv.org/pdf/1406.4729)**]. This strategy is further improved using the dense sliding window method in [**24 (http://arxiv.org/pdf/1312.6229)**,**25 (http://www.robots.ox.ac.uk/~vgg/research/very_deep/)**].
- The authors first apply the convolutional layers on the resized full image and obtain the last convolutional feature map. In the feature map, each 14×14 window is pooled using the SPP layer [**11 (http://arxiv.org/pdf/1406.4729)**].
- The fc layers are then applied on the pooled features to compute the scores. This is also done on the horizontally flipped images. The scores of all dense sliding windows are averaged [**24 (http://arxiv.org/pdf/1312.6229)**,**25 (http://www.robots.ox.ac.uk/~vgg/research/very_deep/)**]. They further combine the results at multiple scales as in [**11 (http://arxiv.org/pdf/1406.4729)**].

Multi-GPU Implementations:

- The paper adopts a simple variant of Krizhevsky's method [**15 (http://arxiv.org/pdf/1404.5997v2.pdf)**] for parallel training on multiple GPUs.
- The paper adopts "data parallelism" [**15 (http://arxiv.org/pdf/1404.5997v2.pdf)**] on the conv layers.
- The GPUs are synchronized before the first fc layer. Then the forward/backward propagations of the fc layers are performed on a single GPU – this means that they do not parallelize the computation of the fc layers. The time cost of the fc layers is low, so it is not necessary to parallelize them. This leads to a simpler implementation than the "model parallelism" in [**15 (http://arxiv.org/pdf/1404.5997v2.pdf)**].
- Besides, model parallelism introduces some overhead due to the communication of filter responses, and is not faster than computing the fc layers on just a single GPU.
- The authors implement the above algorithm on our modification of the Caffe library [**14 (http://ucb-icsi-vision-group.github.io/caffe-paper/caffe.pdf)**]. We do not increase the mini-batch size (128) because the accuracy may be decreased [**15 (http://ucb-icsi-vision-group.github.io/caffe-paper/caffe.pdf)**]. For the large models in this paper, we have observed a 3.8x speedup using 4 GPUs, and a 6.0x speedup using 8 GPUs.

Comparisons with human performance:

- Russakovsky et al. [**22 (http://arxiv.org/pdf/1409.0575v3.pdf)**] recently reported that human performance yields a 5.1% top-5 error on the ImageNet dataset. This number is achieved by a human annotator who is well trained on the validation images to be better aware of the existence of relevant classes.

- When annotating the test images, the human annotator is given a special interface, where each class title is accompanied by a row of 13 example training images. The reported human performance is estimated on a random subset of 1500 test images.
- The classification result (4.94%), reported in the paper, exceeds the reported human-level performance. Up to now, the result is the first published instance of surpassing humans on this visual recognition challenge.
- The analysis in [**22 (http://arxiv.org/pdf/1409.0575v3.pdf)**] reveals that the two major types of human errors come from fine-grained recognition and class unawareness. The investigation in [**22 (http://arxiv.org/pdf/1409.0575v3.pdf)**] suggests that algorithms can do a better job on fine-grained recognition (e.g., 120 species of dogs in the dataset).
- While humans can easily recognize these objects as a bird, a dog, and a flower, it is nontrivial for most humans to tell their species. On the negative side, the algorithm still makes mistakes in cases that are not difficult for humans, especially for those requiring context understanding or high-level knowledge (e.g., the "spotlight" images).
- While the algorithm produces a superior result on this particular dataset, the authors admit this does not indicate that machine vision outperforms human vision on object recognition in general.
- On recognizing elementary object categories (i.e., common objects or concepts in daily lives) such as the Pascal VOC task [6], machines still have obvious errors in cases that are trivial for humans. Nevertheless, the results show the tremendous potential of machine algorithms to match human-level performance on visual recognition.

My Review:

- One interesting aspect about this paper is the reported classification performance of the deep learning algorithm that surpasess human-level performance (though caution should be taken carefully).
- Yet it is not easy to find, what actually drive this impressive performance: the use of PReLU? wider and deeper structure? better initialization? or better design? It would be nice if the authors resolve the improvement by each of these factors in steps, piece by piece, a kind of ablation study.
- While proper initialization using PReLU allow a very deep structure to converge, whereas structure using 'Xavier' initialization cannot converge, the authors also stated that deeper models have only diminishing improvement or even degradation on accuracy.
- It seems we still have no sound theoretical basis how the PReLU propagates the distinguishing capability all the way down from the input to the output.

Posted in **artificial intelligence**, **computer science**, **computer vision**, **data mining**, **deep learning**, **machine learning**, **machine vision**, **review**, **reviews**, **soft computing** / Tagged **artificial intelligence**, **computer science**, **computer vision**, **data mining**, **deep learning**, **ImageNet**, **machine learning**, **machine vision**, **neural network**, **neural networks**, **review**, **reviews**, **soft computing**, **visualization** / **Leave a comment**

- Follow

Follow “Deep Learning for Big Data”