neurala

How Lifelong-DNN™ Solves for Inherent Problems with Traditional DNNs

A TECHNICAL BRIEF

15 January 2019

Neurala, Inc
51 Sleeper Street, Floor 3
Boston, MA 02210

1.617.418.6161
info@neurala.com
www.neurala.com

TABLE OF CONTENTS

neurala

Deep Learning knowledge representation lacks the flexibility to enable the real time learning necessary to overcome catastrophic forgetting and its shortcomings in transfer learning and updating networks with new data. Neurala has developed a proprietary suite of Lifelong AI technologies which directly address these problems in concert with proprietary DNN implementations. The resulting hybrid architecture, Lifelong-DNN, is explained in this paper specifically as applied to the recognition problem in computer vision.

DEEP LEARNING ARCHITECTURES AND SHORTCOMINGS

Rooted in theoretical work from the 1960s, today's Deep Learning and deep neural network algorithms (Collectively referred to as DNNs herein) represent the sub-field of Artificial Intelligence (AI) that is delivering the biggest wins across industries and use cases. These are large-scale systems that grossly simplify and simulate networks of interconnected "neurons" trained to execute specific tasks. Such tasks include recognizing scenes or objects in images and video, natural language processing and time series analysis such as financial fraud detection.

While the data and the tasks may vary, these systems derive their power from the ability to learn from the data – as opposed to being preprogrammed to perform a function – and the distributed nature of knowledge representation across millions of weights or parameters. They also overwhelmingly use a learning formalism of error backpropagation initially suggested in the 1960s and fully crystallized in the 1980s. Backpropagation of error computes the error term by comparing the neural network output with the ground truth, then propagates it backwards through the network to adjust weights in a slow gradient descent towards the final state of the network where the error is minimized.

Data can be images, text or data feeds from a variety of sources, but the principle is the same: the algorithm optimizes the network output by iteratively adjusting the weight of each neuron in the network until optimal performance is achieved. This enables AI systems based on backpropagation to match and sometimes surpass human-level performance in an ever-increasing list of tasks, from playing chess to detecting an intruder in a security camera.

CATASTROPHIC FORGETTING

The excellent performance of DNN comes at a price: a distributed representation indicates that all weights are important for correct behavior. Attempts to add new information to the network require modification of these weights, which in turn disrupts the previous knowledge quite drastically, causing the problem known as *catastrophic forgetting*. Furthermore, the slow gradient descent nature of backpropagation-based learning leads to the need of multiple presentations of each input. This ensures that the network converges to an acceptable solution. Finally, since the complete set of data is usually too large to complete gradient descent for all inputs, these inputs are split into batches stochastically. To ensure convergence of resulting stochastic gradient descent, the steps taken on each iteration of training (i.e. learning rate of the system) must be very small, which results in slow learning.

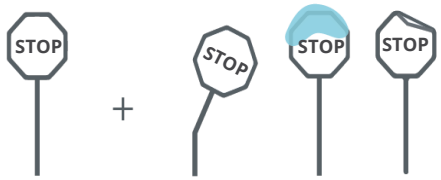
Thus, to achieve high accuracy performance using DNNs, researchers take two steps:

1. Make learning slow by changing weights by only a small amount and repeating the process over many iterations, typically many hundreds of thousands or millions, where each input is presented repeatedly and its error continuously reduced, a little at a time.

2. Freeze learning after the target performance is reached to avoid compromising information already learned and completely prevent the addition of new information.

INABILITY TO UPDATE NETWORKS IN REAL TIME

As a consequence of the algorithmic fixes backpropagation puts in place to overcome catastrophic forgetting, DNNs are unsuitable for real-time (or continuous) learning, a training regimen where information can be added as new data is encountered in a deployed scenario. For example, if a system is trained to detect stop signs and it has difficulty recognizing a partially snow-occluded, bent stop sign, information cannot be added on-the-fly to improve classification.



This makes each DNN only as good as they were trained before being deployed; nothing new can be learned during day-to-day operation. To bypass this limitation within a conventional deep learning paradigm one needs to create a new set of data that contains the data from original training as well as examples of new information intermixed together and retrain the neural network on this extended data set.

In addition to training time and high compute server requirements, backpropagation-based systems come with an additional drawback: they require storage of all or at least a good chunk of input data for possible future re-training. For example, if a DNN has been trained on 1,000 classes and needs to learn an additional class, the examples of all 1,001 classes need to be presented for thousands or millions of iterations. What happens if one does not have (or cannot legally keep) images of the original 1,000 classes? Retraining can't occur, and the network can't be updated.

TRANSFER LEARNING

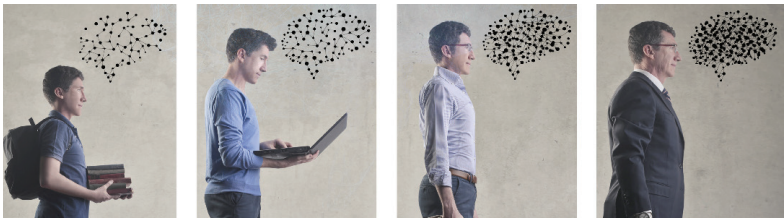
The delicate training regime required for DNNs results in a network that is brittle and unable to transfer its learning beyond the examples it has been trained with. As a result, DNNs require massive datasets, often tens of thousands of examples for each output

that needs to be learned. Without the ability to update these networks in any robust or efficient way after deployment, we are left with the unfeasible tasks of thinking through all possible scenarios of deployment and collecting that relevant data before building the network. Or, worse and more costly, the developer is faced with retraining the entire network from scratch each time new information needs to be learned.

NEURALA'S LIFELONG-DNN TECHNOLOGY SOLVES KEY SHORTCOMINGS OF DNNS

Neurala has built and commercialized an alternative solution, an AI that learns day after day like biological brains do. Most of us received some sort of standardized education where we went to school to learn a set of important skills. As humans, we learn and improve our skills and knowledge daily beyond our standardized education, even when our brains age and become less agile, and, more importantly, we do this quickly: for most of what we store in memory, a few learning episodes are enough, thankfully. If our brains didn't learn this quickly, we would not have survived as a species.

If humans learned like a traditional DNN, we would only know what we were taught in school and would never add new knowledge or understanding. That is how today's AI works, and its limitations are obvious.



Neurala's technology is inspired by brain neurophysiology and mimics in software the ability of cortical and subcortical circuits to work in tandem to add new information on the fly.

LIFELONG-DNN SOLVES CATASTROPHIC FORGETTING THROUGH COMPLEMENTARY LEARNING

Since the early 1990s, researchers have pointed out how there exist multiple, often complementary learning systems in the brain that could be combined to overcome individual systems deficiencies.

For instance, McClelland and colleagues in 1995[1] suggested that backpropagation based neural networks can overcome catastrophic forgetting by utilizing a dual learning system similar to interactions between slow learning neocortical circuits (e.g., the ones found in the 6-layered structure of the cerebral cortex) and fast learning hippocampal network in the human brain. Cortical circuits are characterized as having highly distributed representation, high data compression and generalization, and slow learning similar to back-propagation based neural networks. Hippocampal circuits are known for recurrent connectivity, sparse representations, and fast Hebbian learning.

Lifelong-DNN builds such a dual learning system on top of conventional DNN as described below.

in the figure on page 7: ***the backbone (top bracket) and the head (bottom bracket).***

The backbone defines the quality of feature extraction and its complexity generally dominates the inference speed and quality. The same backbone can be used for different tasks, for example ResNet-101[2] backbone can be found in classification, detection, and segmentation neural networks.

The head defines which task the neural network performs. The complexity of the head changes with the task. Classification heads are quite simple, while segmentation heads usually involve many layers to upsample the result to proper resolution.

 NOTES: 1 http://www.image-net.org 2 See McClelland J.L., McNaughton B.L., and O'Reilly R.C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. Psychol. Rev. 102: 419-457. 3 https://ieeexplore.ieee.org/document/7780459/citations#citations

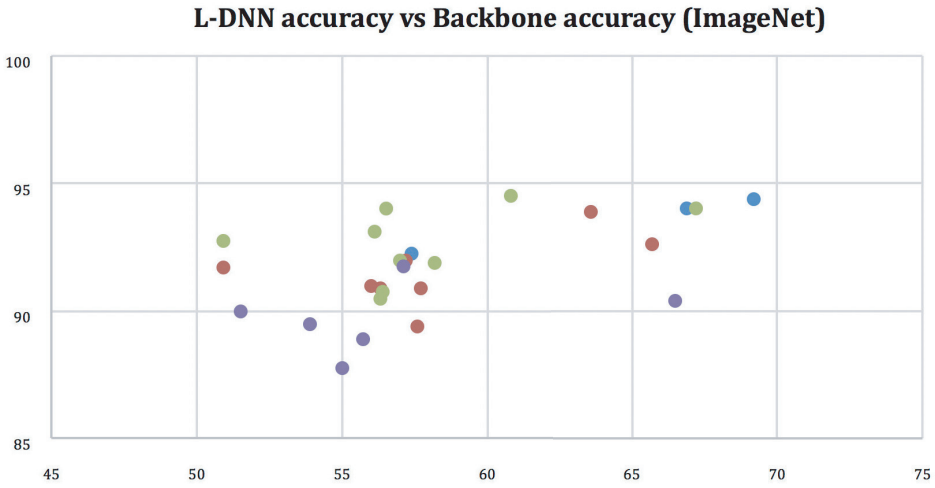
Fast on-the-fly learning is a task, hence Lifelong-DNN modifications are concentrated in the head, although they influence the backbone design and tuning. Head design is the subject of many possible manipulations: introduction of custom layers, new loss functions, special training procedure tweaks, and more. Lifelong-DNN design takes advantage of all of these manipulations and builds a fast-learning head based on the properties of hippocampal system in the human brain: recurrence, sparsity, and Hebbian learning.

The Lifelong-DNN system takes advantage of the fact that weights in the backbone DNN are excellent feature extractors. In order to connect the fast learning head to the backbone, some of the DNN's upper layers may be ignored, modified, or stripped, depending on the use case. For instance, the original DNN usually includes a number of fully connected, averaging, and pooling layers plus a cost layer that is used to enable the gradient descent technique to optimize its weights during training. These layers are used during backbone pre-training or for getting direct predictions from the DNN but aren't necessary for generating an input for the new fast learning head of Lifelong-DNN.

LIFELONG-DNN: 'HEAD' AND 'BACKBONE' INTERPLAY

The quality of the features and their dependency on a particular object set are of critical importance for Lifelong-DNN performance. The better the feature set and the more it is tied to generic 'objectness' rather than a particular object, the higher performance Lifelong-DNN will achieve on the objects that were never used for the backbone pre-training. Feature quality depends on backbone architecture and training. Neurala has focused comprehensively on investigating, comparing, modifying and including in our architecture state-of-the-art mobile backbones like SqueezeNet and MobileNet, as well as proprietary designs of backbones. Additionally, possible manipulations of the loss functions and training procedures can be used to maximize the feature quality and object independence.

The plot above shows four different training/loss modifications (in different colors) for a selection of backbones and compares the accuracy of Lifelong-DNN results on a custom dataset to original backbone ImageNet[4] accuracy. In this study the custom training has 14 training images per each of 10 classes, 5 of which do and 5 of which do not have similar ImageNet classes. The images used are of similar size and quality to ImageNet images. The validation set has 16 images per class. Performance on ImageNet has very small



correlation with Lifelong-DNN performance, and the backbone that by itself is not a top-notch performer on ImageNet can have high quality of features to ensure Lifelong-DNN success.

One important requirement for Lifelong-DNN is the similarity of the feature space between the dataset used to pretrain the backbone and the data used to train and test Lifelong-DNN. For example, the above backbones are pretrained on ImageNet, and Lifelong-DNN performs well on the images that are similar in size and quality to ImageNet images, but does not perform as well on CIFAR-10[5] images that are of much smaller sizes or MNIST[6] digits that are grayscale and have a different feature set in addition to the scale change. On the other hand, if the backbone is pretrained on a subset of MNIST dataset, for example nine digits out of 10, Lifelong-DNN learning of the missing digits is both fast and accurate.

This means that Lifelong-DNN has the ability to improve performance on multiple DNN architectures.

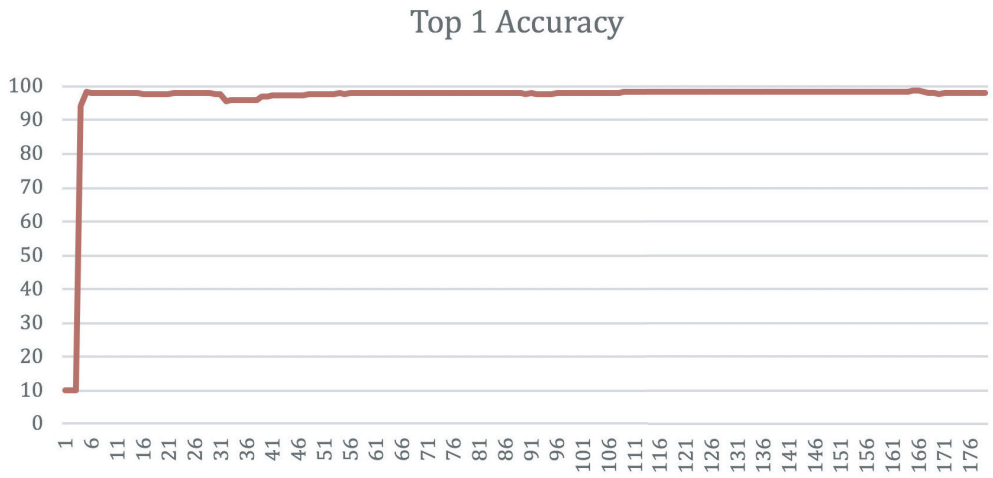
LIFELONG-DNN ENABLES NETWORK UPDATES IN REAL TIME

Below is the confusion matrix for Lifelong-DNN that has a backbone pretrained on MNIST dataset with digit '3' excluded. Lifelong-DNN is trained on all 10 digits and digit '3' is highlighted in the confusion matrix in green with red ovals marking the highest errors (9 is confused for 3 and 3 is confused for 5). The overall accuracy of Lifelong-DNN is 98.4%, which is 1.1% below the accuracy of the same DNN trained directly on all 10 MNIST digits, but the training time of Lifelong-DNN is in a different league compared to the original DNN training as shown in the plot Below.

 NOTES: 5 https://www.cs.toronto.edu/~kriz/cifar.html 6 http://yann.lecun.com/exdb/mnist/

Confusion Matrix

9.79	0	0	0	0	0	0	0.01	0	0
0	11.3	0.02	0.02	0.03	0	0.01	0	0	0
0	0	10.2	0.02	0	0	0.02	0.01	0.02	0.01
0.03	0	0.03	9.63	0	0.36	0	0.02	0.01	0
0	0.02	0	0	9.59	0	0.12	0	0	0.09
0.02	0	0	0.03	0	8.86	0.01	0	0	0
0.03	0	0.01	0	0	0	9.53	0	0.01	0
0	0.03	0.08	0	0	0	0	10.2	0	0
0	0	0.01	0.03	0	0.01	0	0	9.68	0.01
0.01	0	0	0.33	0.04	0.01	0	0.01	0.06	9.63



Here each batch (x-axis) has 10 images, one per digit, and by batch 5 the top 1 validation accuracy is 98.22%, so the training could finish in as little as 0.4 seconds comparing to 6 hours on the same hardware that the same DNN will take to train with backpropagation.

L-DNN vs. DNN learning has a 50,000 times faster speed performance. L-DNN is differentiated to competing techniques when new information needs to be added to an already acquired one.

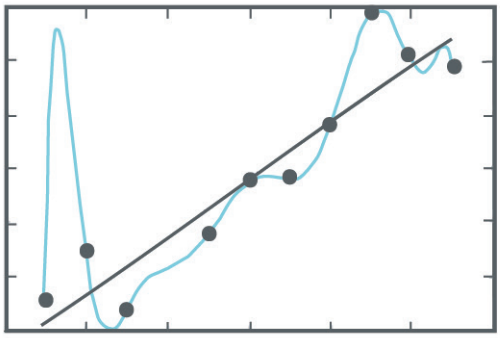
LIFELONG DNN SOLVES TRANSFER LEARNING ISSUES THROUGH MEMORY CONSOLIDATION

Avoiding overfitting and preserving network size: Lifelong-DNN memory consolidation

With such a fast training, overfitting can become a major concern in conventional DNNs. As the plot on the next page shows, it does slightly affect Lifelong-DNN (around batch 33), but the network recovers after a few more batches of training.

The reason for such low overfitting and good recovery is the mechanism of memory consolidation that Lifelong-DNN uses during training. Consolidation utilizes a high degree of redundancy in DNN data representation. The learning metasystem monitors the weight changes and detects outliers for each class. When the number of such outliers crosses the automatic threshold, the consolidation mechanism is triggered.

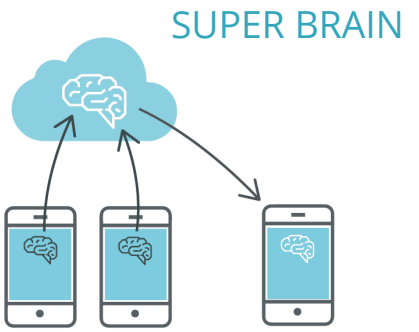
Lifelong-DNN consolidation is based on statistical analysis of the weights in the fast-learning head. Consolidation tightens the clusters for each class by removing the redundant weight vectors and resetting these weights to initial values while preserving the true outliers. Effectiveness of consolidation depends on the feature quality: removing points in sparse clusters leads to the loss of more information than removing the same number of points in the already tight clusters. Consolidation also allows Lifelong-DNN to learn more objects using the same amount of device memory.



Example of overfitting.

Consolidation and federated learning: Brain Melding™ consolidates, or merges, brains across devices

One of the most interesting features of consolidation is that it can be done across multiple Lifelong-DNNs™. Naturally, these Lifelong-DNNs™ need to have the same backbone and consistency between labeling schemes (i.e. calling the same object by different labels will confuse the system). If these requirements are met, then one can consolidate heads from multiple Lifelong-DNNs™ into a single head that will combine the knowledge of all the original heads. This process is called Brain Melding™ and it has far



reaching implications for federated learning and information sharing.

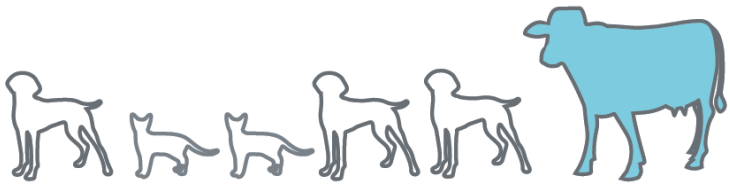
Adapting to real-world scenarios

The real-life deployment of a neural network that allows additional learning like

Lifelong-DNN does differ significantly from the lab dataset training of unpredictability of incoming information. In real life, it is a rule, not an exception, when new objects are likely to be presented one by one without shuffling. Furthermore, the network is very likely to be engaged in inference on images that contain no known objects at all. The latter case "handled" by traditional DNNs by shows the best guess, which frequently is not a desired behavior. A much better solution is for networks to respond to objects that have not been learned.

Lifelong-DNN adapts to real-life conditions through a combination of memory consolidation and an ability to recognize unknown objects by declaring them as "Nothing I Know"—a separate entity compared to known classes. An implementation of this concept works as an implicitly dynamic threshold that favors predictions in which the internal knowledge distribution is clearly focused on a common category as opposed to flatly distributed over several categories.

In other words, when the neural network indicates that there is a clear winner among known object classes, it recognizes the object as belonging to the known class. But when multiple different objects have similar activations and there is no clear winner, the system reports the object as unknown. This thresholding is also useful in the training

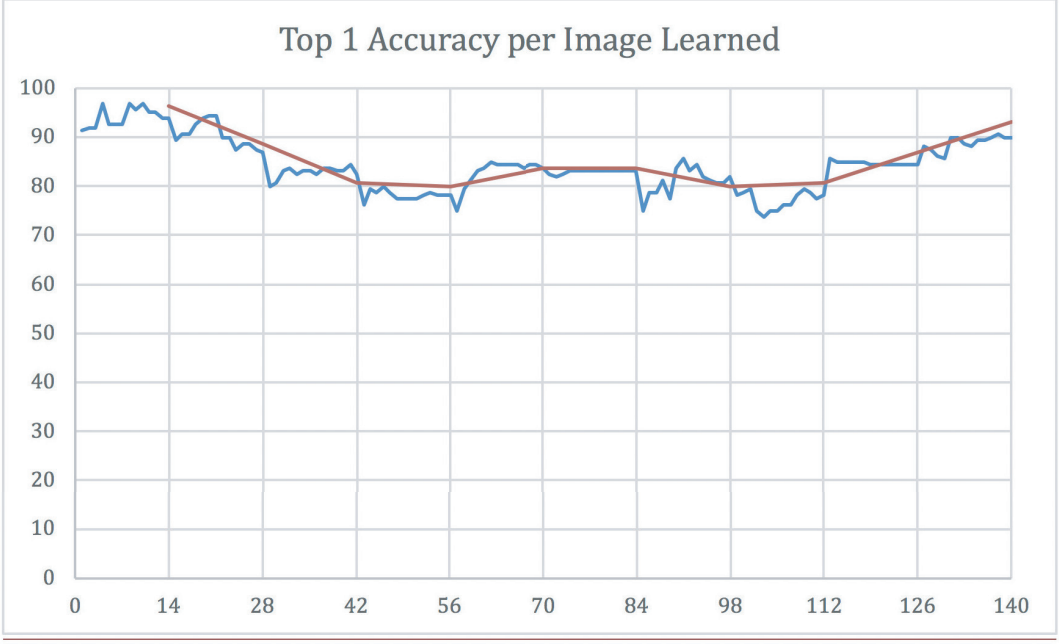


DNN: 55% DOG
L-DNN: UNKNOWN

mode when it filters out well learned object representations that do not require much learning from confusing representations that need attention. This gives the system the opportunity to define and learn the object that has been classified as "unknown," improving the system's intelligence.

The plot below emulates the real-life learning with Lifelong-DNN by feeding each of 14 training images for each of 10 classes one by one and validating after each image.

The blue line shows top 1 accuracy on the validation set after each image is trained. The vertical grid separates the training by classes (14 training images in each). The red line shows the validation results after forced consolidation that happens after each class is fully learned. Most of the times (8 out of 10), this consolidation further boosts the accuracy of the system.



CONCLUSION

While Deep Learning has grown into an industry with real world applications, its usefulness is limited by its significant shortcomings. By eliminating the problems of catastrophic forgetting and allowing transfer learning and continuous learning, Neurala's Lifelong AI brings Deep Learning to the next level. Developing memory consolidation, enabling network updates in real-time, and being adaptive to real-world situations are critical updates to traditional Deep Learning methodologies. Through merging proprietary Lifelong AI with proprietary DNN implementations, Neurala has shifted the paradigm of the recognition problem in computer vision.

ABOUT THE COMPANY

Between them, our three founders have four PhDs, 19 patents, and decades of experience working together to create groundbreaking AI. It started in a coffee shop in 2006: Max said to Anatoli, “Someday we’ll be able to run these models on a cell phone.” Anatoli promptly spit out his coffee. Just a few short years later, we created The Neurala Brain, our award-winning and patented AI technology that can run on the edge and on the lightest of devices. It’s based on advanced research work cofounders Versace, Gorshechnikov and Ames conducted for NASA, DARPA and the Air Force Research Labs. In 2013, Neurala emerged from stealth mode and joined the Techstars program to deploy our AI at a commercial scale.

Now, we work with some of the most innovative companies in the world to bring intelligence to products and devices as diverse as drones and smartphones. We are harnessing the power of the brain to find solutions to problems that change the world in meaningful ways.

Neurala’s groundbreaking technology has been hailed one of the most innovative in the world by several organizations, including Draper Ventures, CB Insights, the Edison Awards, Netexplo / UNESCO, Softech INTL, AUVSI, BostInno, and MassTLC. Visit www.neurala.com

DRAPER VENTURE NETWORK
2016 Most Innovative Company

Softech

EDISON AWARDS WINNER

Top 100 Most Disruptive Companies in the World in 2017 - Disruptor Daily

CB INSIGHTS ai.100 2018

UNESCO netexplo AWARD WINNER

DEFENSE INNOVATION AWARD 2017

CloudNOW
CloudNOW 6th annual ‘Top Women in Cloud’
Innovation Awards hosted at Google

neurala

Neurala, Inc
51 Sleeper Street, Floor 3
Boston, MA 02210

1.617.418.6161
info@neurala.com
www.neurala.com