To Appear In Artificial Intelligence in the Age of Neural Networks and Brain Computing (second edition), R. Kozma, C. Alippi, Y. Choe, and F. Morabito (Eds.), New York, 2023. Elsevier.

Evolving Deep Neural Networks

Risto Miikkulainen[1,2], Jason Liang[1], Elliot Meyerson[1], Aditya Rawal[3], Dan Fink[1], Olivier Francon[1], Bala Raju[4], Hormoz Shahrzad[1], Arshak Navruzyan[5], Nigel Duffy[6], Babak Hodjat[1]

[1]Cognizant AI Labs, [2]The University of Texas at Austin, [3]AWS AI Labs,[4]Jasper, [5]Launchpad.ai, [6]EY

Abstract

The success of deep learning depends on finding an architecture to fit the task. As deep learning has scaled up to more challenging tasks, the architectures have become difficult to design by hand. This paper proposes an automated method, CoDeepNEAT, for optimizing deep learning architectures through evolution. By extending existing neuroevolution methods to topology, components, and hyperparameters, this method achieves results comparable to the best human designs in standard benchmarks in object recognition and language modeling. It also supports building a real-world application of automated image captioning on a magazine website. Given the anticipated increases in available computing power, evolution of deep networks is a promising approach to constructing deep learning applications in the future.

1 Introduction

Large databases (i.e. Big Data) and large amounts of computing power have become readily available since the 2000s. As a result, it has become possible to scale up machine learning systems. Interestingly, not only have these systems been successful in such scaleup, but they have become more powerful. Some ideas that did not quite work before, now do, with million times more compute and data. For instance, deep learning neural networks (DNNs), i.e. convolutional neural networks (LeCun et al. 1998) and recurrent neural networks (in particular long short-term memory, or LSTM (Hochreiter and Schmidhuber 1997)), which have existed since the 1990s, have improved state-of-the-art significantly in computer vision, speech, language processing, and many other areas (Collobert and Weston 2008; Graves et al. 2013; Szegedy et al. 2016b).

As DNNs have been scaled up and improved, they have become much more complex. A new challenge has therefore emerged: How to configure such systems? Human engineers can optimize a handful of configuration parameters through experimentation, but DNNs have complex topologies and hundreds of hyperparameters. Moreover, such design choices matter; often success depends on finding the right architecture for the problem. Much of the recent work in deep learning has indeed focused on proposing different hand-designed architectures on new problems (Che et al. 2018; Devlin et al. 2019; He et al. 2016; Huang et al. 2017; Ng et al. 2015; Tan and Le 2019).

The complexity challenge is not unique to neural networks. Software and many other engineered systems have become too complex for humans to optimize fully. As a result, a new way of thinking about such design has started to emerge. In this approach, humans are responsible for the high-level design, and the details are left for computational optimization systems to figure out. For instance, humans write the overall design of a software system, and the parameters and low-level code is optimized automatically (Hoos 2012); humans write imperfect versions of programs, and evolutionary algorithms are then used to repair them (Goues

et al. 2012); humans define the space of possible web designs, and evolution is used to find effective ones (Miikkulainen et al. 2017).

This same approach can be applied to the design of DNN architectures. This problem includes three challenges: how to design the components of the architecture, how to put them together into a full network topology, and how to set the hyperparameters for the components and the global design. These three aspects need to be optimized separately for each new task.

This paper develops an approach for automatic design of DNNs. It is based on the existing neuroevolution technique of NEAT (Stanley and Miikkulainen 2002), which has been successful in evolving topologies and weights of relatively small recurrent networks in the past. In this paper, NEAT is extended to the coevolutionary optimization of components, topologies, and hyperparameters. The fitness of the evolved networks is determined based on how well they can be trained, through gradient descent, to perform in the task. The approach is demonstrated in the standard benchmark tasks of object recognition and language modeling, and in a real-world application of captioning images on a magazine website.

The results show that the approach discovers designs that are comparable to the state of the art, and does it automatically without much development effort. The approach is computationally extremely demanding—with more computational power, it is likely to be more effective and possibly surpass human design. Such power is now becoming available in various forms of cloud computing and grid computing, thereby making evolutionary optimization of neural networks a promising approach for the future.

2 Background and Related Work

Neuroevolution techniques have been applied successfully to sequential decision tasks for three decades (Floreano et al. 2008; Lehman and Miikkulainen 2013; Montana and Davis 1989; Stanley et al. 2019; Yao 1999). In such tasks there is no gradient available, so instead of gradient descent, evolution is used to optimize the weights of the neural network. Neuroevolution is a good approach in particular to POMDP (partially observable Markov decision process) problems because of recurrency: It is possible to evolve recurrent connections to allow disambiguating hidden states.

The weights can be optimized using various evolutionary techniques. Genetic algorithms are a natural choice because crossover is a good match with neural networks: they recombine parts of existing neural networks to find better ones. CMA-ES (Igel 2003), a technique for continuous optimization, works well on optimizing the weights as well because it can capture interactions between them. Other approaches such as SANE, ESP, and CoSyNE evolve partial neural networks and combine them into fully functional networks (Gomez and Miikkulainen 1997; Gomez et al. 2008; Moriarty and Miikkulainen 1997). Further, techniques such as Cellular Encoding (Gruau and Whitley 1993) and NEAT (Stanley and Miikkulainen 2002) have been developed to evolve the topology of the neural network, which is particularly effective in determining the required recurrence. Neuroevolution techniques have been shown to work well in many tasks in control, robotics, constructing intelligent agents for games, and artificial life (Lehman and Miikkulainen 2013). However, because of the large number of weights to be optimized, they are generally limited to relatively small networks.

Evolution has been combined with gradient-descent based learning in several ways, making it possible to utilize much larger networks. These methods are still usually applied to sequential decision tasks, but gradients from a related task (such as prediction of the next sensory inputs) are used to help search. Much of the work is based on utilizing the Baldwin effect, where learning only affects the selection (Hinton and Nowlan 1987). Computationally, it is possible to utilize Lamarckian evolution as well, i.e. encode the learned weight changes back into the genome (Gruau and Whitley 1993). However, care must be taken

to maintain diversity so that evolution can continue to innovate when all individuals are learning similar behavior.

Evolution of DNNs departs from this prior work in that it is applied to supervised domains where gradients are available, and evolution is used only to optimize the design of the neural network. Deep neuroevolution is thus more closely related to bilevel (or multilevel) optimization techniques (Sinha et al. 2014). The idea is to use an evolutionary optimization process at a high level to optimize the parameters of a low-level evolutionary optimization process.

Consider for instance the problem of controlling a helicopter through aileron, elevator, rudder, and rotor inputs. This is a challenging benchmark from the 2000s for which various reinforcement learning approaches have been developed (Abbeel et al. 2007; Bagnell and Schneider 2001; Ng et al. 2004). One of the most successful ones is single-level neuroevolution, where the helicopter is controlled by a neural network that is evolved through genetic algorithms (Koppejan and Whiteson 2011). The eight parameters of the neuroevolution method (such mutation and crossover rate, probability, and amount and population and elite size) are optimized by hand. It would be difficult to include more parameters because the parameters interact nonlinearly. A large part of the parameter space thus remains unexplored in the single-level neuroevolution approach. However, a bilevel approach, where a high-level evolutionary process is employed to optimize these parameters, can search this space more effectively (Liang and Miikkulainen 2015). With bilevel evolution, the number of parameters optimized could be extended to 15, which resulted in significantly better performance. In this manner, evolution was harnessed in this example task to optimize a system design that was too complex to be optimized by hand.

Recently, several studies have applied this idea to optimizing DNNs. Originally, due to significant computational demands, they focused on specific parts of the design. For instance, Loshchilov et al. (2016) used CMA-ES to optimize the hyperparameters of existing DNNs obtaining state-of-the-art results at the time on object recognition. Fernando et al. (2016) evolved a CPPN (compositional pattern-producing network (Stanley 2007)) to output the weights of an auto-encoder neural network. The autoencoder was then trained further through gradient descent, forming gradients for the CPPN training, and its trained weights were then incorporated back into the CPPN genome through Lamarckian adaptation. A related approach was proposed by Zoph and Le (2017): the topology and hyperparameters of a deep network and LSTM network were modified through policy iteration. More recently, Such et al. (2017) demonstrated the power of this approach in reinforcement learning tasks, and Real et al. (2019) in image classification.

Building on this foundation, a systematic approach to evolving DNNs is developed in this paper. First, the standard NEAT neuroevolution method is applied to the topology and hyperparameters of convolutional neural networks, and then extended to evolution of components as well, achieving results comparable to state of the art in the CIFAR-10 image classification benchmark. Second, a similar method is used to evolve the structure of LSTM networks in language modeling, showing that even small innovations in the components can have a significant effect on performance. Third, the approach is used to build a real-world application on captioning images in the website for an online magazine.

3 Evolution of Deep Learning Architectures

NEAT neuroevolution method (Stanley and Miikkulainen 2002) is first extended to evolving network topology and hyperparameters of deep neural networks in DeepNEAT, and then further to coevolution of modules and blueprints for combining them in CoDeepNEAT. The approach is tested in the standard CIFAR-10 benchmark of object recognition, and found to be comparable to the state of the art at the time the experiments were run.

3.1 Extending NEAT to Deep Networks

DeepNEAT is a most immediate extension of the standard neural network topology-evolution method NEAT to DNN. It follows the same fundamental process as NEAT: First, a population of chromosomes (each represented by a graph) with minimal complexity is created. Over generations, structure (i.e. nodes and edges) is added to the graph incrementally through mutation. During crossover, historical markings are used to determine how genes of two chromosomes can be lined up. The population is divided into species (i.e. subpopulations) based on a similarity metric. Each species grows proportionally to its fitness and evolution occurs separately in each species.

DeepNEAT differs from NEAT in that each node in the chromosome no longer represents a neuron, but a layer in a DNN. Each node contains a table of real and binary-valued hyperparameters that are mutated through uniform Gaussian distribution and random bit-flipping, respectively. These hyperparameters determine the type of a layer (such as convolutional, fully connected, or recurrent) and the properties of that layer (such as number of neurons, kernel size, and activation function). The edges in the chromosome are no longer marked with weights; instead they simply indicate how the nodes (layers) are connected together. To construct a DNN from a DeepNEAT chromosome, one simply needs to traverse the chromosome graph, replacing each node with the corresponding layer. The chromosome also contains a set of global hyperparameters applicable to the entire network (such as learning rate, training algorithm, and data preprocessing).

When arbitrary connectivity is allowed between layers, additional complexity is required. If the current layer has multiple parent layers, a merge layer must be applied to the parents in order to ensure that the parent layer's output is the same size as the current layer's input. Typically, this adjustment is done through a concatenation or element-wise sum operation. If the parent layers have mismatched output sizes, all of the parent layers must be downsampled to parent layer with the smallest output size. The specific method for downsampling is domain dependent. For example, in image classification, a max-pooling layer is inserted after specific parent layers; in image captioning, a fully connected bottleneck layer will serve this function.

During fitness evaluation, each chromosome is converted into a DNN. These DNNs are then trained for a fixed number of epochs. After training, a metric that indicates the network's performance is returned back to DeepNEAT and assigned as fitness to the corresponding chromosome in the population.

While DeepNEAT can be used to evolve DNNs, the resulting structures are often complex and unprincipled. They contrast with typical DNN architectures that utilize repetition of basic components. DeepNEAT is therefore extend to evolution of modules and blueprints next.

3.2 Cooperative Coevolution of Modules and Blueprints

Many of the most successful DNNs, such as GoogLeNet and ResNet are composed of modules that are repeated multiple times (He et al. 2016; Szegedy et al. 2016b). These modules often themselves have complicated structure with branching and merging of various layers. Inspired by this observation, a variant of DeepNEAT, called Coevolution DeepNEAT (CoDeepNEAT), is proposed. The algorithm behind CoDeepNEAT is inspired mainly by Hierarchical SANE (Moriarty and Miikkulainen 1997) but is also influenced by component-evolution approaches ESP (Gomez and Miikkulainen 1999) and CoSyNE (Gomez et al. 2008).

In CoDeepNEAT, two populations of modules and blueprints are evolved separately, using the same methods as described above for DeepNEAT. The blueprint chromosome is a graph where each node contains a pointer to a particular module species. In turn, each module chromosome is a graph that represents a small DNN. Both graphs and the nodes in them are evolved. During fitness evaluation, the modules and blueprints are combined to create a larger assembled network (Figure 1). Each node in the blueprint is replaced with a module chosen randomly from the species to which that node points. If multiple blueprint nodes point to the

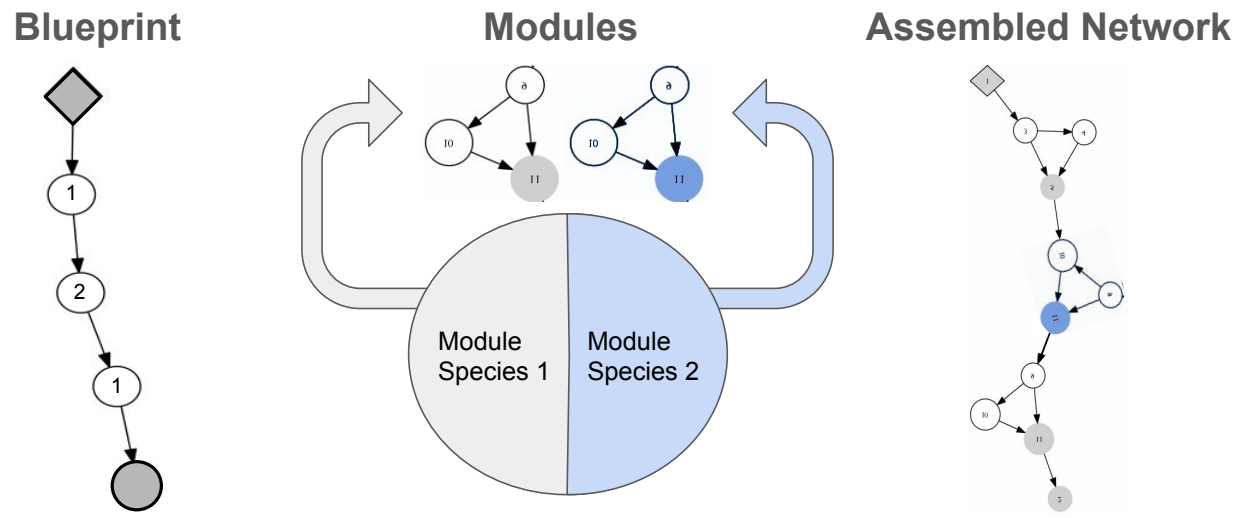

Figure 1: A visualization of how CoDeepNEAT assembles networks for fitness evaluation. Modules and blueprints are assembled together into a network through replacement of blueprint nodes with corresponding modules. This approach allows evolving repetitive and deep structures seen in many successful DNNs.

same module species, then the same module is used in all of them. The assembled networks are evaluated the a manner similar to DeepNEAT, but the fitnesses of the assembled networks are attributed back to blueprints and modules as the average fitness of all the assembled networks containing that blueprint or module.

CoDeepNEAT can evolve repetitive modular structure efficiently. Furthermore, because small mutations in the modules and blueprints often lead to large changes in the assembled network structure, CoDeepNEAT can explore more diverse and deeper architectures than DeepNEAT. An example application to the CIFAR-10 domain is presented next.

3.3 Evolving DNNs in the CIFAR-10 Benchmark

In this experiment, CoDeepNEAT was used to evolve the topology of a convolutional neural network (CNN) to maximize its classification performance on the CIFAR-10 dataset, a common image classification benchmark. The dataset consists of 50,000 training images and 10,000 testing images. The images consist of 32x32 color pixels and belong to one of 10 classes. For comparison, the neural network layer types were restricted to those used by Snoek et al. (2015) in their Bayesian optimization of CNN hyperparameters. Also following Snoek et al., data augmentation consisted of converting the images from RGB to HSV color space, adding random perturbations, distortions, and crops, and converting them back to RGB color space.

CoDeepNEAT was initialized with populations of 25 blueprints and 45 modules. From these two populations, 100 CNNs were assembled for fitness evaluation in every generation. Each node in the module chromosome represents a convolutional layer. Its hyperparameters determine the various properties of the layer and whether additional max-pooling or dropout layers are attached (Table 1). In addition, a set of global hyperparameters were evolved for the assembled network. During fitness evaluation, the 50,000 images were split into a training set of 42,500 samples and a validation set of 7,500 samples. Since training a DNN is computationally very expensive, each network was trained for eight epochs on the training set. The validation set was then used to determine classification accuracy, i.e. the fitness of the network. After 72

Node Hyperparameter	Range
Number of Filters	$[32, 256]$
Dropout Rate	$[0, 0.7]$
Initial Weight Scaling	$[0, 2.0]$
Kernel Size	$\{1, 3\}$
Max Pooling	{True, False}
Global Hyperparameter	**Range**
Learning Rate	$[0.0001, 0.1]$
Momentum	$[0.68, 0.99]$
Hue Shift	$[0, 45]$
Saturation/Value Shift	$[0, 0.5]$
Saturation/Value Scale	$[0, 0.5]$
Cropped Image Size	$[26, 32]$
Spatial Scaling	$[0, 0.3]$
Random Horizontal Flips	{True, False}
Variance Normalization	{True, False}
Nesterov Accelerated Gradient	{True, False}

Table 1: Node and global hyperparameters evolved in the CIFAR-10 domain.

generations of evolution, the best network in the population was returned.

After evolution was complete, the best network was trained on all 50,000 training images for 300 epochs, and the classification error measured. This error was 7.3%, comparable to the 6.4% error reported by Snoek et al. (2015). Interestingly, because only limited training could be done during evolution, the best network evolved by CoDeepNEAT trains very fast. While the network of Snoek et al. takes over 30 epochs to reach 20% test error and over 200 epochs to converge, the best network from evolution takes only 12 epochs to reach 20% test error and around 120 epochs to converge. This network utilizes the same modules multiple times, resulting in a deep and repetitive structure typical of many successful DNNs (Figure 2).

4 Evolution of LSTM Architectures

Recurrent neural networks, in particular those utilizing LSTM nodes, is another powerful approach to DNN. Much of the power comes from repetition of LSTM modules and the connectivity between them. In this section, CoDeepNEAT is extended with mutations that allow searching for such connectivity, and the approach is evaluated in the standard benchmark task of language modeling.

4.1 Extending CoDeepNEAT to LSTMs

Long Short Term Memory (LSTM) consists of gated memory cells that can integrate information over longer time scales (as compared to simply using recurrent connections in a neural network). LSTMs have been shown powerful in supervised sequence processing tasks such as speech recognition (Graves and Jaitly 2014) and machine translation (Bahdanau et al. 2015).

Recent research on LSTMs has focused in two directions: Finding variations of individual LSTM memory unit architecture (Bayer et al. 2009; Cho et al. 2014; Greff et al. 2017; Jozefowicz et al. 2015; Rawal and Miikkulainen 2020; Zoph and Le 2017), and discovering new ways of stitching LSTM layers into a

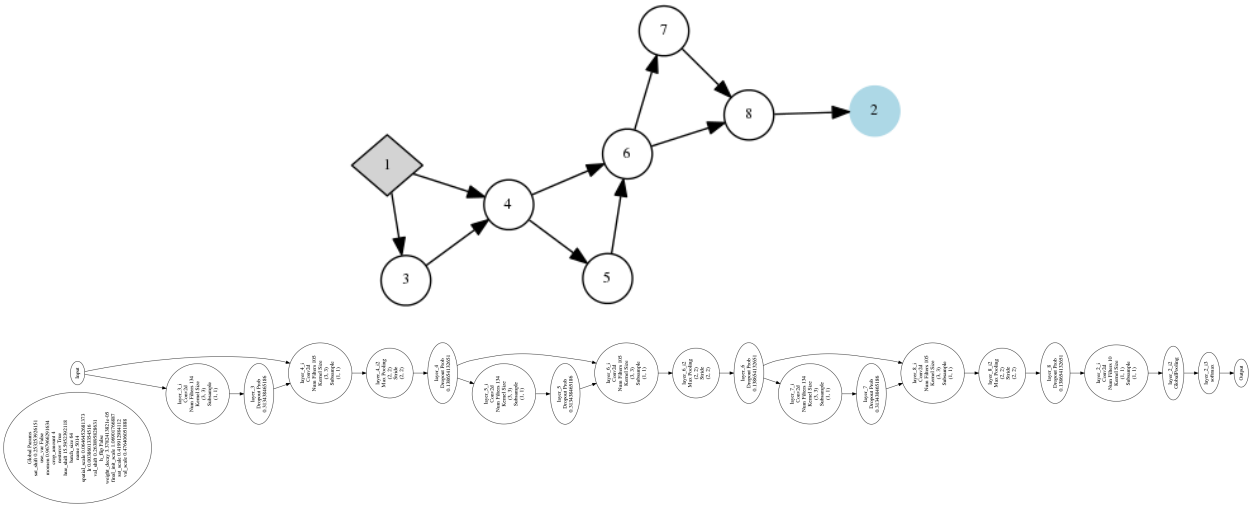

Figure 2: Top: Simplified visualization of the best network evolved by CoDeepNEAT for the CIFAR-10 domain. Node 1 is the input layer, while Node 2 is the output layer. The network has repetitive structure because its blueprint reuses same module in multiple places. Bottom: A more detailed visualization of the same network.

network (Chung et al. 2015; Kalchbrenner et al. 2015; Zilly et al. 2016). Both approaches have improved performance over vanilla LSTMs, with best recent results achieved through network design. The CoDeepNEAT method incorporates both approaches: neuroevolution searches for both new LSTM units and their connectivity across multiple layers at the same time.

CoDeepNEAT was slightly modified to make it easier to find novel connectivities between LSTM layers. Multiple LSTM layers are flattened into a neural network graph that is then modified by neuroevolution. There are two types of mutations: one enables or disables a connection between LSTM layers, and the other adds or removes skip connections between two LSTM nodes. Recently, skip connections have led to performance improvements in deep neural networks, which suggests that they could be useful for LSTM networks as well. Thus, neuroevolution modifies both the high-level network topology and the low-level LSTM connections.

In each generation, a population of these network graphs (i.e. blueprints), consisting of LSTM variants (i.e. modules with possible skip connections), is created. The individual networks are then trained and tested with the supervised data of the task. The experimental setup and the language modeling task are described next.

4.2 Evolving DNNs in the Language Modeling Benchmark

One standard benchmark task for LSTM network is language modeling, i.e. predicting the next word in a large text corpus. The benchmark utilizes the Penn Tree Bank (PTB) dataset (Marcus et al. 1993), which consists of 929k training words, 73k validation words, and 82k test words. It has 10k words in its vocabulary.

A population of 50 LSTM networks was initialized with uniformly random initial connection weights within [-0.05, 0.05]. Each network consisted of two recurrent layers (vanilla LSTM or its variants) with 650 hidden nodes in each layer. The network was unrolled in time upto 35 steps. The hidden states were initialized to zero. The final hidden states of the current minibatch was used as the initial hidden state of

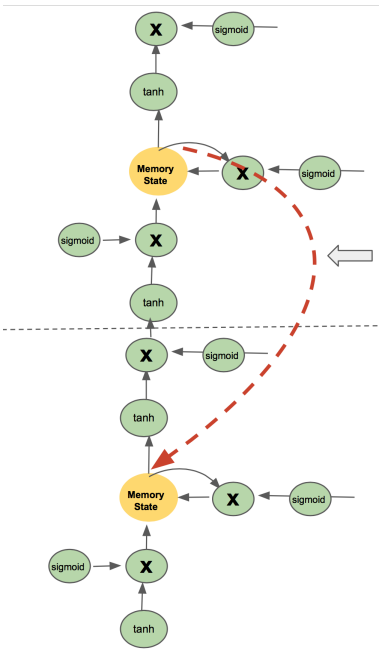

Figure 3: The best-performing LSTM variant after 25 generations of neuroevolution. It includes a novel skip connection between the two memory cells, resulting in 5% improvement over the vanilla LSTM baseline. Such improvements are difficult to discover by hand; CoDeepNEAT with LSTM-specific mutation searches for them automatically.

the subsequent minibatch (successive minibatches sequentially traverse the training set). The size of each minibatch was 20. For fitness evaluation, each network was trained for 39 epochs. A learning rate decay of 0.8 was applied at the end of every six epochs; the dropout rate was 0.5. The gradients were clipped if their maximum norm (normalized by minibatch size) exceeded 5. Training a single network took about 200 minutes on a GeForce GTX 980 GPU card.

After 25 generations of neuroevolution, the best network improved the performance on PTB dataset by 5% (test-perplexity score 78) as compared to the vanilla LSTM (Zaremba et al. 2014). As shown in Figure3, this LSTM variant consists of a feedback skip connection between the memory cells of two LSTM layers. This result is interesting because it is similar to a recent hand-designed architecture that also outperforms vanilla LSTM (Chung et al. 2015).

The initial results thus demonstrate that CoDeepNEAT with just two LSTM-specific mutations can automatically discover improved LSTM variants. It is likely that expanding the search space with more mutation types and layer and connection types would lead to further improvements.

5 Application Case Study: Image Captioning for the Blind

In a real-world case study, the vision and language capabilities of CoDeepNEAT were combined to build a real-time online image-captioning system. In this application, CoDeepNEAT searches for architectures that learn to integrate image and text representations to produce captions that blind users can access through existing screen readers. This application was implemented for a major online magazine website. Evolved networks were trained with the open-source MSCOCO image-captioning dataset (Chen et al. 2015), along

Global Hyperparameter	**Range**
Learning Rate	$[0.0001, 0.1]$
Momentum	$[0.68, 0.99]$
Shared Embedding Size	$[128, 512]$
Embedding Dropout	$[0, 0.7]$
LSTM Recurrent Dropout	{True, False}
Nesterov Momentum	{True, False}
Weight Initialization	{Glorot normal, He normal}
Node Hyperparameter	**Range**
Layer Type	{Dense, LSTM}
Merge Method	{Sum, Concat}
Layer Size	$\{128, 256\}$
Layer Activation	{ReLU, Linear}
Layer Dropout	$[0, 0.7]$

Table 2: Node and global hyperparameters evolved for the image captioning case study.

with a new dataset collected for this website.

5.1 Evolving DNNs for Image Captioning

Deep learning has recently provided state-of-the-art performance in image captioning, and several diverse architectures have been suggested (Karpathy and Fei-Fei 2015; Vedantam et al. 2017; Vinyals et al. 2015; Xu et al. 2015; You et al. 2016). The input to an image-captioning system is a raw image, and the output is a text caption intended to describe the contents of the image. In deep learning approaches, a convolutional network is usually used to process the image, and recurrent units, often LSTMs, to generate coherent sentences with long-range dependencies.

As is common in existing approaches, the evolved system uses a pretrained ImageNet model (Szegedy et al. 2016b) to produce initial image embeddings. The evolved network takes an image embedding as input, along with a one-hot text input. As usual, in training the text input contains the previous word of the ground truth caption; in inference it contains the previous word generated by the model (Karpathy and Fei-Fei 2015; Vinyals et al. 2015).

In the initial CoDeepNEAT population the image and text inputs are fed to a shared embedding layer, which is densely connected to a softmax output over words. From this simple starting point, CoDeepNEAT evolves architectures that include fully connected layers, LSTM layers, sum layers, concatenation layers, and sets of hyperparameters associated with each layer, along with a set of global hyperparameters (Table 2). In particular, the well-known Show and Tell image-captioning architecture (Vinyals et al. 2015) is in this search space, providing a baseline with which evolution results can be compared. These components and the glue that connects them are evolved as described in Section 3.2, with 100 networks trained in each generation. Since there is no single best accepted metric for evaluating captions, the fitness function is the mean across three metrics (BLEU, METEOR, and CIDEr; (Chen et al. 2015)) normalized by their baseline values. Fitness is computed over a holdout set of 5000 images, i.e. 25,000 image-caption pairs.

To keep the computational cost reasonable, during evolution the networks are trained for only six epochs, and only with a random 100,000 image subset of the 500,000 MSCOCO image-caption pairs. As a result, there is evolutionary pressure towards networks that converge quickly: The best resulting architectures train to near convergence 6 times faster than the baseline Show and Tell model (Vinyals et al. 2015). After

Figure 4: An iconic image from an online magazine captioned by an evolved model. The model provides a suitably detailed description without any unnecessary context.

evolution, the optimized learning rate is scaled by one-fifth to compensate for the subsampling.

5.2 Building the Application

The images in MSCOCO are chosen to depict "common objects in context". The focus is on a relatively small set of objects and their interactions in a relatively small set of settings. The internet as a whole, and the online magazine website in particular, contain many images that cannot be classified as "common objects in context". Other types of images from the magazine include staged portraits of people, infographics, cartoons, abstract designs, and *iconic* images, i.e. images of one or multiple objects *out of context* such as on a white or patterned background. Therefore, an additional dataset of 17,000 image-caption pairs was constructed for the case study, targeting iconic images in particular. Four thousand images were first scraped from the magazine website, and 1000 of them were identified as iconic. Then, 16,000 images that were visually similar to those 1000 were retrieved automatically from a large image repository. A single ground truth caption for each of these 17K images was generated by human subjects through MightyAI[1]. The holdout set for evaluation consisted of 100 of the original 1000 iconic images, along with 3000 other images.

During evolution, networks were trained and evaluated only on the MSCOCO data. The best architecture from evolution was then trained from scratch on both MSCOCO and MightyAI datasets in an iterative alternating approach: one epoch on MSCOCO, followed by five epochs on MightyAI, until maximum performance was reached on the MightyAI holdout data. Beam search was then used to generate captions from the fully trained models. Performance achieved using the MightyAI data demonstrates the ability of evolved architectures to generalize to domains towards which they were not evolved.

Once the model was fully-trained, it was placed on a server where it can be queried with images to caption. A JavaScript snippet was written that a developer can embed in his/her site to automatically query the model to caption all images on a page. This snippet runs in an existing Chrome extension for custom scripts and automatically captions images as the user browses the web. These tools add captions to the 'alt' field of images, which screen readers can then read to blind internet users (Figure 4).

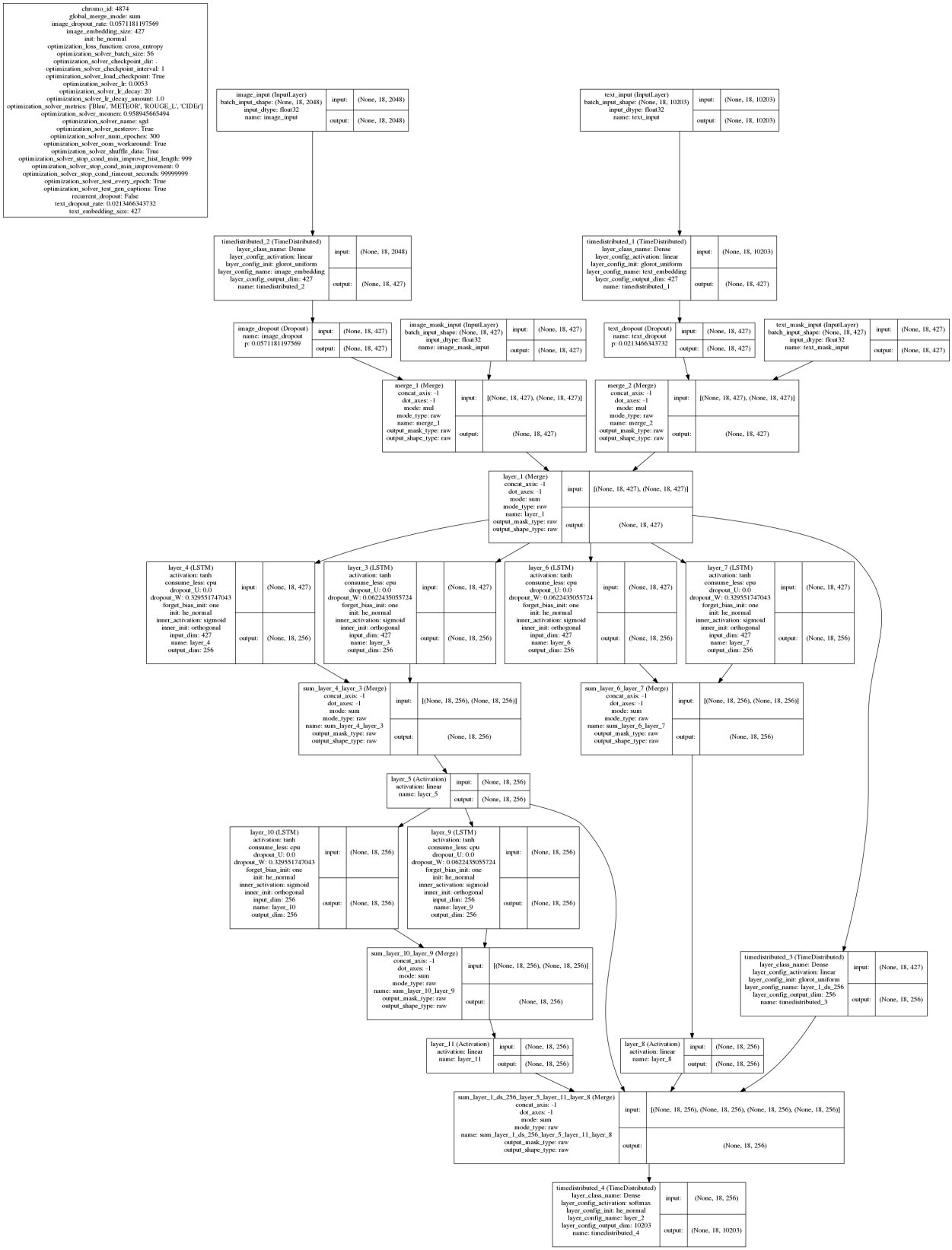

Figure 5: The most fit architecture found by evolution. It consists of three LSTM-based modules as well as a parallel densely connected pathway. Multiple pathways are robust during search and can represent multiple hypotheses during performance. Such a high-level design is different from most hand-designed architectures and represents an innovation of evolutionary search.

Model	BLEU-4	CIDEr	METEOR
DNGO (Snoek et al. 2015)	26.7	—	—
Baseline (Vinyals et al. 2015)	27.7	85.5	23.7
Evolved	**29.1**	**88.0**	**23.8**

Table 3: The evolved network improves over the hand-designed baseline when trained on MSCOCO alone.

5.3 Image-Captioning Results

Trained in parallel on about 100 GPUs, each generation took around one hour to complete. The most fit architecture was discovered on generation 37 (Figure 5). Among the components in its unique structure are six LSTM layers, four summing merge layers, and several skip connections. A single module consisting of two LSTM layers merged by a sum is repeated three times. There is a path from the input through dense layers to the output that bypasses all LSTM layers, providing the softmax with a more direct view of the current input. The motif of skip connections with a summing merge is similar to residual architectures that are currently popular in deep learning (He et al. 2016; Szegedy et al. 2016a). However, the high-level structure of multiple parallel pathways is an innovation of evolutionary search and is rarely seen in human-designed architectures. During search, such parallel pathways provide robust performance as the architecture varies, and are thus likely to be discovered. During performance, they possibly represent multiple parallel hypotheses that are combined to the most likely final answer. This architecture performs better than the hand-tuned baseline (Vinyals et al. 2015) when trained on the MSCOCO data alone (Table 3).

However, a more important result is the performance of this network on the magazine website. Because no suitable automatic metrics exist for the types of captions collected for the magazine website (and existing metrics are very noisy when there is only one reference caption), captions generated by the evolved model on all 3100 holdout images were manually evaluated as correct, mostly-correct, mostly-incorrect, and incorrect (Figure 6). Figure 7 shows some examples of good and bad captions for these images.

The model is not perfect, but the results are promising. The evolved network is correct or mostly correct on 63% of iconic images and 31% of all images. There are many known improvements that can be implemented, including ensembling diverse architectures generated by evolution, fine-tuning of the ImageNet model, using a more recent ImageNet model, and performing beam search or scheduled sampling during training (Vinyals et al. 2016); preliminary experiments with ensembling alone suggest improvements of about 20%. For this application, it is also important to include methods for automatically evaluation caption quality and filtering captions that would give an incorrect impression to a blind user. However, even without these additions, the results demonstrate that it is now possible to develop practical applications through evolving DNNs.

6 Discussion and Future Work

The results in this paper show that the evolutionary approach to optimizing deep neural networks is feasible: The results are comparable to hand-designed architectures in benchmark tasks, and it is possible to build real-world applications based on the approach. It is important to note that the approach has not yet been pushed to its full potential. It takes a couple of days to train each deep neural network on a typical GPU, and over the course of evolution, thousands of them need to be trained. Therefore, the results are limited by the available computational power.

[1] https://mty.ai/computer-vision/

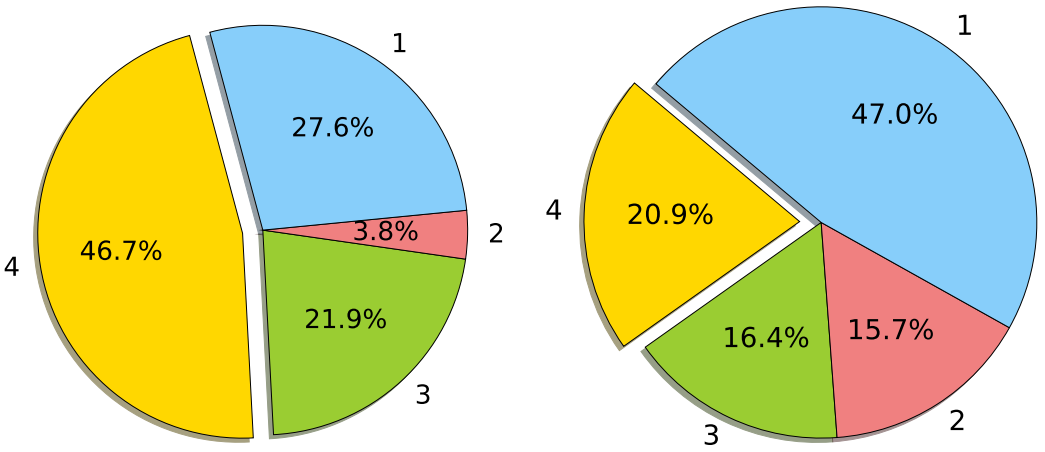

Figure 6: Results for captions generated by an evolved model for the online magazine images rated from 1 to 4, with 4=Correct, 3=Mostly Correct, 2=Mostly Incorrect, 1=Incorrect. Left: On iconic images, the model is able to get about one half correct; Right: On all images, the model gets about one fifth correct. The superior performance on iconic images shows that it is useful to build supplementary training sets for specific image types.

Interestingly, since it was necessary to train networks only partially during evolution, evolution is biased towards discovering fast learners instead of top performers. This is an interesting result on its own: Evolution can be guided with goals other than simply accuracy, including training time, execution time, or memory requirements of the network. On the other hand, if it was possible to train the networks further, it would be possible to identify top performers more reliably, and final performance would likely improve. In order to speed up evolution, other techniques may be developed. For instance, it may be possible to seed the population with various state-of-the-art architectures and modules, instead of having to re-discover them during evolution.

Significantly more computational resources are likely to become available in the near future. Cloud-based services offer GPU computation with a reasonable cost, and they are becoming cheaper. Not many approaches can take advantage of such power, but evolution of deep learning neural networks can. The search space of different components and topologies can be extended, and more hyperparameters be optimized. While cloud services such as AutoML provide some such functionality, and benefit from the increasing computing power as well, they currently represent brute-force approaches that are unlikely to scale as well. Given the results in this paper, the evolutionary approach is likely to discover designs that are superior to those that can be developed by hand today; it is also likely to make it possible to apply deep learning to a wider array of tasks and applications in the future.

7 Conclusion

Evolutionary optimization makes it possible to design more complex deep learning architectures than can be designed by hand. The topology, components, and hyperparameters of the architecture can all be optimized simultaneously to fit the requirements of the task, resulting in superior performance. Such automated design can make new applications of deep learning possible in vision, speech, language, and other areas. The approach is computationally demanding, and thus offers one way to put the anticipated increases in computing power to good use.

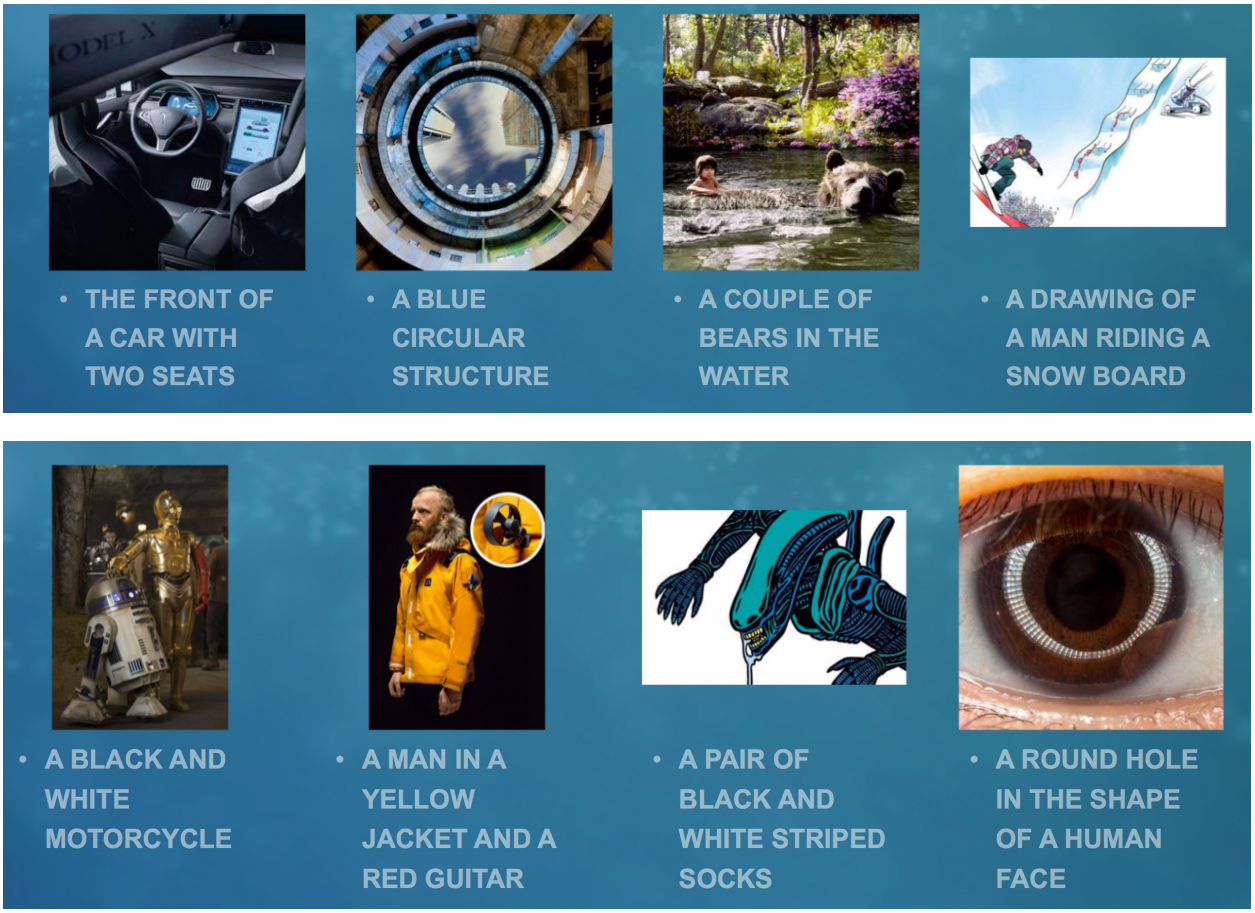

Figure 7: Top: Four good captions. The model is able to abstract about ambiguous images and even describe drawings, along with photos of objects in context. Bottom: Four bad captions. When it fails, the output of the model still contains some correct sense of the image.

References

Abbeel, P., Coates, A., Quigley, M., and Ng, A. Y. (2007). An application of reinforcement learning to aerobatic helicopter flight. In *Advances in Neural Information Processing Systems 19*.

Bagnell, J., and Schneider, J. (2001). Autonomous helicopter control using reinforcement learning policy search methods. In *Proceedings of the International Conference on Robotics and Automation 2001*. IEEE.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *In ICLR*.

Bayer, J., Wierstra, D., Togelius, J., and Schmidhuber, J. (2009). Evolving memory cell structures for sequence learning. In *In Artificial Neural Networks ICANN*, 755–764.

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8:6085.

Chen, X., Fang, H., Lin, T. Y., Vedantam, R., Gupta, S., Dollar, P., and Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. *arXiv:1504.00325*.

Cho, K., van Merrienboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2015). Gated feedback recurrent neural networks. In *Proceedings of the 32nd International Conference on Machine Learning*.

Collobert, R., and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, 160–167. ACM.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Fernando, C., Banarse, D., Besse, F., Jaderberg, M., Pfau, D., Reynolds, M., Lactot, M., and Wierstra, D. (2016). Convolution by evolution: Differentiable pattern producing networks. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2016)*. New York, NY: ACM.

Floreano, D., Dürr, P., and Mattiussi, C. (2008). Neuroevolution: From architectures to learning. *Evolutionary Intelligence*, 1:47–62.

Gomez, F., and Miikkulainen, R. (1997). Incremental evolution of complex general behavior. *Adaptive Behavior*, 5:317–342.

Gomez, F., and Miikkulainen, R. (1999). Solving non-Markovian control tasks with neuroevolution. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, 1356–1361. San Francisco: Morgan Kaufmann.

Gomez, F., Schmidhuber, J., and Miikkulainen, R. (2008). Accelerated neural evolution through cooperatively coevolved synapses. *Journal of Machine Learning Research*, 9:937–965.

Goues, C. L., Nguyen, T., Forrest, S., and Weimer, W. (2012). Genprog: Automatic bug correction in real programs. *ACM Transactions on Software Engineering*, 38.

Graves, A., and Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In *In Proc. 31st ICML*, 1764–1772.

Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645–6649. IEEE.

Greff, K., Srivastava, R. K., Koutnk, J., Steunebrink, B. R., and Schmidhuber, J. (2017). LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28:222–2232.

Gruau, F., and Whitley, D. (1993). Adding learning to the cellular development of neural networks: Evolution and the Baldwin effect. *Evolutionary Computation*, 1:213–233.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hinton, G. E., and Nowlan, S. J. (1987). How learning can guide evolution. *Complex Systems*, 1:495–502.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9:1735–1780.

Hoos, H. (2012). Programming by optimization. *Communications of the ACM*, 55:70–80.

Huang, G., Liu, Z., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Igel, C. (2003). Neuroevolution for reinforcement learning using evolution strategies. In *Proceedings of the 2003 Congress on Evolutionary Computation*, 2588–2595. Piscataway, NJ: IEEE Press.

Jozefowicz, R., Zaremba, W., and Sutskever, I. (2015). An empirical exploration of recurrent network architectures. In *In Proceedings of the 32nd International Conference on Machine Learning*, 2342–2350.

Kalchbrenner, N., Danihelka, I., and Graves, A. (2015). Grid long short-term memory. *CoRR*, abs/1507.01526.

Karpathy, A., and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proc. of CVPR*, 3128–3137.

Koppejan, R., and Whiteson, S. (2011). Neuroevolutionary reinforcement learning for generalized control of simulated helicopters. *Evolutionary Intelligence*, 4:219–241.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:22782324.

Lehman, J., and Miikkulainen, R. (2013). Neuroevolution. *Scholarpedia*, 8(6):30977.

Liang, J. Z., and Miikkulainen, R. (2015). Evolutionary bilevel optimization for complex control tasks. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2015)*. Madrid, Spain.

Loshchilov, I., and Hutter, F. (2016). CMA-ES for hyperparameter optimization of deep neural networks. In *International Conference on Learning Representations (ICLR-2016) Workshop Track*.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2).

Miikkulainen, R., Iscoe, N., Shagrin, A., Cordell, R., Nazari, S., Schoolland, C., Brundage, M., Epstein, J., Dean, R., and Lamba, G. (2017). Conversion rate optimization through evolutionary computation. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2017)*, 1193–1199. New York, NY: ACM.

Montana, D. J., and Davis, L. (1989). Training feedforward neural networks using genetic algorithms. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, 762–767. San Francisco: Morgan Kaufmann.

Moriarty, D. E., and Miikkulainen, R. (1997). Forming neural networks through efficient and adaptive co-evolution. *Evolutionary Computation*, 5:373–399.

Ng, A. Y., Kim, H. J., Jordan, M., and Sastry, S. (2004). Autonomous helicopter flight via reinforcement learning. In *Advances in Neural Information Processing Systems 16*.

Ng, J. Y., Hausknecht, M. J., Vijayanarasimhan, S., Vinyals, O., Monga, R., and Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4694–4702.

Rawal, A., and Miikkulainen, R. (2020). Discovering gated recurrent neural network architectures. In Iba, H., and Noman, N., editors, *Deep Neural Evolution Deep Learning with Evolutionary Computation*. Springer.

Real, E., Aggarwal, A., Huang, Y., and Le, Q. V. (2019). Regularized evolution for image classifier architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI.

Sinha, A., Malo, P., Xu, P., and Deb, K. (2014). A bilevel optimization approach to automated parameter tuning. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2014)*. Vancouver, BC, Canada.

Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M., Prabhat, M., and Adams, R. P. (2015). Scalable bayesian optimization using deep neural networks. In *Proc. of ICML*, 2171–2180.

Stanley, K. (2007). Compositional pattern producing networks: A novel abstraction of development. *Genetic Programming and Evolvable Machines*, 8(2):131–162.

Stanley, K. O., Clune, J., Lehman, J., and Miikkulainen, R. (2019). Designing neural networks through evolutionary algorithms. *Nature Machine Intelligence*, 1:2435.

Stanley, K. O., and Miikkulainen, R. (2002). Evolving Neural Networks Through Augmenting Topologies. *Evolutionary Computation*, 10:99–127.

Such, F. P., Madhavan, V., Conti, E., Lehman, J., Stanley, K. O., and Clune, J. (2017). Deep neuroevolution: Genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning. *arXiv:1712.06567*.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. (2016a). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., , and Wojna, Z. (2016b). Rethinking the inception architecture for computer vision. In *Proc. of CVPR*, 2818–2826.

Tan, M., and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 6105–6114.

Vedantam, R., Bengio, S., Murphy, K., Parikh, D., and Chechik, G. (2017). Context-aware captions from context-agnostic supervision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3156–3164.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2016). Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *Trans. on Pattern Analysis and Machine Intelligence*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salkhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proc. of ICML*, 77–81.

Yao, X. (1999). Evolving artificial neural networks. *Proceedings of the IEEE*, 87(9):1423–1447.

You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. (2016). Image captioning with semantic attention. In *Proc. of CVPR*, 4651–4659.

Zaremba, W., Sutskever, I., and Vinyals, O. (2014). Recurrent neural network regularization. *arXiv:1409.2329*.

Zilly, J. G., Srivastava, R. K., Koutník, J., and Schmidhuber, J. (2016). Recurrent highway networks. *CoRR*, abs/1607.03474.

Zoph, B., and Le, Q. V. (2017). Neural architecture search with reinforcement learning. In *International Conference on Learning Representations*.