Taking inspiration from biological evolution, we explore the idea of “Can deep neural networks evolve naturally over successive generations into highly efficient deep neural networks?” by introducing the notion of synthesizing new highly efficient, yet powerful deep neural networks over successive generations via an evolutionary process from ancestor deep neural networks.

The architectural traits of ancestor deep neural networks are encoded using synaptic probability models, which can be viewed as the ‘DNA’ of these networks. New descendant networks with differing network architectures are synthesized based on these synaptic probability models from the ancestor networks and computational environmental factor models, in a random manner to mimic heredity, natural selection, and random mutation. These offspring networks are then trained into fully functional networks, like one would train a newborn, and have more efficient, more diverse network architectures than their ancestor networks, while achieving powerful modeling capabilities. Experimental results for the task of visual saliency demonstrated that the synthesized ‘evolved’ offspring networks can achieve state-of-the-art performance while having network architectures that are significantly more efficient (with a staggering ~48-fold decrease in synapses by the fourth generation) compared to the original ancestor network.
Deep learning, especially deep neural networks [1–4] have shown considerable promise through tremendous results in recent years, significantly improving the accuracy of a variety of challenging problems when compared to other machine learning methods [5–10]. However, deep neural networks require high performance computing systems due to the tremendous quantity of computational layers they possess, leading to a massive quantity of parameters to learn and compute. This issue of architectural complexity has increased greatly in recent years [7,11,12], driven by the demand for increasingly deeper and larger deep neural networks to boost modeling accuracy. As such, it has become increasingly more difficult to take advantage of such complex deep neural networks in scenarios where computational and energy resources are scarce.

To enable the widespread use of deep learning, there has been a recent drive towards obtaining highly-efficient deep neural networks with strong modeling power. Much of the work in obtaining efficient deep neural networks have focused on deterministically compressing trained deep neural networks [13–17]. Rather than attempting to take an existing deep neural network and compress it into a smaller representation heuristically, we instead consider a radically different idea: Can deep neural networks evolve naturally over successive generations into highly efficient deep neural networks? The ability to mimic the biological evolutionary process for the task of producing highly-efficient deep neural networks over successive generations can have considerable benefits. As such, we entertain a radically different notion for producing highly-efficient deep neural networks by introducing the evolutionary synthesis of deep neural networks over successive generations based on ancestor deep neural networks.

The proposed evolutionary synthesis of deep neural networks is primarily inspired by real biological evolutionary mechanisms. In nature, traits that are passed down from generation to generation through DNA may change over successive generations due to factors such as natural selection and random mutation, giving rise to diversity and enhanced traits in later generations. To realize the idea of evolutionary synthesis for producing deep neural networks, we introduce a number of computational constructs to mimic the following mechanisms of biological evolution: i) Heredity, ii) Natural Selection, and iii) Random Mutation.

Heredity. Here, we mimic the idea of heredity by encoding the architectural traits of deep neural networks
in the form of synaptic probability models, which are used to pass down traits from generation to generation. One can view these synaptic probability models as the ‘DNA’ of the networks. Let $H = (N, S)$ denote the possible architecture of a deep neural network, with $N$ denoting the set of possible neurons and $S$ denoting the set of possible synapses, with $s_k \in S$ denoting a synapse between two neurons $(n_i, n_j) \in \mathcal{N}$. One can encode the architectural traits of a deep neural network as $P(H_g|H_{g-1})$, which denotes the conditional probability of the architecture of a network in generation $g$ (denoted by $H_g$), given the architecture of its ancestor network in generation $g-1$ (denoted by $H_{g-1}$).

If we were to treat areas of strong synapses in an ancestor network as desirable traits to be inherited by descendant networks, where descendant networks have a higher probability of having similar areas of strong synapses as its ancestor network, one can instead encode the architectural traits of a deep neural network as the synaptic probability $P(S_g|W_{g-1})$, where $w_k \in W$ encodes the synaptic strength of each synapse $s_k$. Modeling $P(S_g|W_{g-1})$ as an exponential distribution, with the probability of each synapse in the network assumed to be independently distributed, one arrives at

$$P(S_{g,i}|W_{g-1,i}) = \prod_i \exp \left( \frac{w_{g-1,i}}{Z} - 1 \right),$$

where $Z$ is a normalization constant.

**Natural Selection and Random Mutation.** The ideas of natural selection and random mutation are mimicked through the introduction of a network synthesis process for synthesizing descendant networks, which takes into account not only the synaptic probability model encoding the architectural traits of the ancestor network, but also an environmental factor model to mimic the environmental conditions that help drive natural selection, in a random manner that drives random mutation. More specifically, a synapse is synthesized randomly between two possible neurons in a descendant network based on $P(H_g|W_{g-1})$ and an environmental factor model $\mathcal{F}(\mathcal{E})$, with the neurons in the descendant network synthesized subsequently based on the set of synthesized synapses. As such, the architecture of a descendant network at generation $g$ can be synthesized randomly via synthesis
probability $P(H_g)$, which can be expressed by

$$P(H_g) = \mathcal{F}(\mathcal{E}) \cdot P(S_g|W_g-1).$$  \hfill (2)

The environmental factor model $\mathcal{F}(\mathcal{E})$ can be the combination of quantitative environmental conditions that are imposed upon the descendant networks that they must adapt to.

To have a better intuitive understanding, let us examine an illustrative example of how one can impose environmental conditions using $\mathcal{F}(\mathcal{E})$ to promote the evolution of highly efficient deep neural networks.

Efficiency-driven Evolutionary Synthesis. One of the main environmental factors in encouraging energy efficiency during evolution is to restrict the resources available. For example, in a study by Moran et al. [18], it was proposed that the eyeless Mexican cavefish lost its vision system over generations due to the high energetic cost of neural tissue and low food availability in subterranean habitats. Their study demonstrated that the cost of vision is about 15% of resting metabolism for a 1-g eyed phenotype, thus losing their vision system through evolution has significant energy savings and thus improves survivability. As such, we are inspired to computationally restrict resources available to descendant networks to encourage the evolution of highly-efficient deep neural networks.

Considering the aforementioned example, the descendant networks must take on network architectures with more efficient energy consumption than this original ancestor network to be able to survive. The main factor in energy consumption is the quantity of synapses and neurons in the network. Therefore, to mimic environmental constraints that encourage the evolution of highly-efficient deep neural networks, we introduce an environmental constraint $\mathcal{F}(\mathcal{E}) = C$ that probabilistically constrains the quantity of synapses that can be synthesized in the descendant network (which in effect also constrains the quantity of neurons that can be synthesized), such that descendant networks are forced to evolve more efficient network architectures than their ancestor networks.

Therefore, given $P(S_g|W_{g-1})$ and $\mathcal{F}(\mathcal{E}) = C$, the synthesis probability $P(H_g)$ can be formulated as

$$P(H_g) = C \cdot P(S_g|W_{g-1}),$$  \hfill (3)

where $C$ is the highest percentage of synapses desired in the descendant network. The random element of the network synthesis process mimics the random mutation process and promotes network architectural diversity.
The synthesized descendant networks are then trained into fully-functional networks, like one would train a newborn, and the evolutionary synthesis process is repeated for producing successive generations of descendant networks. An overview of the proposed evolutionary synthesis of highly-efficient deep neural networks is demonstrated in Figure 1.

**Experiments.** To investigate the efficacy of the proposed evolutionary synthesis of highly-efficient deep neural networks, experiments were performed using the MSRA-B [19] and HKU-IS datasets [20] for the task of visual saliency. This task was chosen given the importance for biological beings to detect objects of interest (e.g., prey, food, predators) for survival in complex visual environments, and can provide interesting insights into the evolution of networks. Three generations of descendant deep neural networks (second, third, and fourth generations) were synthesized within an artificially constrained environment beyond the original, first-generation ancestor network. The environmental constraint imposed during synthesis in this study is that the descendant networks should not have more than 40% of the total number of synapses that its direct ancestor network possesses (i.e., $C = 0.4$), thus encouraging the evolution of highly-efficient deep neural networks. The network architecture of the original, first generation ancestor network used in this study, and details on the tested datasets and performance metrics are provided in Supplementary Materials.

**Results and Discussion.** Figure 2 shows the $F_\beta$ vs architectural efficiency of the descendant networks for MSRA-B and HKU-IS datasets for the task of visual saliency (see Supplementary Materials for full numerical MAE, $F_\beta$ score, number of synapses, and architectural efficiency results). First, it can be observed that the performance differences from one generation of descendant networks to the next generation are small for MSRA-B (<3% between first generation and the fourth generation), while the performance differences are small for HKU-IS between the first two generations (<0.5%) before larger performance differences in the third and fourth generations (<8% between the first and fourth generations). These results indicate that the modeling power of the ancestor network are well-preserved in the descendant networks.

Second, it can be observed that the descendant networks in the second and third generations can achieve state-of-the-art $F_\beta$ scores for MSRA-B (0.876 at second generation and 0.861 at third generation, compared to 0.865 as
reported by Li et al. [20] for their state-of-the art visual saliency method), while having network architectures that are significantly more efficient compared to the first generation ancestor network (~18-fold decrease in synapses). A similar trend was observed for HKU-IS, though persisting only in the second generation (0.826 compared to 0.8 reported in [20], while achieving a ~4-fold decrease in synapses over ancestor network). What is more remarkable is that the descendant network at the fourth generation maintains strong Fβ scores (0.850 for MSRA-B and 0.753 for HKU-IS), while having network architectures that are incredibly efficient (~48-fold decrease in synapses) compared to the first generation ancestor network. This ~48-fold increase in architectural efficiency while maintaining modeling power clearly show the efficacy of producing highly-efficient deep neural networks over successive generations via the proposed evolutionary synthesis.

To gain additional insights, Figure 3 show example visual saliency maps generated by the descendant networks at different generations. It can be observed that the descendant networks at all generations consistently identified the objects of interest in the scene as visually salient. It is also interesting to observe that by the fourth generation, with a ~48-fold decrease in synapses compared to the first generation ancestor network, the ability to distinguish fine-grained visual saliency starts to diminish. These observations are interesting in that, similar to biological evolution, they show that the descendant networks evolved over successive generations in such a way that important traits (e.g., general ability to identify salient objects) are retained from its ancestors while less important traits (e.g., ability to distinguish fine-grained saliency) diminish in favor of adapting to environmental constraints (e.g., growing highly-efficient architectures due to imposed constraints).

These experimental results show that, by taking inspiration from biological evolution, the proposed evolutionary synthesis of deep neural networks can lead to the natural evolution of deep neural networks over successive generations into highly efficient, yet powerful deep neural networks, and thus a promising direction for future exploration in deep learning.

References and Notes

1. Y. LeCun, Y. Bengio, G. Hinton, Nature (2015).
2. A. Graves, A.-r. Mohamed, G. Hinton, *IEEE international conference on acoustics, speech and signal processing* (2013), pp. 6645–6649.

3. Y. Bengio, *Foundations and trends® in Machine Learning* 2, 1 (2009).

4. J. J. Tompson, A. Jain, Y. LeCun, C. Bregler, *Advances in neural information processing systems* (2014), pp. 1799–1807.

5. A. Krizhevsky, I. Sutskever, G. E. Hinton, *Advances in neural information processing systems (NIPS)* (2012).

6. C. Farabet, C. Couprie, L. Najman, Y. LeCun, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2013).

7. K. Simonyan, A. Zisserman, *CoRR, abs/1409.1556* (2014).

8. G. Hinton, et al., *IEEE Signal Processing Magazine* (2012).

9. A. Hannun, et al., *CoRR, abs/1412.5567* (2014).

10. D. Amodei, et al., *CoRR, abs/1512.02595* (2015).

11. R. K. Srivastava, K. Greff, J. Schmidhuber, *Advances in Neural Information Processing Systems (NIPS)* (2015), pp. 2368–2376.

12. C. Szegedy, et al., *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 1–9.

13. Y. LeCun, J. S. Denker, S. A. Solla, R. E. Howard, L. D. Jackel, *Advances in Neural Information Processing Systems (NIPS)* (1989).

14. Y. Gong, L. Liu, M. Yang, L. Bourdev, *CoRR, abs/1412.6115* (2014).

15. S. Han, H. Mao, W. J. Dally, *CoRR, abs/1510.00149* (2015).

16. S. Han, J. Pool, J. Tran, W. Dally, *Advances in Neural Information Processing Systems (NIPS)* (2015).
17. W. Chen, J. T. Wilson, S. Tyree, K. Q. Weinberger, Y. Chen, CoRR, abs/1504.04788 (2015).

18. D. Moran, R. Softley, E. J. Warrant, Science advances (2015).

19. T. Liu, et al., IEEE Transactions on Pattern Analysis and Machine Intelligence (2011).

20. G. Li, Y. Yu, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015).

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada, Canada Research Chairs Program, the Ontario Ministry of Research and Innovation, and the Nvidia Hardware Grant Program.

Author Contributions

A.W. conceived the concept of evolutionary synthesis of deep neural networks proposed in this study. M.S. and A.W. formulated the evolutionary synthesis process proposed in this study. A.M. implemented the evolutionary synthesis process and performed all experiments in this study. AW and MS analyzed the experimental results. A.W., M.S., and A.M. all participated in writing this paper.

Competing Interests

The authors declare that they have no competing interests.

Data and Material Availability

All data related to this paper may be requested from the authors by emailing either mjshafiee@uwaterloo.ca or alexander.wong@uwaterloo.ca
Figure 1: Evolutionary synthesis process of highly-efficient deep neural networks.

Figure 2: $F_\beta$ vs architectural efficiency of the descendant deep neural networks for MSRA-B and HKU-IS datasets. 'Gx' denotes the descendant neural network at generation $x$.

Figure 3: Example test images from the tested datasets, and the corresponding visual saliency maps generated by the descendant deep neural networks at different generations.
Supplementary Material

Network architecture. The network architecture of the original, first generation ancestor network used in this study builds upon the VGG16 very deep convolutional neural network architecture \cite{7} for the purpose of image segmentation as follows. The outputs of the c3, c4, and c5 stacks from the VGG16 architecture are fed into newly added c6, c7, c8 stacks, respectively. The output of the c7 and c8 stacks are then fed into d1 and d2 stacks. The concatenated outputs of the c6, d1, and d2 stacks are then fed into the c9 stack. The output of the c5 stack is fed into c10 and c11 stacks. Finally, the combined output of the c9, c10 and c11 stacks are fed into a softmax layer to produce final segmentation result. The details of different stacks are as follows: c1: 2 convolutional layers of 64, \(3 \times 3\) local receptive fields, c2: 2 convolutional layers of 128, \(3 \times 3\) local receptive fields, c3: 3 convolutional layers of 256, \(3 \times 3\) local receptive fields, c4: 3 convolutional layers of 512, \(3 \times 3\) local receptive fields, c5: 3 convolutional layers of 512, \(3 \times 3\) local receptive fields, c6: 1 convolutional layers of 256, \(3 \times 3\) local receptive fields, c7 and c8: 1 convolutional layers of 512, \(3 \times 3\) local receptive fields, c9: 1 convolutional layers of 384, \(1 \times 1\) local receptive fields, c10 and c11: 2 convolutional layers of 512, \(11 \times 11\) local receptive fields and 384, \(1 \times 1\) local receptive fields, d1 and d2 are deconvolutional layers.

Datasets. The MSRA-B dataset \cite{19} consists of 5000 natural images and their corresponding ground truth maps where the salient objects in the images are segmented with pixel-wise annotation. The dataset is divided into training, validation and testing groups containing 2500, 500 and 2000 images, respectively. Figure 5 illustrates some of the example images from the dataset with their corresponding ground truths. The HKU-IS dataset \cite{20} consists of 4447 natural images and their corresponding ground truth maps where the salient objects in the images are segmented with pixel-wise annotation. The entire dataset is used as a testing group for the descendant networks trained on the training group of the MSRA-B dataset. Figure 5 illustrates some of the example images from the dataset with their corresponding ground truths.

Performance metrics. To evaluate the performance of the evolved descendant deep neural networks at different generations, the MAE, \(F_\beta\) score (where \(\beta^2=0.3\) \cite{20}) metrics were computed for each of the descendant deep neural networks across the 2000 test images of the MSRA-B dataset that were not used for training. As a reference, the
Figure 4: MSRA-B image dataset: This dataset contains 5000 natural images divided into 2500, 500 and 2000 images as training, validation and test samples, respectively. The ground truth maps are provided with pixel-wise annotation. Examples of images and their corresponding ground truth maps in the MSRA-B image dataset are shown here.

Figure 5: HKU-IS image dataset: This dataset contains 4447 natural images, and the entire dataset is used as a testing group for the descendant deep neural networks trained on the training group of the MSRA-B dataset. The same performance metrics was also computed for the original, first generation ancestor deep neural network.

**Detailed experimental results.** The detailed experimental results describing the number of synapses, architectural efficiency (defined here as the reduction of synapses in the network compared to the original, ancestor deep neural network in the first generation), \( F_\beta \) score, and mean absolute error (MAE) presented in Table 1 and Table 2 for the MSRA-B and HKU-IS datasets, respectively.
Table 1: Performance metrics for different generations of synthesized offspring networks for MSRA-B dataset

| Generation | Number of synapses | Architectural efficiency | $F_β$ score | MAE  |
|------------|--------------------|--------------------------|-------------|------|
| 1          | 63767232           | 1X                       | 0.875       | 0.0743|
| 2          | 15471797           | 4.12X                    | 0.876       | 0.0739|
| 3          | 3603007            | 17.69X                   | 0.861       | 0.0813|
| 4          | 1333010            | 47.83X                   | 0.850       | 0.0863|

Table 2: Performance metrics for different generations of synthesized offspring networks for HKU-IS dataset

| Generation | Number of synapses | Architectural efficiency | $F_β$ score | MAE  |
|------------|--------------------|--------------------------|-------------|------|
| 1          | 63767232           | 1X                       | 0.830       | 0.0914|
| 2          | 15471797           | 4.12X                    | 0.826       | 0.0911|
| 3          | 3603007            | 17.69X                   | 0.775       | 0.1087|
| 4          | 1333010            | 47.83X                   | 0.753       | 0.1190|