Evolving Robust Neural Architectures to Defend from Adversarial Attacks

Shashank Kotyan and Danilo Vasconcellos Vargas
Department of Informatics, Kyushu University, Japan

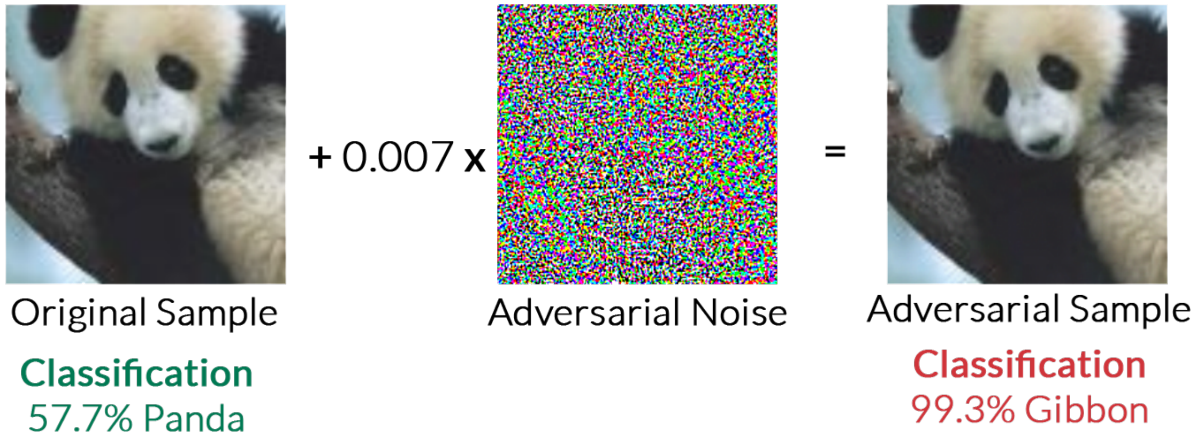
Contents

Background

- Adversarial Attacks
- Adversarial Defence
- Neural Architecture Search

Adversarial Attacks

- Algorithms which prompt failure in neural networks.
- **Practical Application:** Can be applied in real-world scenario.
- **Threats:** Security and safety risks in neural network's



An example of an Adversarial Attack (FGSM) in image classification Goodfellow et al. (2014).

Goodfellow I. J. , Shlens J., and Szegedy C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Adversarial Defence

- Algorithm which mitigate the effect of adversarial attacks.
- Problem: Not consistent, can be ineffective against stronger adversaries.

Neural Architecture Search

- Algorithm which search for best possible architecture of neural network in constraint environment.
- Aim: To develop methods that do not need specialists in order to be applied to a different application.
- Shortcoming: Use of confined exploration area, which spans around the hand-crafted architectures.

- Did you know? Neural Architecture Search has developed one of the SOTA neural networks for image classification problem-NASNet Zoph et al. (2018).

Zoph, B. and Le, Q. V (2018). Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages. 8697–8710.

Neural Architecture Search

Components of Neural Architecture Search

- Search Space:
  - It defines the domain in which the algorithm searches.
  - Most of this search space spans the space, which encompasses the accurate hand-crafted architectures.
- Search Strategy:
  - It defines the policy used to explore the search space effectively in order to find the best feasible solution.
  - Some widely used search strategies are: Random Search, Bayesian Optimisation, Evolutionary Methods, Reinforcement Learning, Gradient Based Methods
- Performance Estimation:
  - It defines the fitness function, which is optimised by the search strategy.

Robust Neural Architecture Search

- Populations
- Mutation Operators
- Evaluation of Architectures
- Niching Scheme
- Evolution

Populations

- Layer Population:

  Containing raw layers (Convolutional and Fully Connected).

- Block Population:

  Containing blocks which are a combination of individuals from layer population

- Model Population:

  Containing architectures which consist of interconnected individuals from block population.

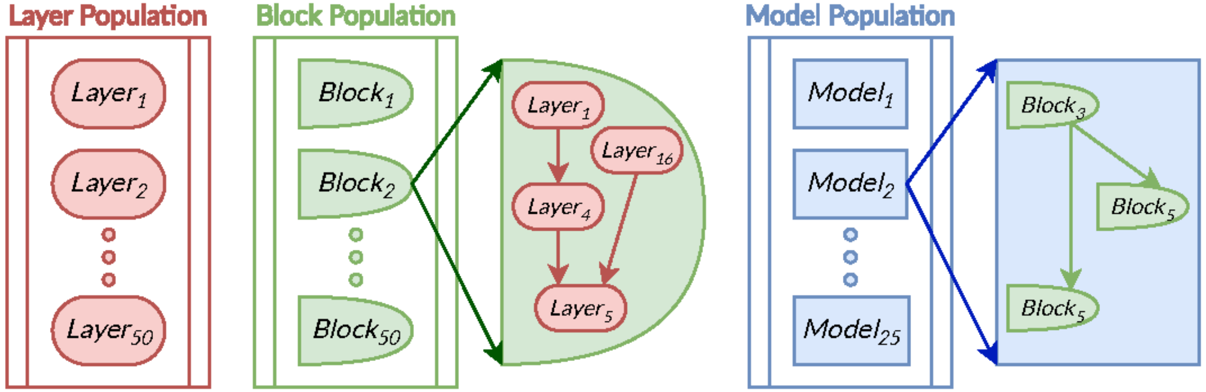


Illustration of the Considered Sub-Populations.

Mutation Operators

- Layer Mutation:

  Changing kernel size | Changing filter size | Changing unit size | Swapping layers

- Block Mutation:

  Adding a layer | Removing a layer | Swapping blocks

  Adding a layer connection | Removing a layer connection

- Model Mutation:

  Adding a block | Removing a block

  Adding a block connection | Removing a block connection

Evaluation of Architectures

Evaluated fitness of the neural network is:

$$Fitness = Accuracy - Robustness$$

- **Accuracy:** Calculated after the model is trained for 50 epochs on the CIFAR-10's entire 100% training dataset on every 10^{th} generation and 2% of the training dataset for every other generation.
- **Robustness:** Calculated using the adversarial samples created from the model-agnostic (black-box) L_0 and L_∞ attacks Kotyan and Vargas (2019).

Kotyan, S., & Vasconcellos Vargas, D. (2019). Adversarial Robustness Assessment: Why both $ L_0 $ and $ L_\infty $ Attacks Are Necessary. arXiv e-prints, arXiv-1906.

Niching Scheme

- To keep a high amount of diversity while exploring in vast search space is achieved by using a novel niching scheme which is based on Spectrum-based niching.
- Here, they use the spectrum as a histogram containing the number of;

Blocks	Total Layers
Dense Layers	Convolution Layers
Block Connections	Total Layer Connections
Dense to Dense Connections	Dense to Convolution Connections
Convolution to Dense Connections	Convolution to Convolution Connections

Evolution

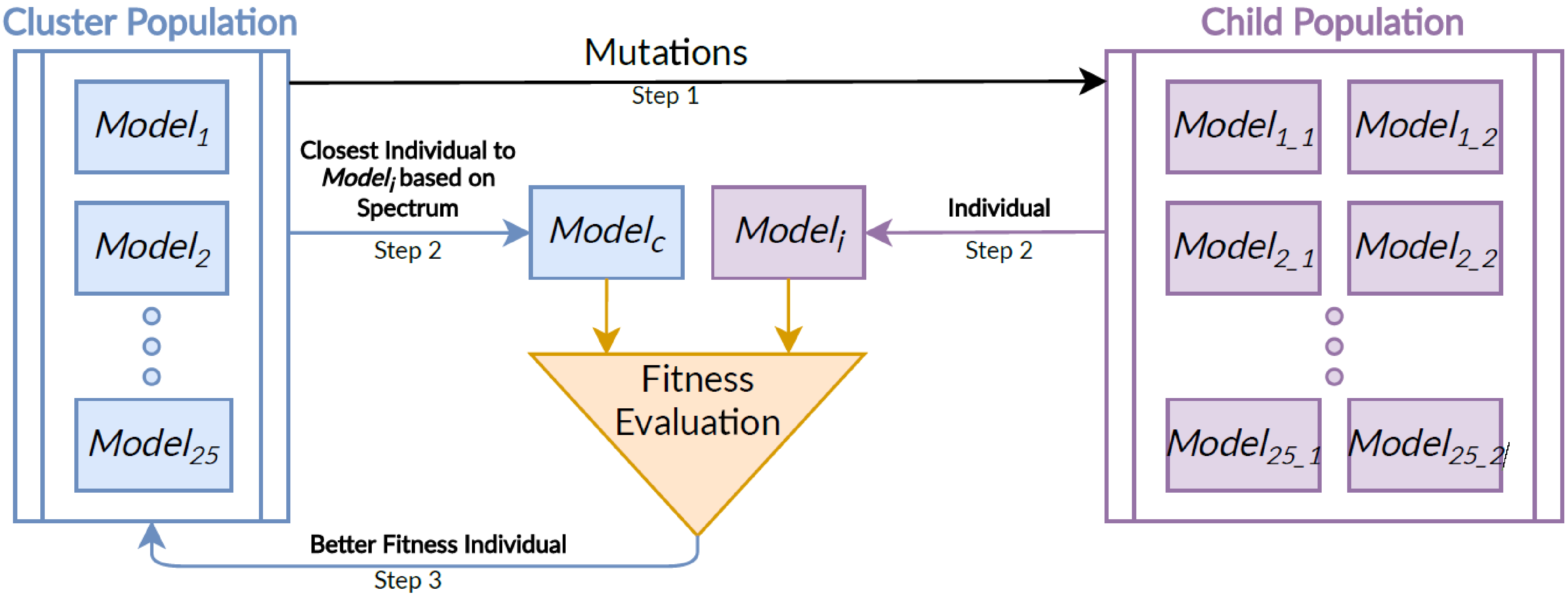
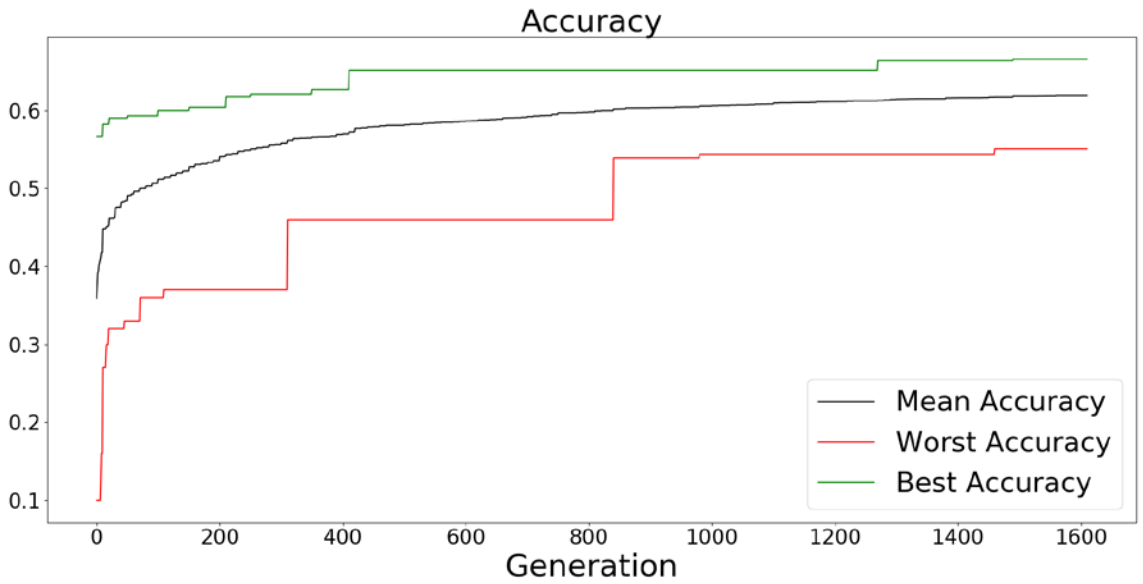


Illustration of the evolution.

Experimental Results

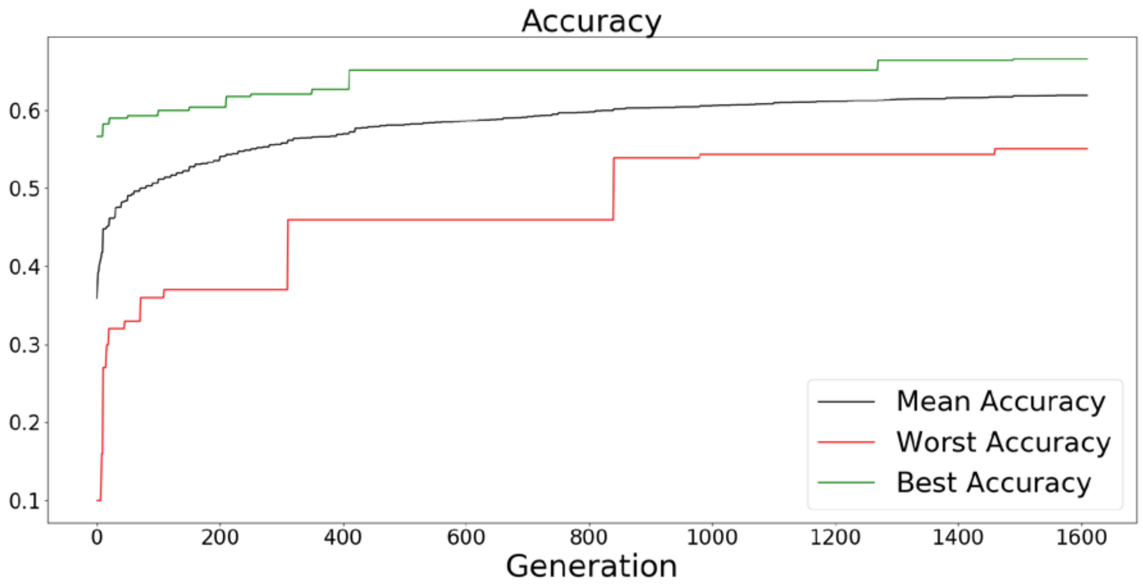
- An unwaveringly improving accuracy curve over generations.
- Suggesting that in evolving, the model steadily intrinsically robust to a comprehensive assortment of adversarial examples.



Accuracy of architectures over generations in evolution.

Experimental Results

- The final evolved robust model is trained with augmented data,
  - Standard Accuracy of 88%
  - Adversarial accuracy (accuracy on adversarial examples) of 58%.
- The current state-of-the-art architectures (ResNet, DenseNet and WideResNet) have 0-10% accuracy on these adversarial samples.

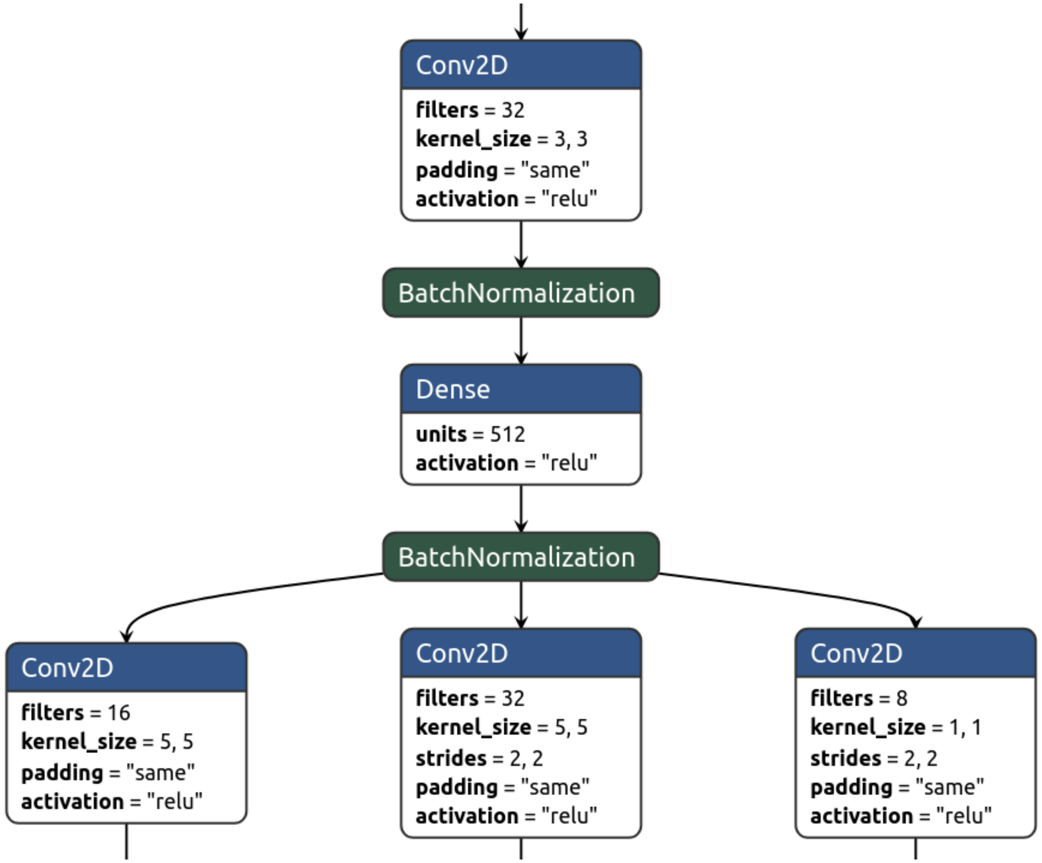


Accuracy of architectures over generations in evolution.

Characteristics of Robust Neural Network

Multiple Bottlenecks and Projections into High- Dimensional Space:

- The robust models have
  - Multiple bottlenecks for feature space
  - Projections of feature space into higher-dimensional space.
- This inference is a clear application of Cover's Theorem [1] which states that,
  - Projecting a feature space into a higher dimensional space makes a feature set linearly-separable.

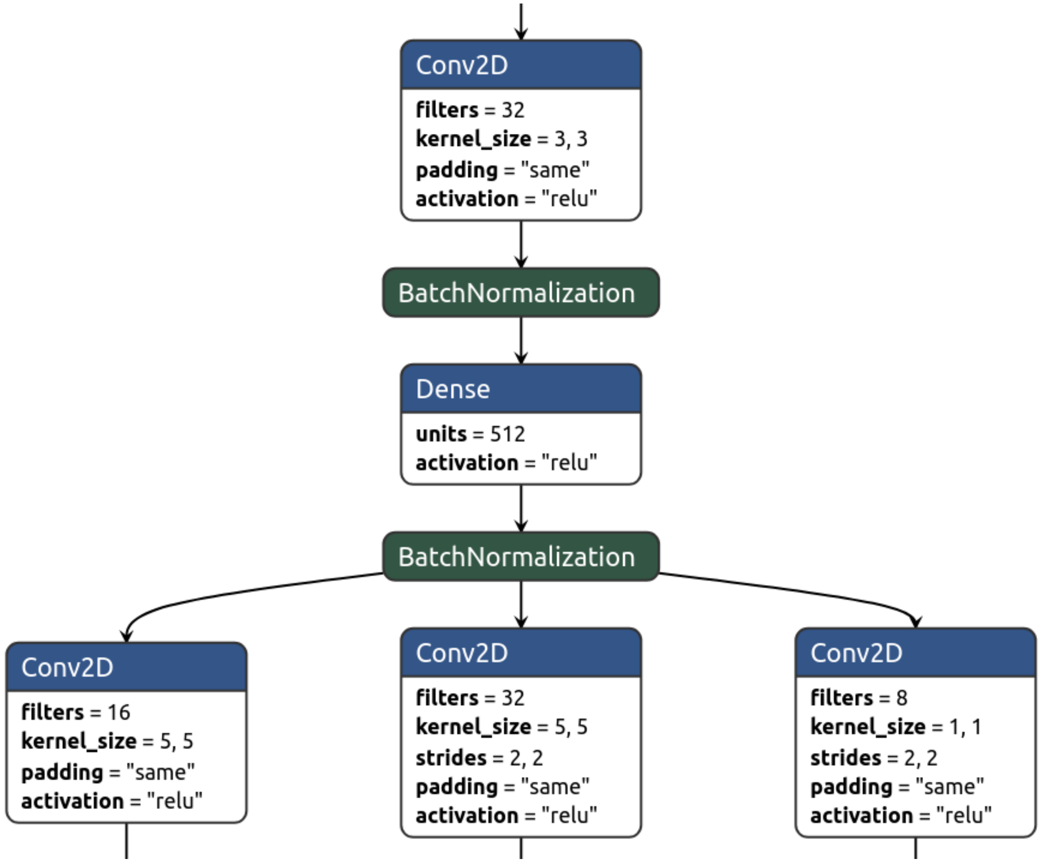


Snippet of a part of a robust architecture.

Characteristics of Robust Neural Network

Paths with Different Constraints:

- A high-dimensional feature space gets split into multiple lower-dimensional feature spaces, each distinct with each other.
- This observation shows that several so-called paths do a separate analysis of the feature space.



Snippet of a part of a robust architecture.

References

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE transactions on electronic computers, (3):326–334.

Goodfellow I. J. , Shlens J., and Szegedy C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Kotyan, S., & Vasconcellos Vargas, D. (2019). Adversarial Robustness Assessment: Why both $ L_0 $ and $ L_\infty $ Attacks Are Necessary. arXiv e-prints, arXiv-1906.

Zoph, B. and Le, Q. V (2018). Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages. 8697–8710.

九州大学
KYUSHU UNIVERSITY

Thank you !

Please feel free to ask questions.