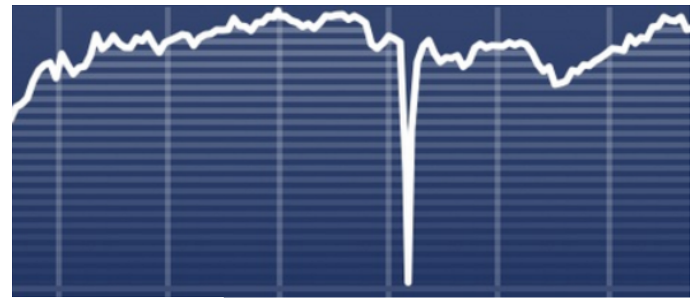
ML Meets Attackers

Syrian hackers compromise @AP:

$\Longrightarrow$



$136 billion drop

ML Meets Attackers

Syrian hackers compromise @AP:

$\Longrightarrow$

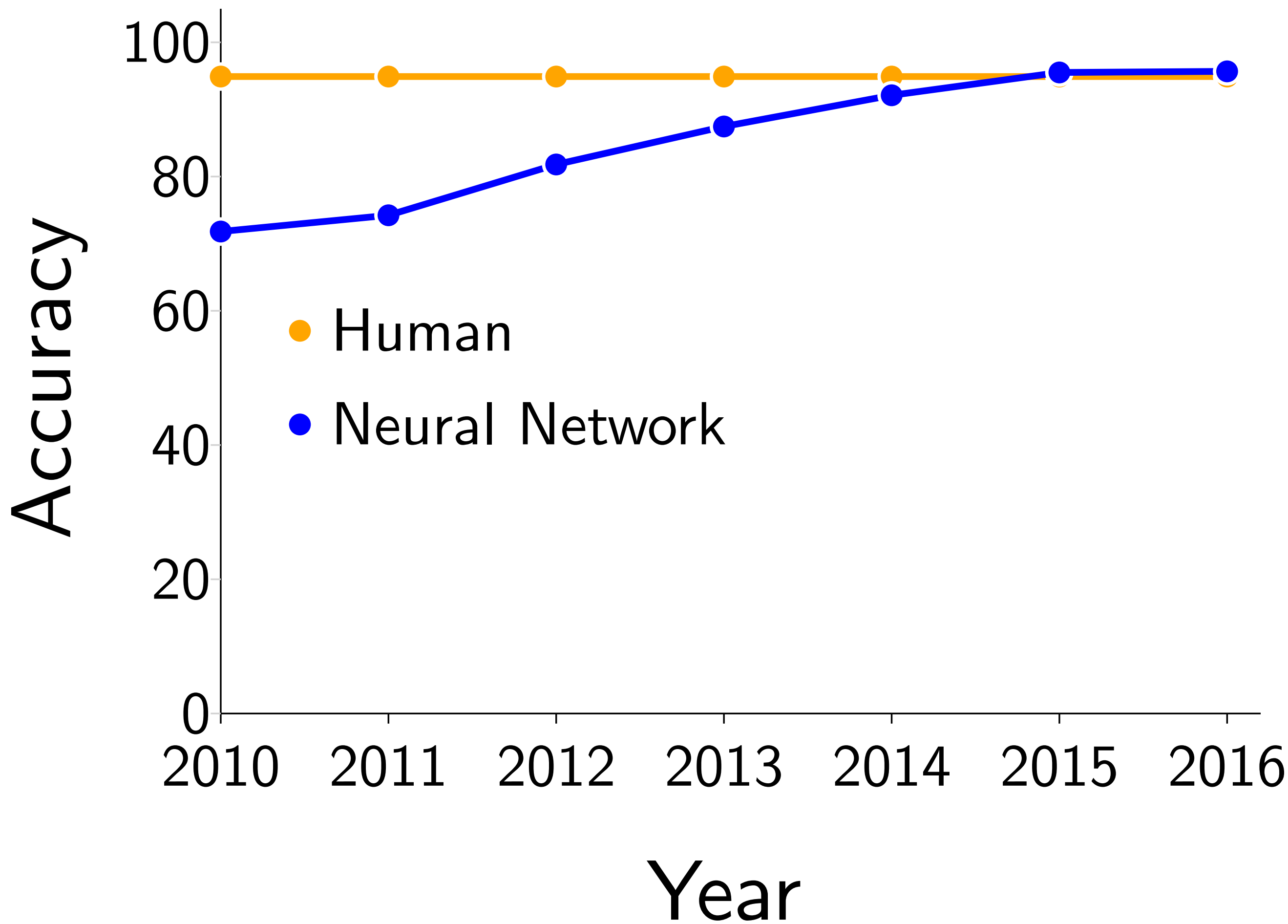
$136 billion drop

Bots influenced U.S., other elections [Marwick & Lewis '17]

- presidential debates, #MacronLeaks
- affect trending topics

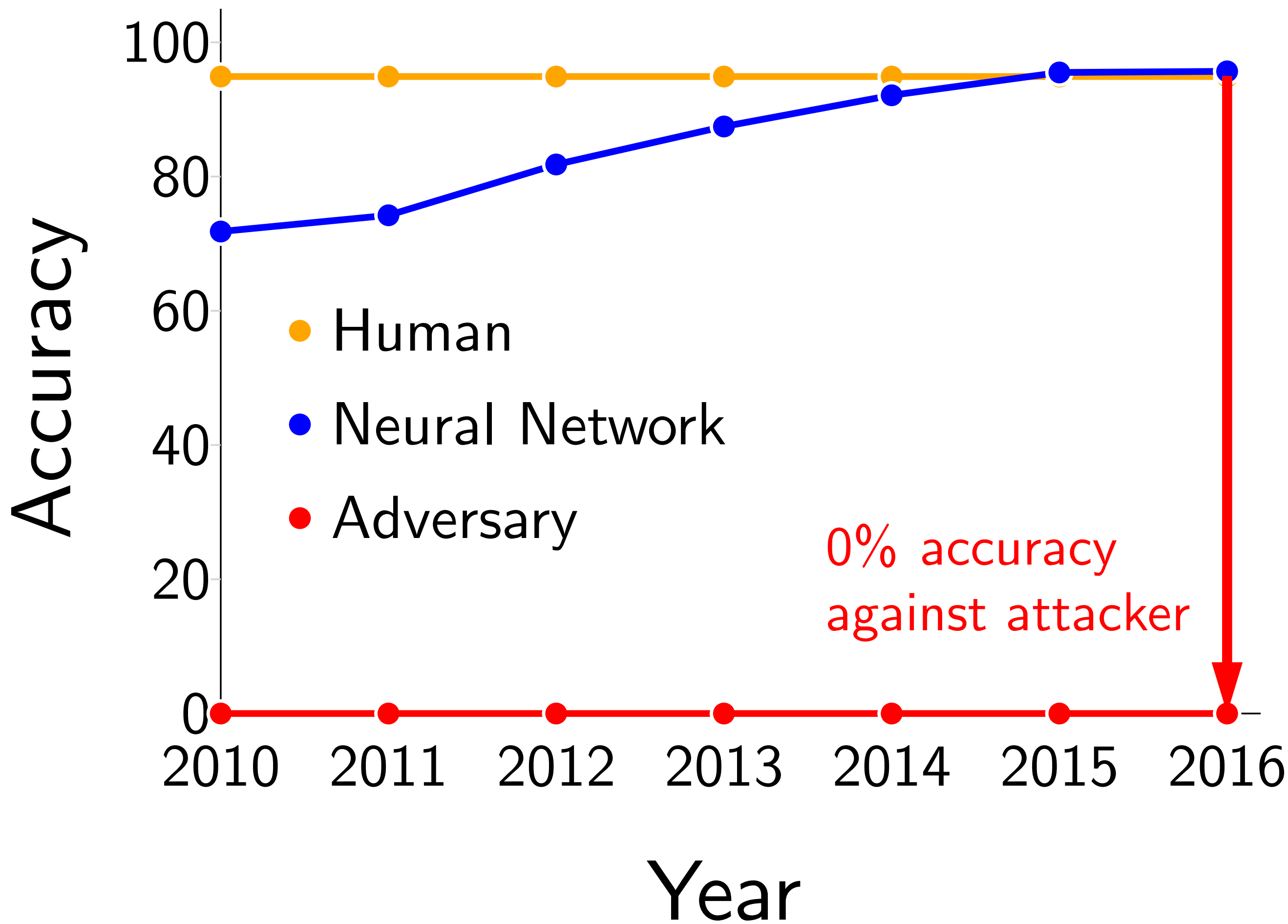
ML: Powerful But Fragile

ML is state-of-the-art in many domains, such as vision:

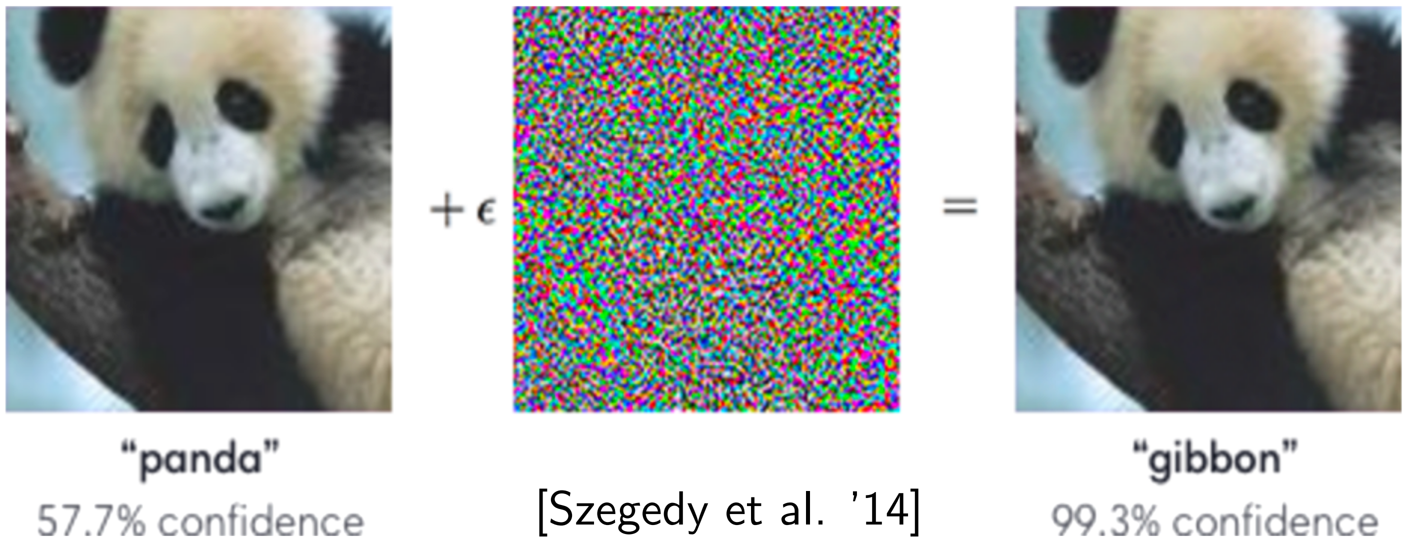


ML: Powerful But Fragile

ML is state-of-the-art in many domains, such as vision:

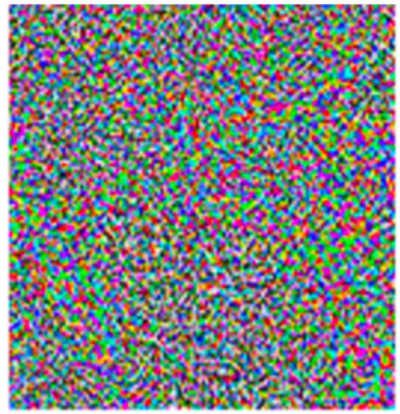


Machine Learning is Insecure



[Szegedy et al. ’14]

Machine Learning is Insecure

 $+\epsilon$  $=$ 

"panda"
57.7% confidence

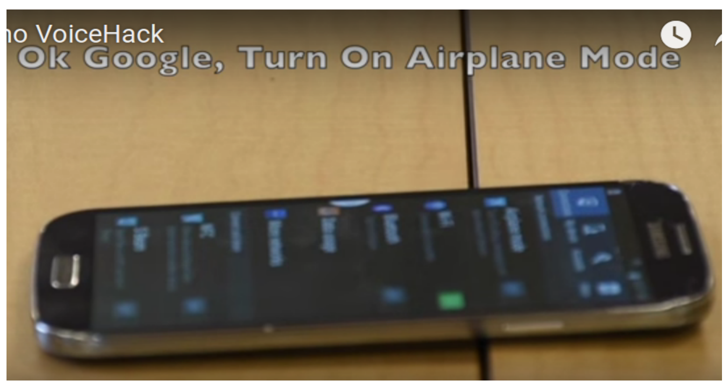
[Szegedy et al. '14]

"gibbon"
99.3% confidence

Self-driving cars:

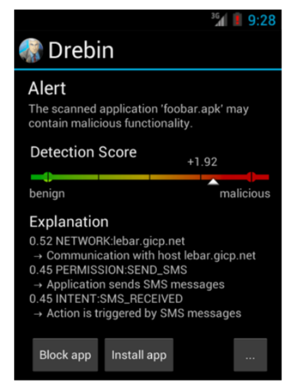
stop → yield
[Evtimov et al. '17]

Speech recognition:



noise → "Ok Google"
[Carlini et al. '16]

Malware:

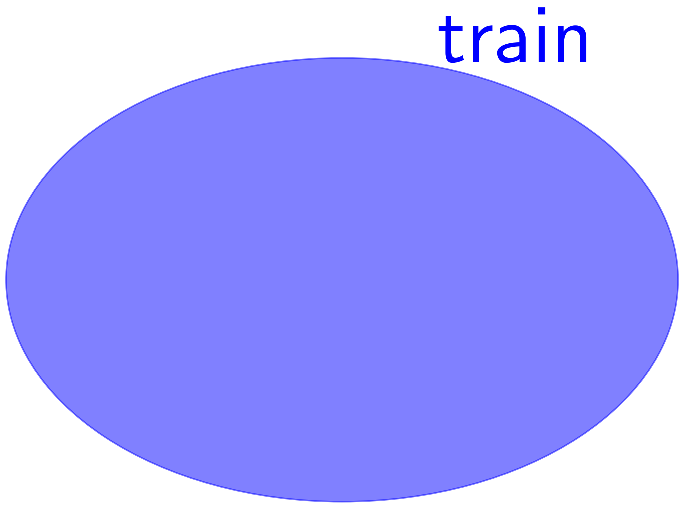


malware → benign
[Grosse et al. '16]

ML Paradigm is Broken

Most ML systems assume:

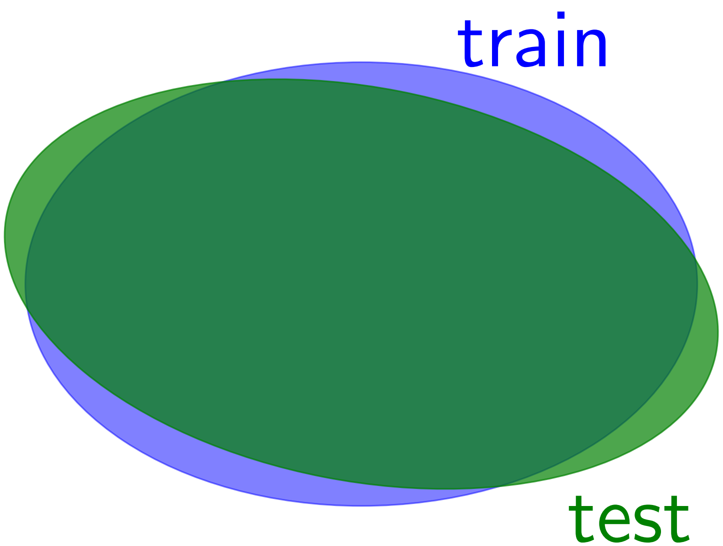
train (data collection) $\approx$ **test (deployment)**



ML Paradigm is Broken

Most ML systems assume:

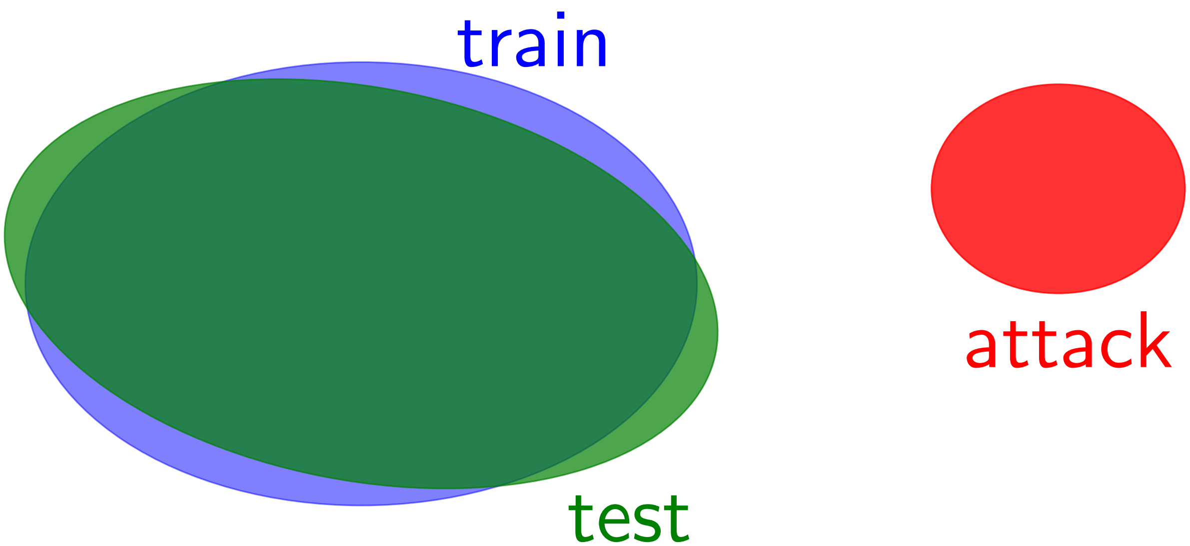
train (data collection) $\approx$ **test (deployment)**



ML Paradigm is Broken

Most ML systems assume:

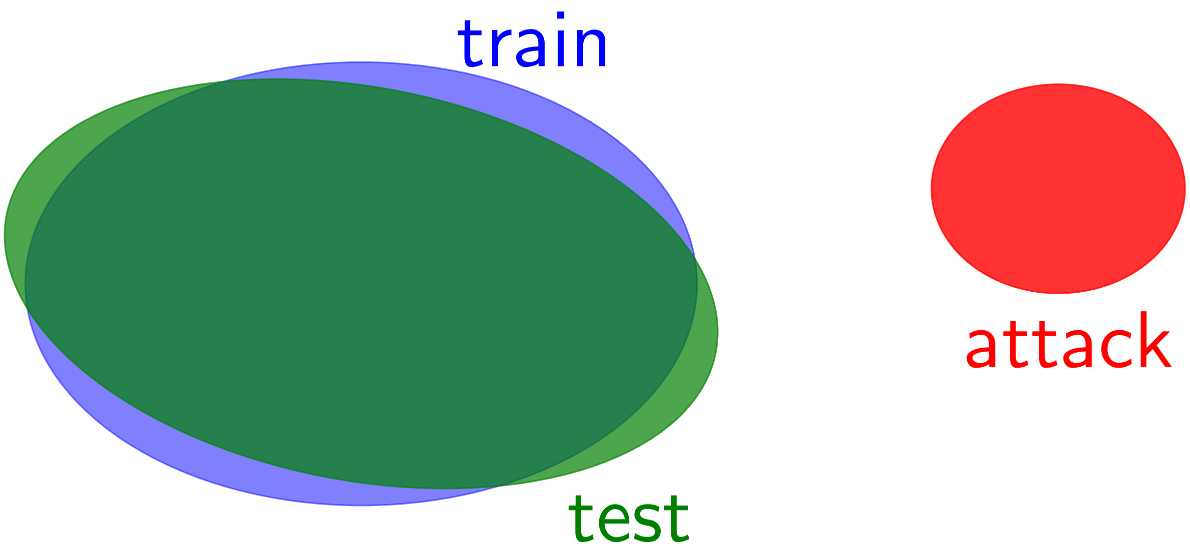
$$\textbf{train (data collection)} \approx \textbf{test (deployment)}$$



ML Paradigm is Broken

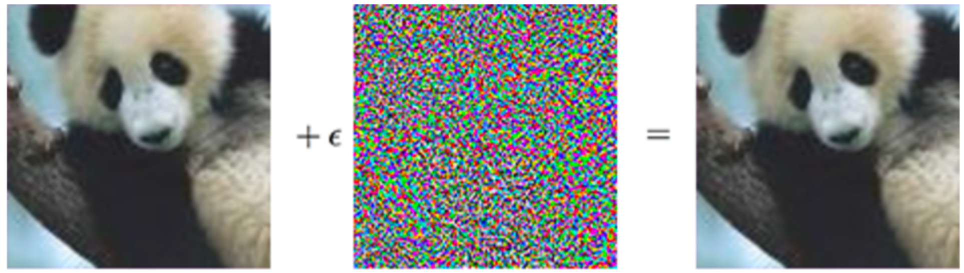
Most ML systems assume:

train (data collection) $\approx$ **test (deployment)**



Attackers can easily violate assumption, create vulnerabilities!

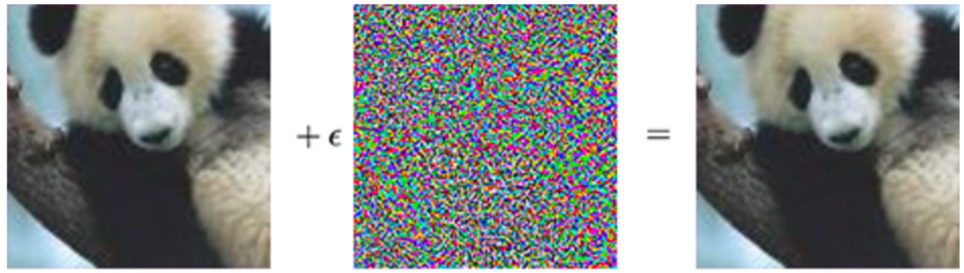
Arms Races



Empirical evaluation against attacks insufficient:

$$\boxed{\begin{array}{c}\text{discovery}\\ \text{[Szegedy et al. '14]}\end{array}}$$

Arms Races



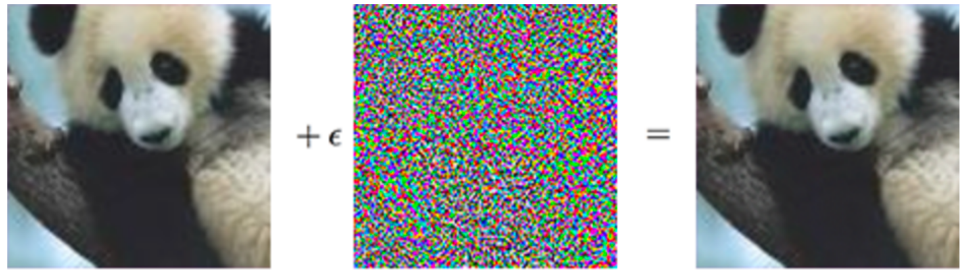
Empirical evaluation against attacks insufficient:

discovery
[Szegedy et al. '14]

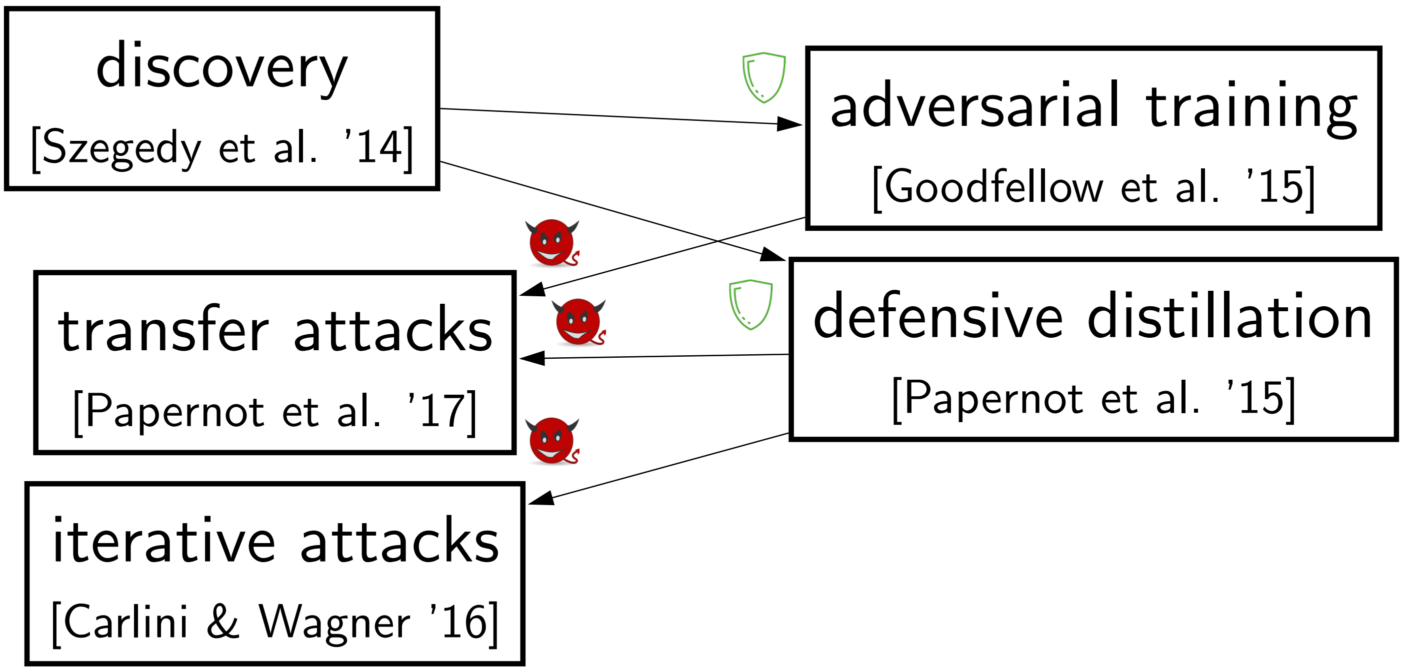
adversarial training
[Goodfellow et al. '15]

defensive distillation
[Papernot et al. '15]

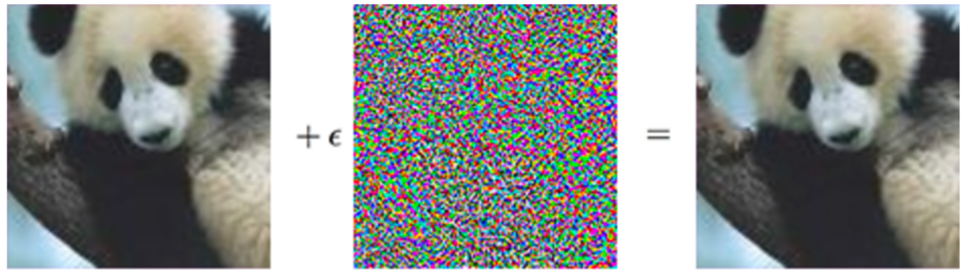
Arms Races



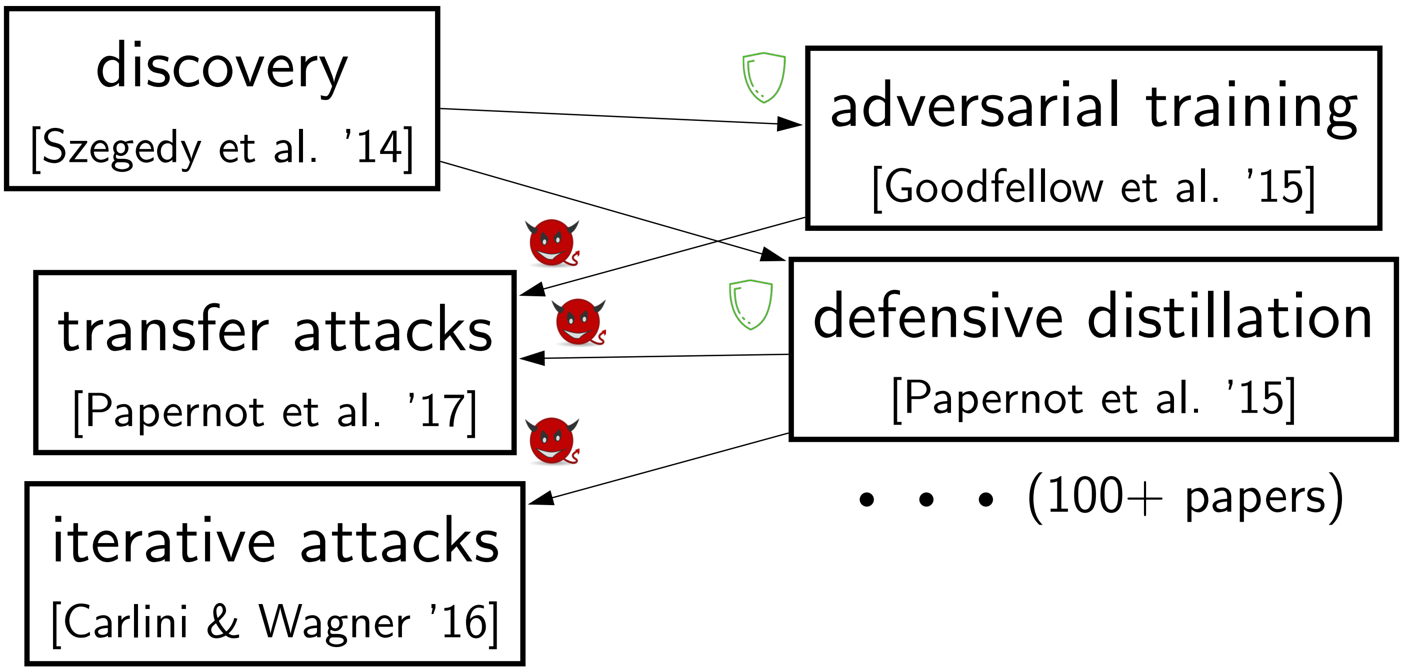
Empirical evaluation against attacks insufficient:



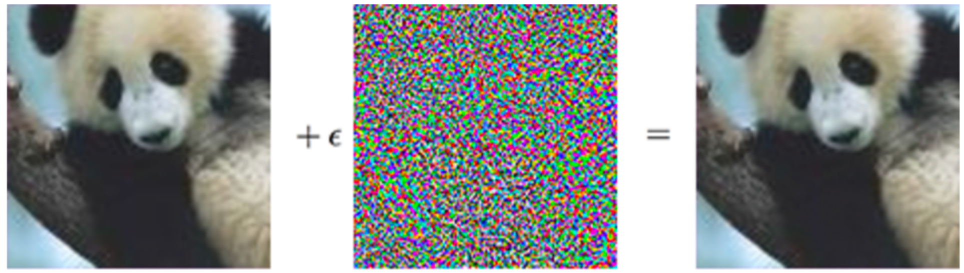
Arms Races



Empirical evaluation against attacks insufficient:



Arms Races



Empirical evaluation against attacks insufficient:

discovery
[Szegedy et al. '14]

adversarial training
[Goodfellow et al. '15]

defensive distillation
[Papernot et al. '15]

transfer attacks
[Papernot et al. '17]

iterative attacks
[Carlini & Wagner '16]

• • • (100+ papers)

bypassing 10 defenses
[Carlini & Wagner '17]

Take-away

Can't just "see what works"–leads to a **security arms race** that defenders often lose!

Take-away

Can't just "see what works"– leads to a **security arms race** that defenders often lose!

Need new methodology to evaluate robustness.

Adversarial Examples are Persistent

Persist despite hundreds of papers trying to avoid them

Adversarial Examples are Persistent

Persist despite hundreds of papers trying to avoid them

stop $\rightarrow$ yield

[Evtimov et al. '17]

turtle $\rightarrow$ rifle

[Athalye et al. '17]

banana $\rightarrow$ toaster

[Brown et al. '17]

Adversarial Examples are Persistent

Persist despite hundreds of papers trying to avoid them

stop $\rightarrow$ yield
[Evtimov et al. '17]

turtle $\rightarrow$ rifle
[Athalye et al. '17]

banana $\rightarrow$ toaster
[Brown et al. '17]

Most defenses fail within weeks **(arms race)**, but a few have lasted.

Adversarial Examples are Persistent

Persist despite hundreds of papers trying to avoid them

stop $\rightarrow$ yield
[Evtimov et al. '17]

turtle $\rightarrow$ rifle
[Athalye et al. '17]

banana $\rightarrow$ toaster
[Brown et al. '17]

Most defenses fail within weeks **(arms race)**, but a few have lasted.

What makes them different?

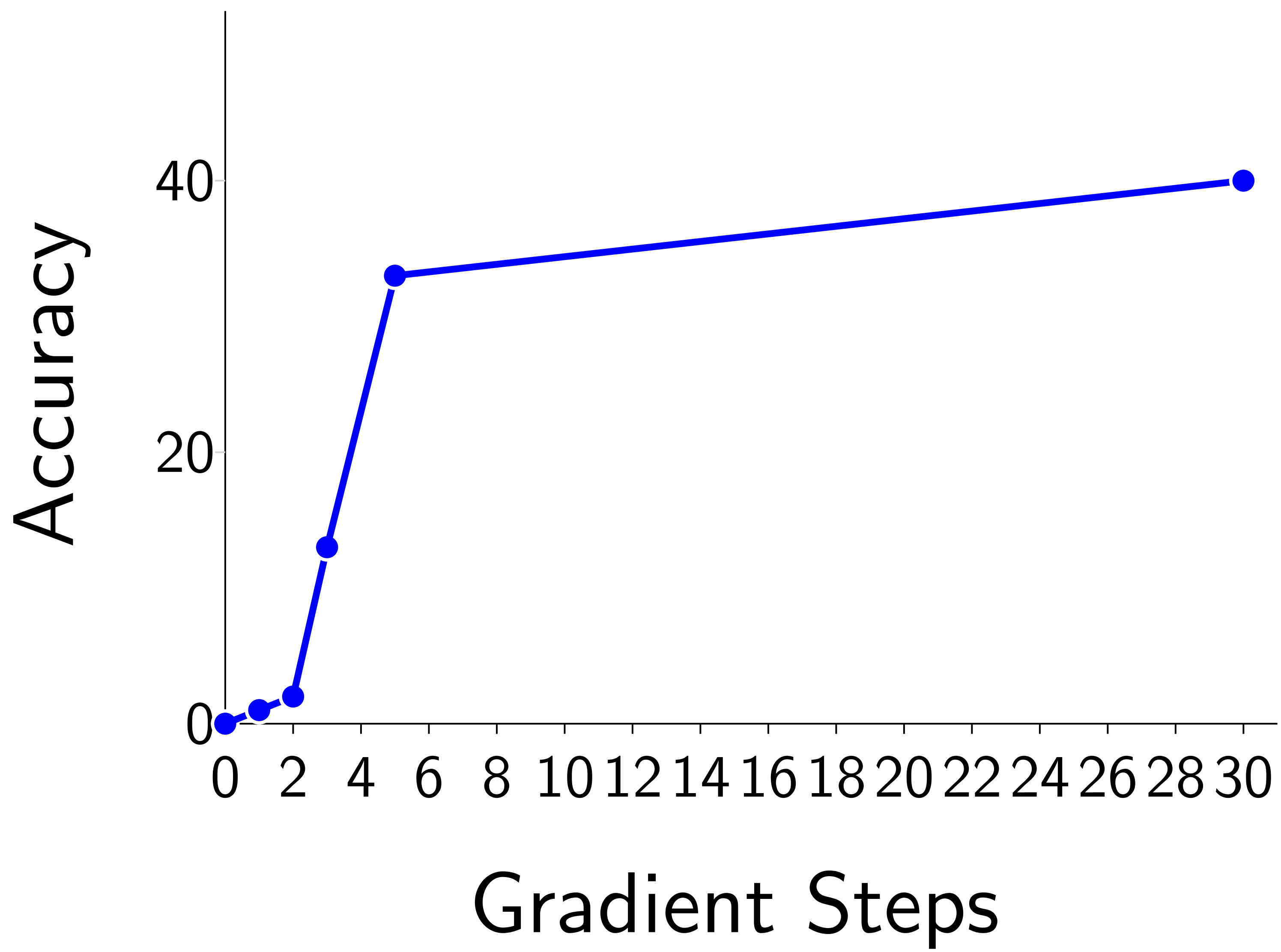
Details of the robust model

Obtained via **adversarial training** (train on adversarial images)

Generate training images via **gradient ascent** on cross-entropy loss

If too few gradient steps, model learns to **fool optimizer** instead of being truly robust

Accuracy vs. gradient steps



Tool: Visualization (Lucid)

The Building Blocks of Interpretability

Interpretability techniques are normally studied in isolation.
We explore the powerful interfaces that arise when you combine them —
and the rich structure of this combinatorial space.

CHOOSE AN INPUT IMAGE

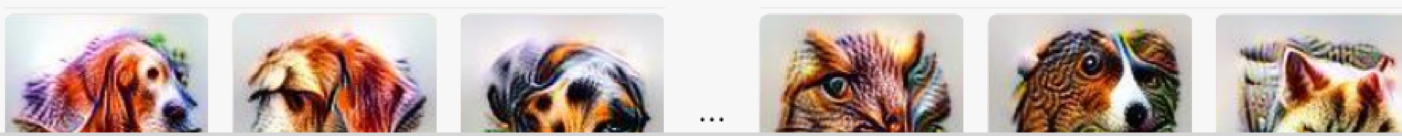
For instance, by combining feature visualization (*what is a neuron looking for?*) with attribution (*how does it affect the output?*), we can explore how the network decides between labels like **Labrador retriever** and **tiger cat**.

Several floppy ear detectors seem to be important when distinguishing dogs, whereas pointy ears are used to classify "tiger cat".

CHANNELS THAT MOST SUPPORT ...

LABRADOR RETRIEVER ▾ ←→ TIGER CAT ▾

feature visualization of channel



Tool: Visualization (Lucid)

Lucid

status alpha | build passing | coverage 82% | python 2.7 | 3.6 | pypi v0.3.8

Lucid is a collection of infrastructure and tools for research in neural network interpretability.

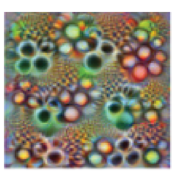
- **Notebooks** -- Get started without any setup!
- **Reading** -- Learn more about visualizing neural nets.
- **Community** -- Want to get involved? Please reach out!
- **Additional Information** -- Licensing, code style, etc.
- **Start Doing Research!** -- Want to get involved? We're trying to research openly!

Notebooks

Start visualizing neural networks ***with no setup***. The following notebooks run right from your browser, thanks to Colaboratory. It's a Jupyter notebook environment that requires no setup to use and runs entirely in the cloud.

You can run the notebooks on your local machine, too. Clone the repository and find them in the `notebooks` subfolder. You will need to run a local instance of the Jupyter notebook environment to execute them.

Tutorial Notebooks



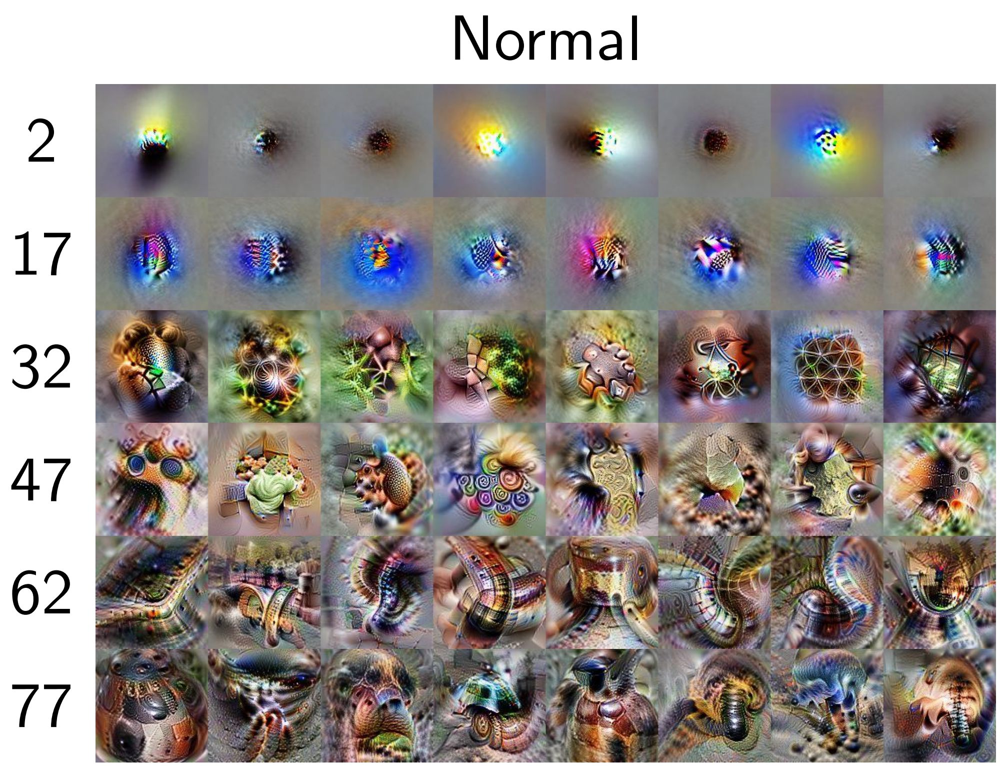
Lucid Tutorial [colab]

Quickly get started using Lucid. Become familiar with changing **objectives, transformations,** and **paramaterization.**

Modelzoo Introduction [colab]

Visualizing Neural Networks

Visualization: find images that maximally excite different neurons.

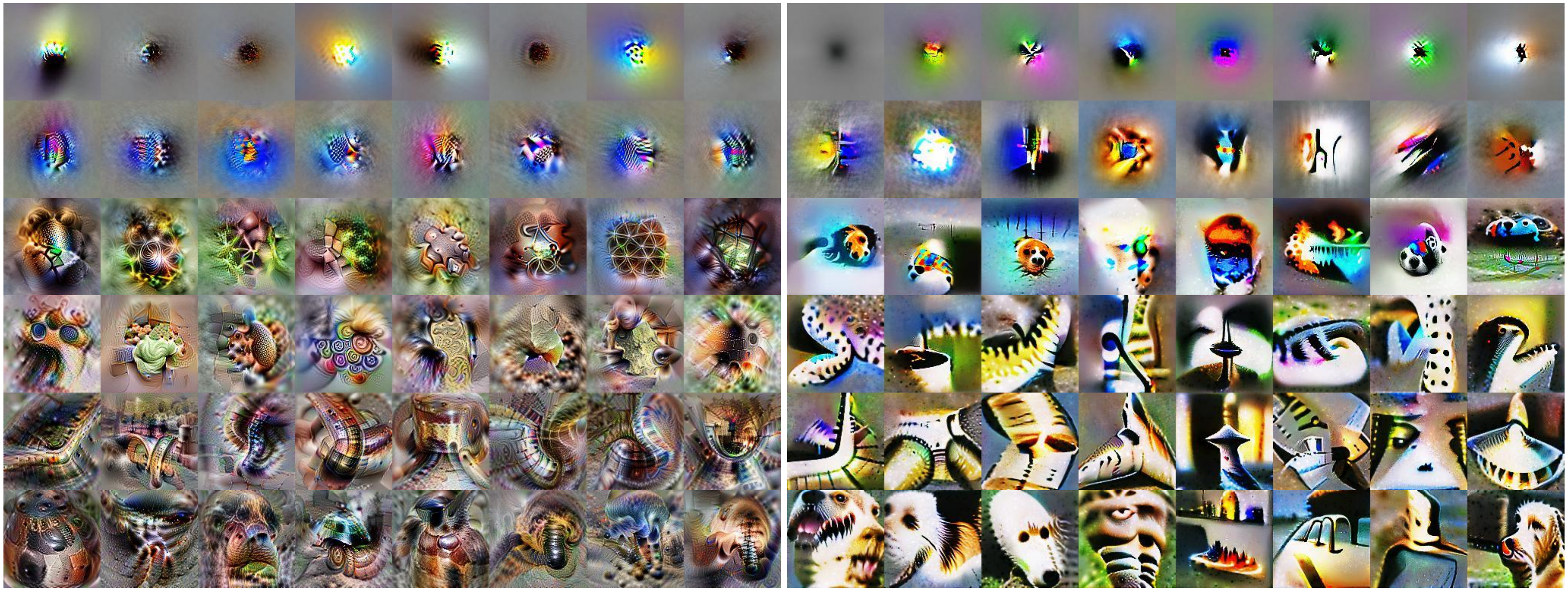


Visualizing Neural Networks

Visualization: find images that maximally excite different neurons.

Normal Robust

2
17
32
47
62
77

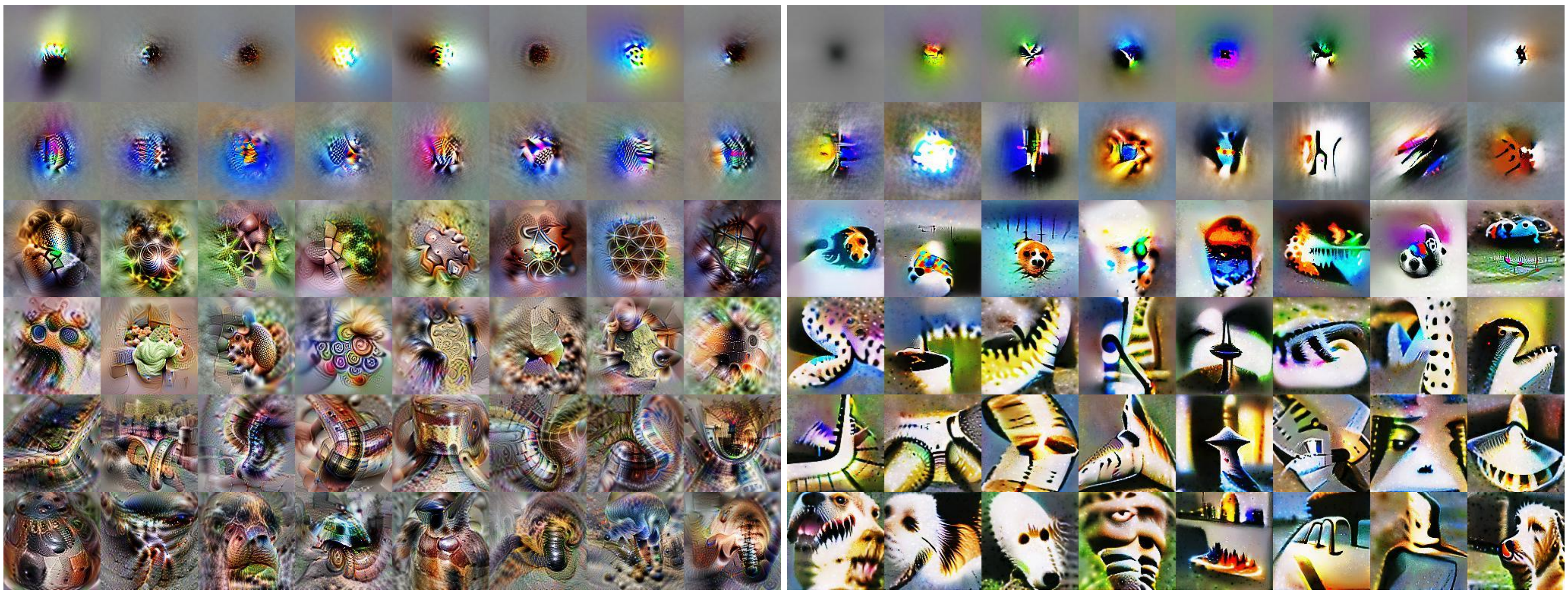


Visualizing Neural Networks

Visualization: find images that maximally excite different neurons.

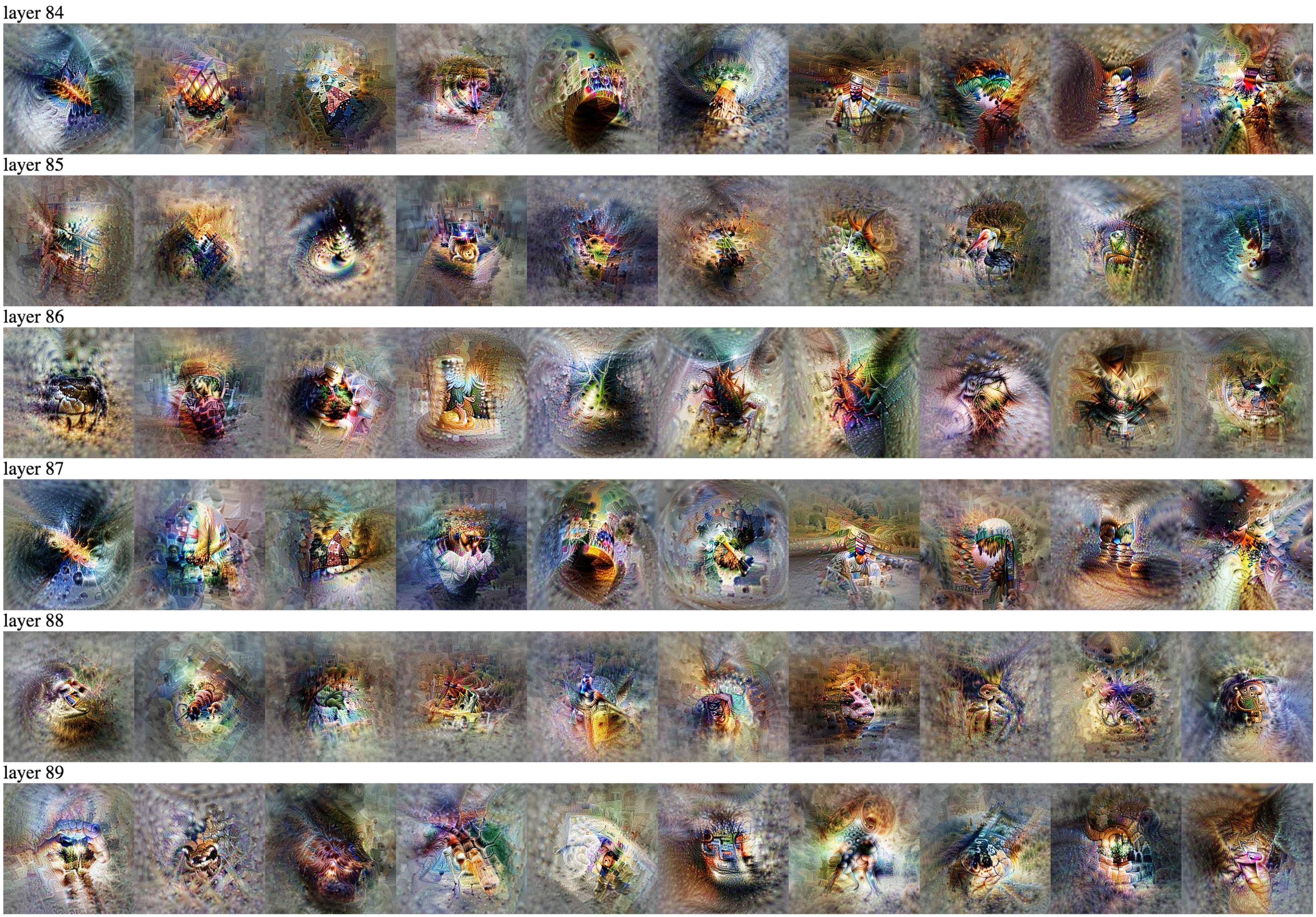
Normal Robust

2
17
32
47
62
77



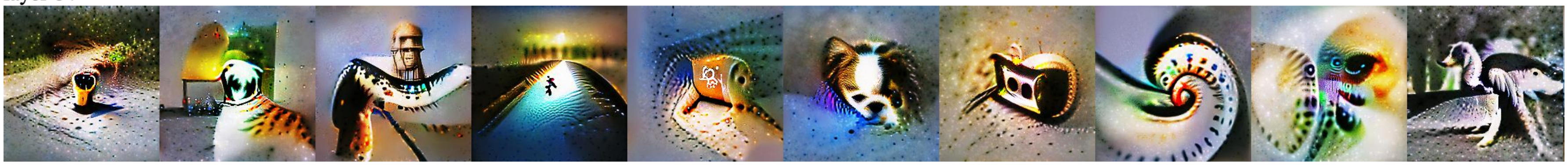
Other non-robust model: 

Regular network (zoomed in)

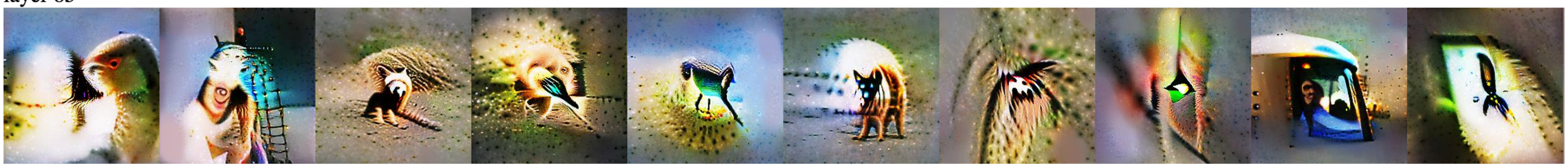


Robust network (zoomed in)

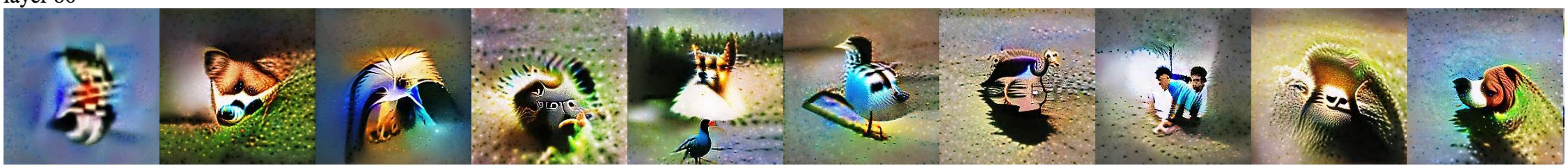
layer 84



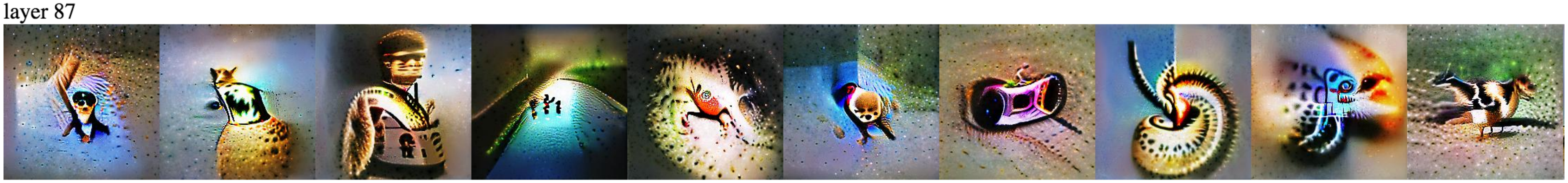
layer 85



layer 86



layer 87

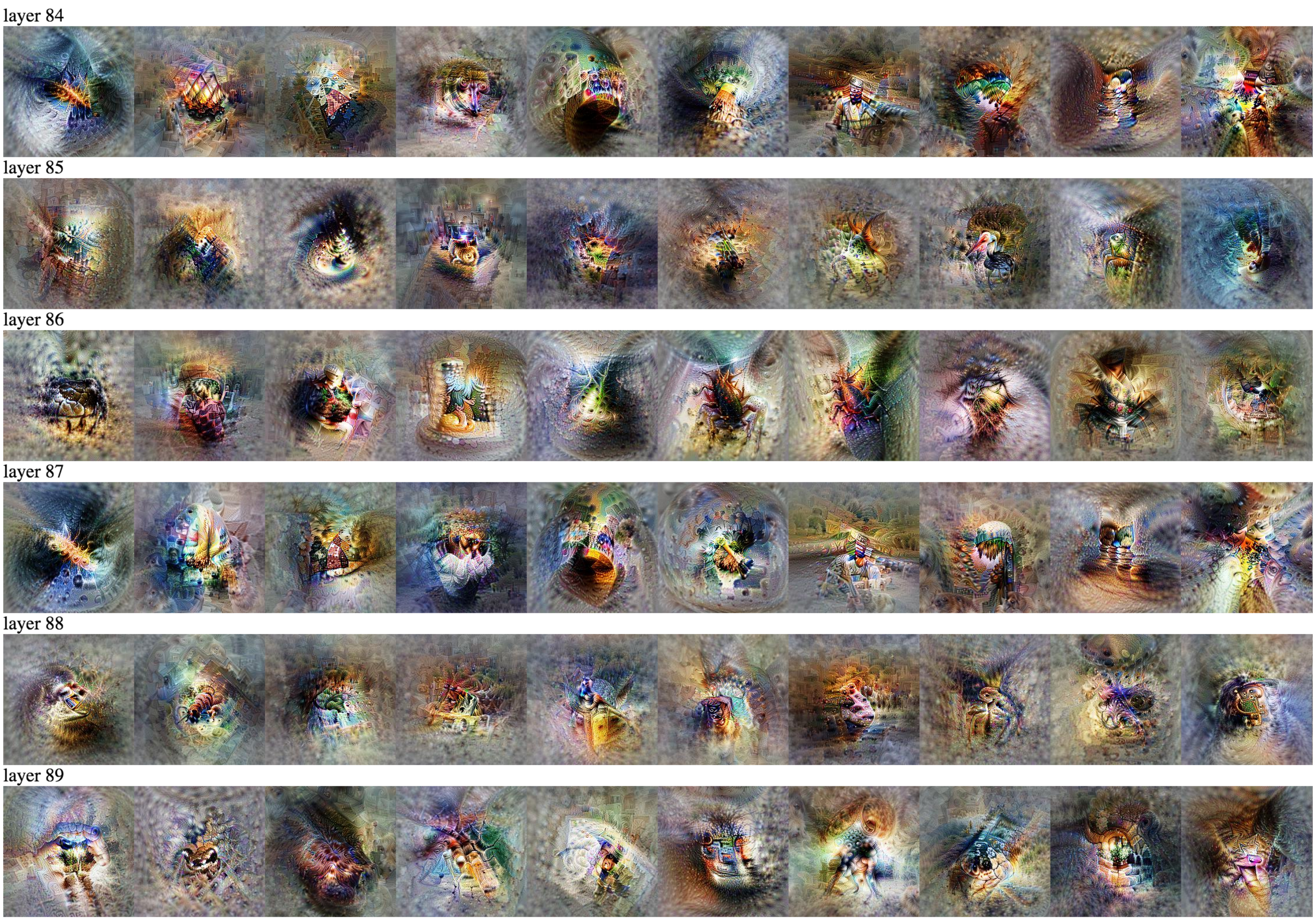


layer 88

layer 89

Neurons vs. gradient steps

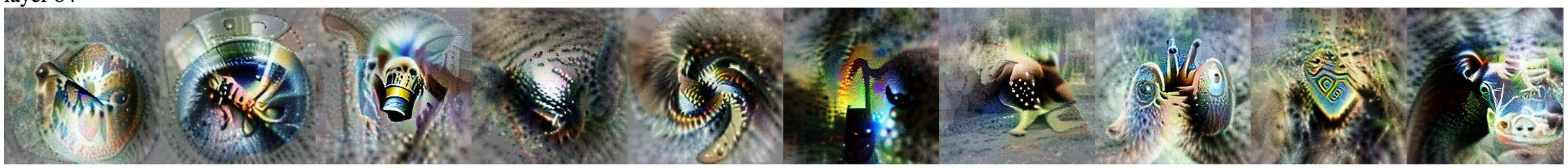
0:



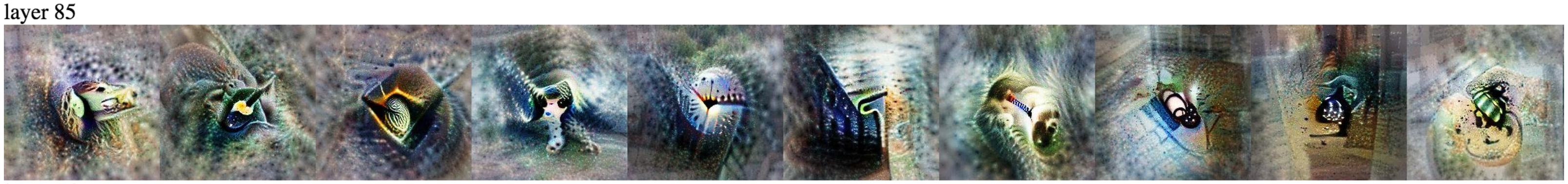
Neurons vs. gradient steps

1:

layer 84

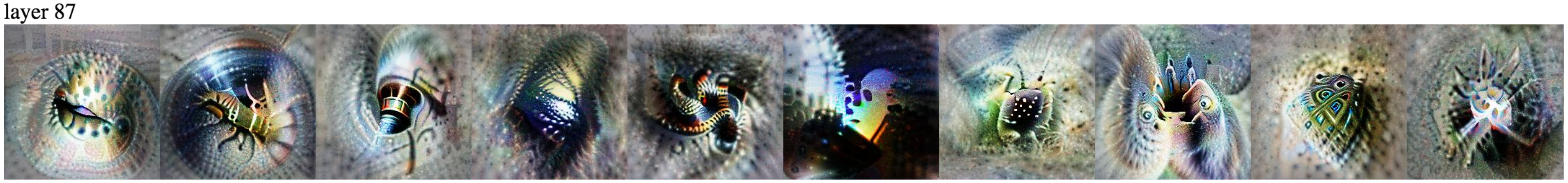


layer 85



layer 86

layer 87



layer 88

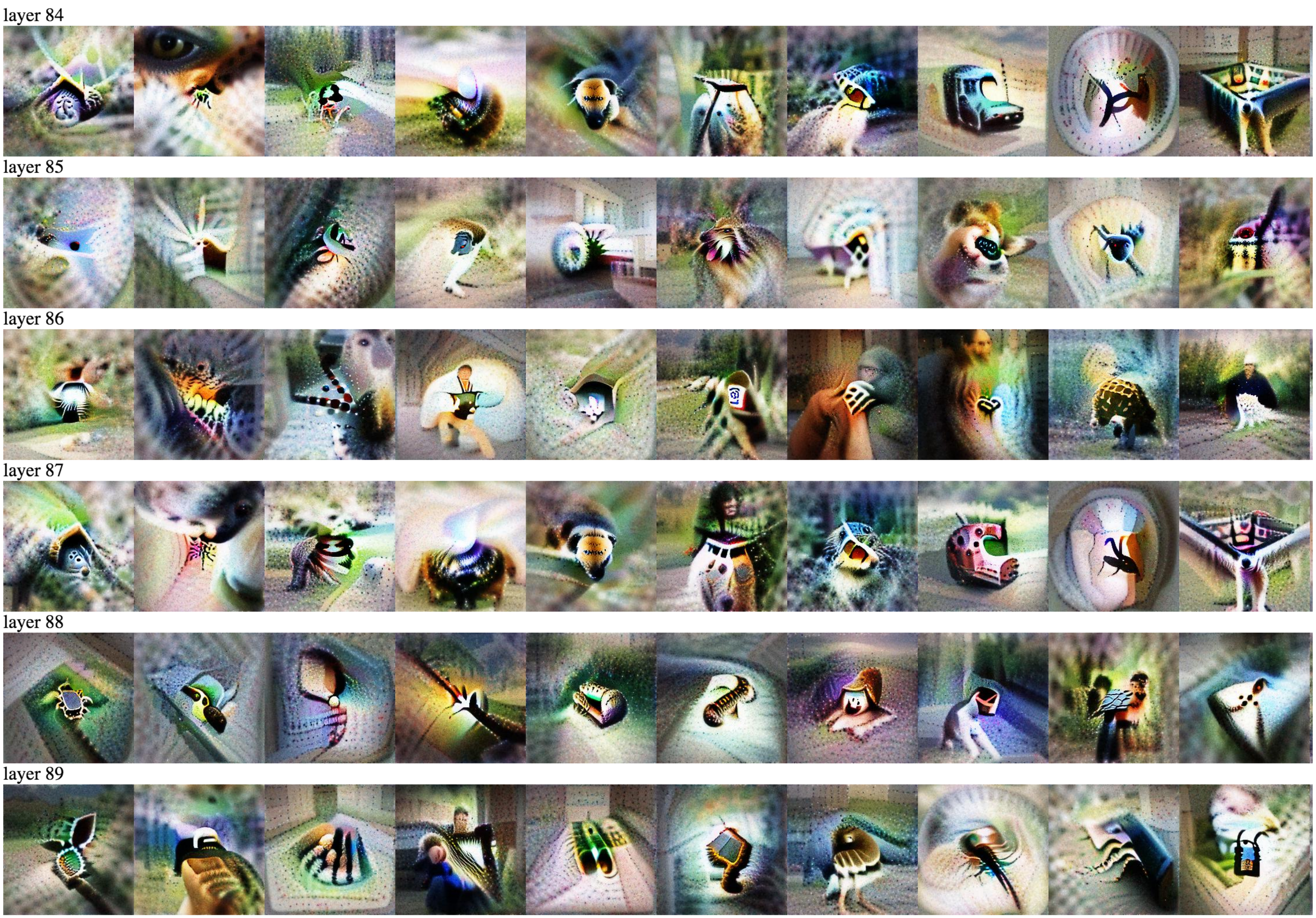
layer 89

Neurons vs. gradient steps

2:

Neurons vs. gradient steps

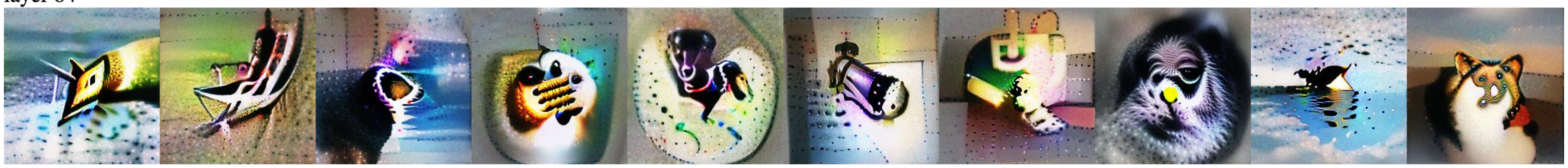
3:



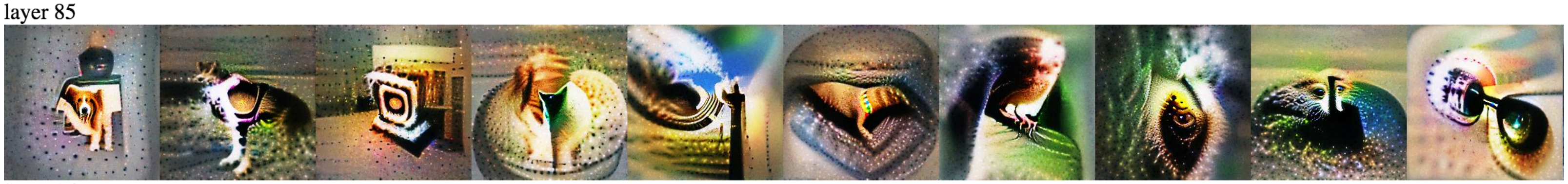
Neurons vs. gradient steps

5:

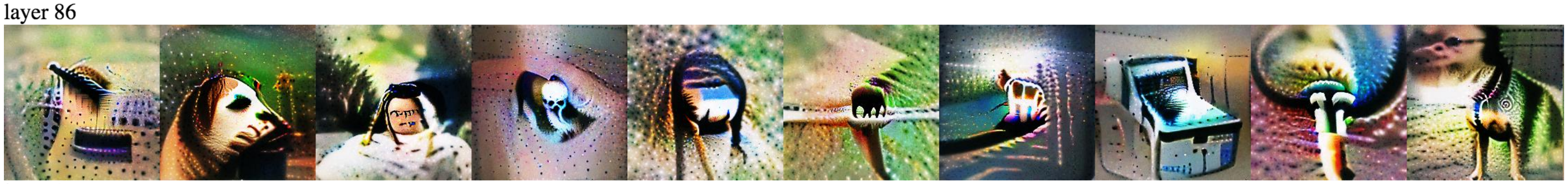
layer 84



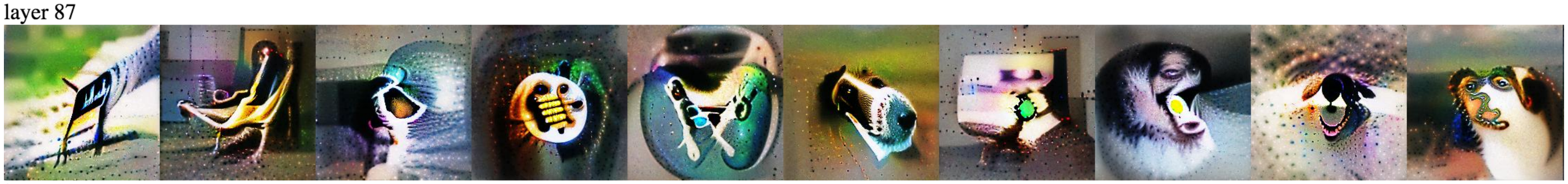
layer 85



layer 86



layer 87



layer 88

layer 89

Neurons vs. gradient steps

30: