Physics: A Gateway to Bayesian Deep Learning

07/11/18

Henri Palacci

@henripal

Columbia University & NeuroLead.AI

#SciPy2018

Austin, Texas July 9-15

Scientific Computing with Python

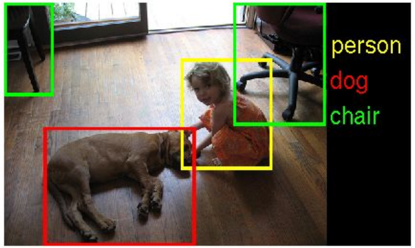
Deep Learning is the state of the art for:

- Speech Recognition and Generation
- Language Recognition and Understanding

Deep Learning is the state of the art for:

- Speech Recognition and Generation
- Language Recognition and Understanding

- Image and Video Processing

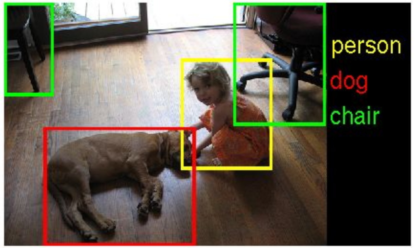


https://www.kaggle.com/c/imagenet-object-detection-from-video-challenge

Deep Learning is the state of the art for:

- Speech Recognition and Generation
- Language Recognition and Understanding

- Image and Video Processing



https://www.kaggle.com/c/imagenet-object-detection-from-video-challenge

- Decision making in controlled environments (games!)

Source Google DeepMind via YouTube

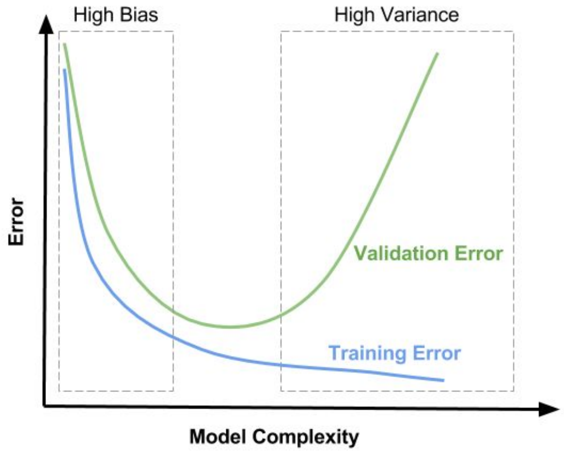
Limitations of Deep Learning

- Deep Learning models:
  - are Black Boxes
  - cannot identify “unusual” data
  - don’t provide error bounds on predictions

Source: reddit, /r/pics, /u/s1lentway

Limitations of Deep Learning

- Generalization in Deep Learning is not well understood



Limitations of Deep Learning

- Generalization in Deep Learning is not well understood

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. “Understanding Deep Learning Requires Rethinking Generalization.” arXiv:1611.03530 [cs], November. http://arxiv.org/abs/1611.03530.

Data: $\mathbf{x}_1, \ldots, \mathbf{x}_m$

Source Karen Zack, twitter.com/teenybiscuit

Dog Mop Dog Mop Dog Mop Dog Mop

Data: $\mathbf{x}_1, \ldots, \mathbf{x}_m$

Source Karen Zack, twitter.com/teenybiscuit

Dog Mop Dog Mop Dog Mop Dog Mop

Our deepnet is a function: **`model`**`(data=x`$_i$`, weights=w)`

`model` returns either **`dog`** or **`mop`**

Data: $\mathbf{x}_1, \ldots, \mathbf{x}_m$

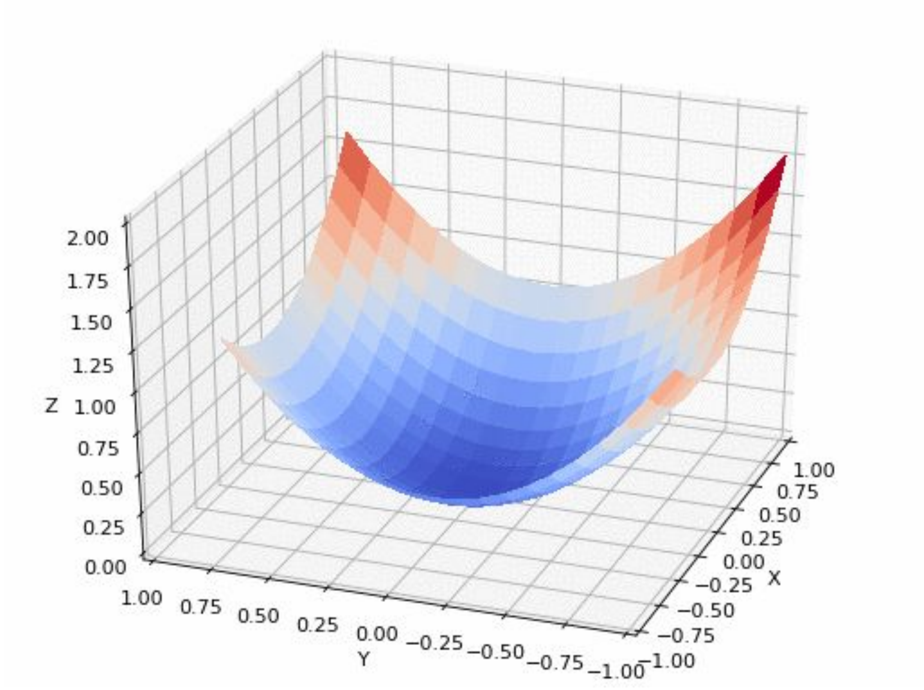
Source Karen Zack, twitter.com/teenybiscuit

Dog Mop Dog Mop Dog Mop Dog Mop

Our deepnet is a function: **`model`**`(data=`x_i`, weights=w)`

`model` returns either **`dog`** or **`mop`**

We ‘learn’ the weights **`w`** by minimizing a loss function on training data.



source: https://github.com/pvigier/gradient-descent

Stochastic Gradient Descent (SGD):

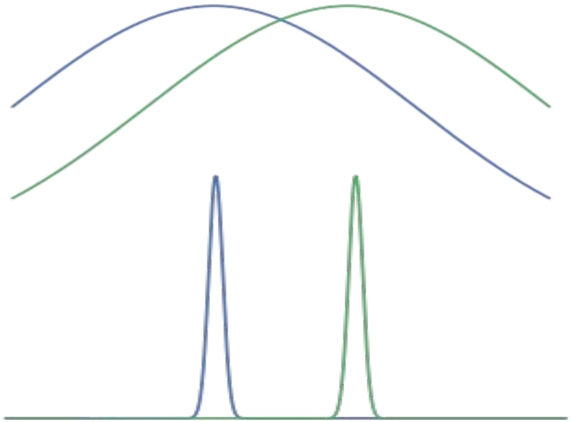
$$w_{t+1} = w_t - \text{lambda} * \text{grad}(\text{loss}(x_i, l_i, w_t))$$

SGD:

- Gives us a (locally) optimal solution, $\mathtt{w}_{\mathtt{min}}$

What if we had a probability distribution for $\mathtt{w}$?

- Error bars
- Helps reduce overfitting
- May help detect “unusual” data



Statistical Mechanics

source: Stephen Curry, https://www.youtube.com/watch?v=ernnQJwaKTs

Statistical Mechanics

Particles are:

- driven to lower energy states

```
force = - grad(energy)
```

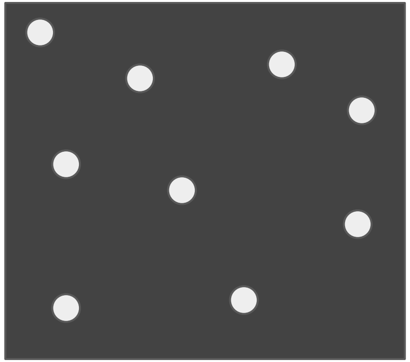
Statistical Mechanics

Particles are:

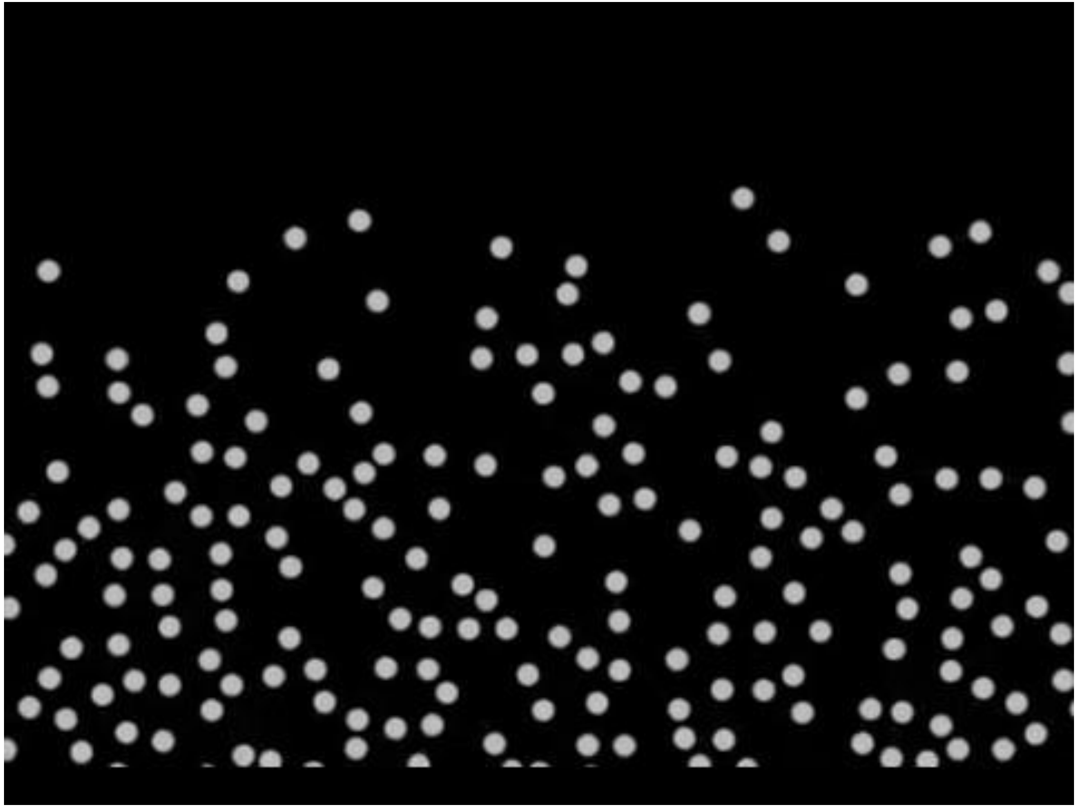
- driven to lower energy states

```
force = - grad(energy)
```

- driven to the more frequently occuring configurations



Statistical Mechanics



Statistical Mechanics

Particles are:

- driven to lower energy states

```
min(energy)
```

- driven to the more frequently occuring configurations

```
max(log(n_configurations))
```

Statistical Mechanics

Particles are:

- driven to lower energy states

```
min(energy)
```

- driven to the more frequently occuring configurations

```
max(entropy)
```

source: twitter.com/acuriouswavefn

Statistical Mechanics

Particles are:

- driven to lower energy states

```
min(energy)
```

- driven to the more frequently occuring configurations

```
max(entropy)
```

- the ‘dial’ between these two forces is the temperature

```
min(energy - T * entropy)
```

Statistical Mechanics

```
min(energy - T * entropy)

(. . .)
```

`P(w) = exp(- energy(w) / T) / Z`

Statistical Mechanics

`min(energy - T * entropy)`

`(. . .)`

`P(w) = exp(- energy(w) / T) / Z`

Langevin Dynamics:

$w_{t+dt} = w_t - dt * \mathrm{grad}(\mathrm{energy}) + \mathbf{sqrt(dt * T) * noise}$

Statistical Physics of Learning

Statistical Physics	**Statistical Learning**
State of particles	Parameters: **w**

Statistical Physics of Learning

Statistical Physics	**Statistical Learning**
State of particles	Parameters: **w**
Energy	Loss

Statistical Physics of Learning

Statistical Physics	**Statistical Learning**
State of particles	Parameters: **w**
Energy	Loss
Relaxation to equilibrium	Loss minimization

Statistical Physics of Learning

Statistical Physics	**Statistical Learning**
State of particles	Parameters: $\mathbf{w}$
Energy	Loss
Relaxation to equilibrium	Loss minimization
Maximum entropy	...

Stochastic Gradient Descent (SGD):

$$w_{t+1} = w_t - \text{lambda} * \text{grad(loss)}$$

Stochastic Gradient Langevin Dynamics (SGLD) (Welling and Teh, 2011):

$$w_{t+1} = w_t - \text{lambda} * \text{grad(loss)} + \text{sqrt(lambda)} * \text{noise}$$

Welling, Max, and Yee W. Teh. 2011. “Bayesian Learning via Stochastic Gradient Langevin Dynamics.” In Proceedings of the 28th International Conference on Machine Learning (ICML-11), 681–88.

Implementing SGLD

- PyTorch SGD: (https://github.com/pytorch/pytorch/blob/master/torch/optim/sgd.py)

```
d_p = p.grad.data

p.data.add_(-group['lr'], d_p)
```

Implementing SGLD

- PyTorch SGD: (https://github.com/pytorch/pytorch/blob/master/torch/optim/sgd.py)

```
d_p = p.grad.data

p.data.add_(-group['lr'], d_p)
```

- PyTorch SGLD:

```
size = d_p.size()
langevin_noise = Normal(torch.zeros(size),
                         torch.ones(size))


p.data.add_(- np.sqrt(lr), np.sqrt(lr) * d_p + langevin_noise.sample())
```

Results - out of sample image detection

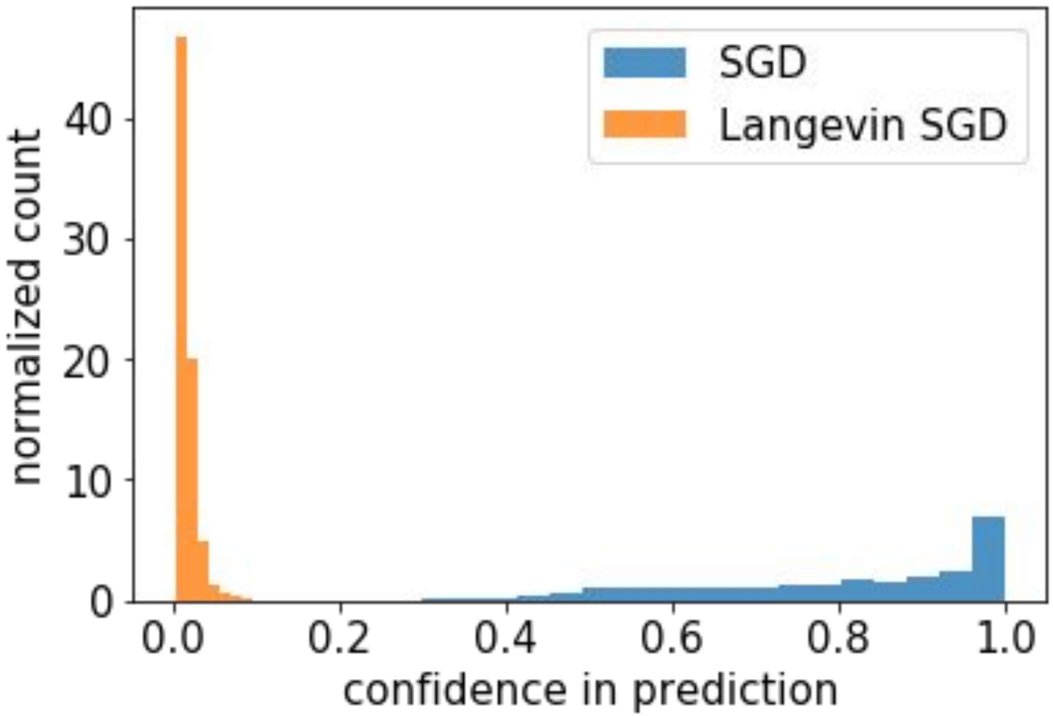
- 60,000 training images
- 10,000 test images

LeCun, Yann, Corinna Cortes, and C. J. Burges. 2010. “MNIST Handwritten Digit Database.” AT&T Labs [Online]. Available: Http://yann. Lecun. Com/exdb/mnist 2.

- notMNIST: 20,000 images

Bulatov, Yaroslav. n.d. “notMNIST Dataset.” Accessed April 24, 2018. http://yaroslavvb.blogspot.com/2011/09/notmnist-dataset.html.

Results - out of sample image detection



Code, links, blog posts, and more:

https://github.com/henripal/sgld

Thank you!

Acknowledgements:
Advisor: Prof. Henry Hess

Funding:
ARO grant “Thermodynamics of Statistical Learning”,
PI: H. Hess, Army Research Office
W911-NF-17-1-0107