Indian Institute of Information Technology, Allahabad

CVBL IIIT Allahabad

Neural Networks
Optimization and Regularization

By

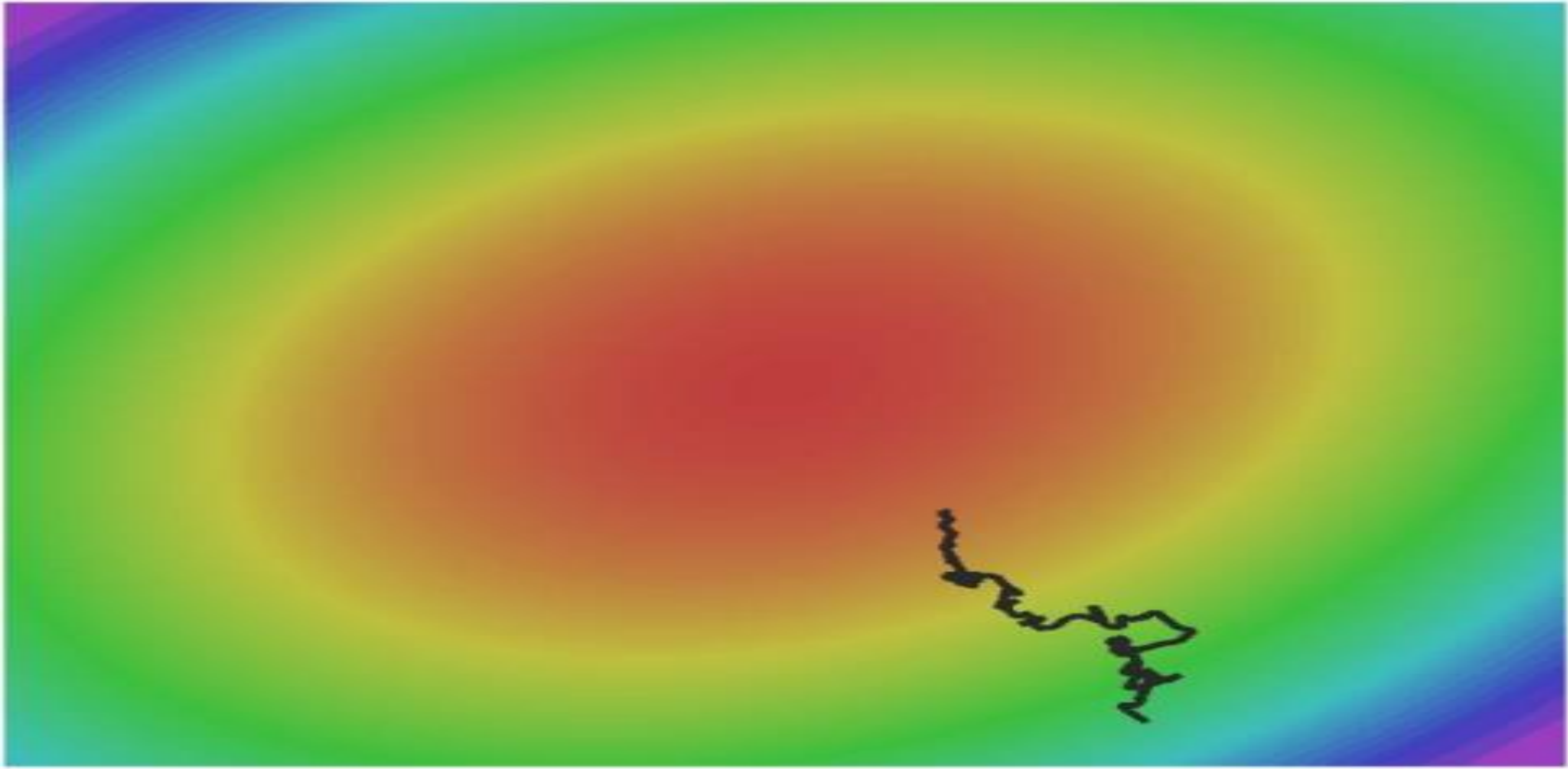
Dr. Shiv Ram Dubey
Assistant Professor
Computer Vision And Biometrics Lab (CVBL)
Department Of Information Technology
Indian Institute Of Information Technology, Allahabad

Email: srdubey@iiita.ac.in Web: https://profile.iiita.ac.in/srdubey/

DISCLAIMER

The content (text, image, and graphics) used in this slide are adopted from many sources for Academic purposes. Broadly, the sources have been given due credit appropriately. However, there is a chance of missing out some original primary sources. The authors of this material do not claim any copyright of such material.

Optimization



Optimization

Source: cs231n

MINI-BATCH SGD

Loop:
1. **Sample** a batch of data
2. **Forward** prop it through the graph (network), get loss
3. **Backprop** to calculate the gradients
4. **Update** the parameters using the gradient

Source: cs231n

STOCHASTIC GRADIENT DESCENT (SGD)

The procedure of repeatedly evaluating the gradient of loss function and then performing a parameter update.

Vanilla (Original) Gradient Descent:

```
while True:
  dx = compute_gradient(x)
  x -= learning_rate * dx
```

Source: cs231n

SGD: PROBLEMS

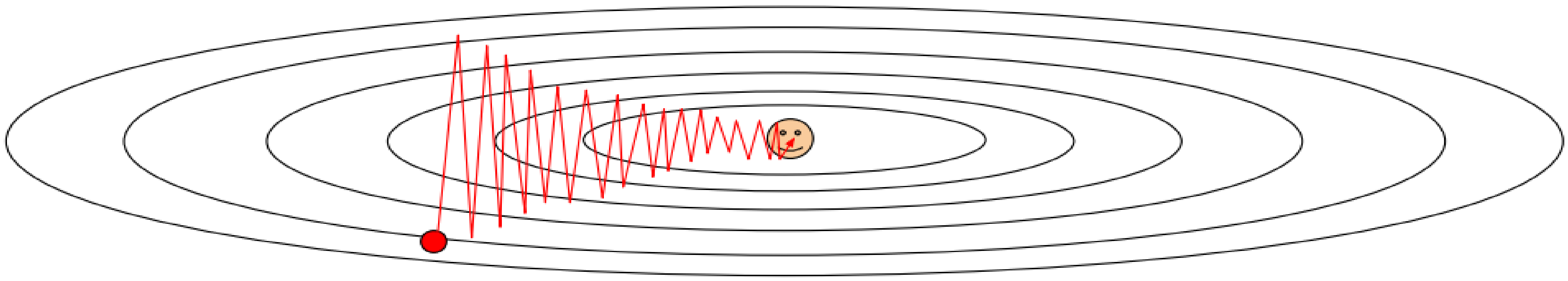
What if loss changes quickly in one direction and slowly in another?

Source: cs231n

SGD: PROBLEMS

What if loss changes quickly in one direction and slowly in another?

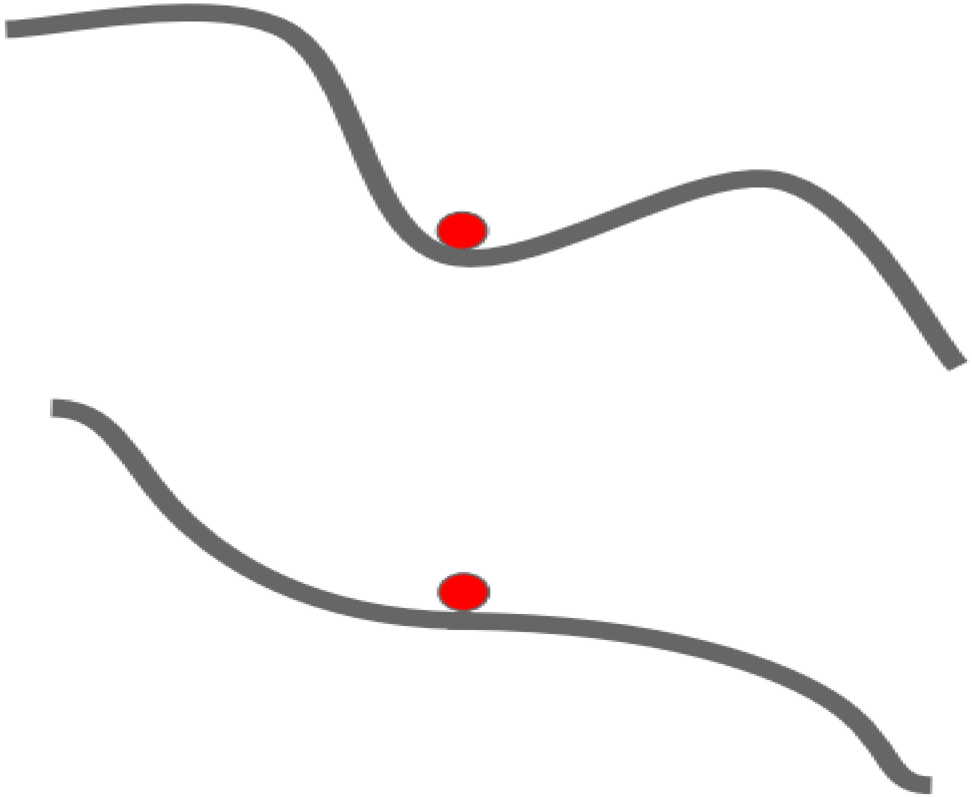
Very slow progress along shallow dimension, jitter along steep direction



Source: cs231n

SGD: PROBLEMS

What if the loss function has a **local minima** or **saddle point**?

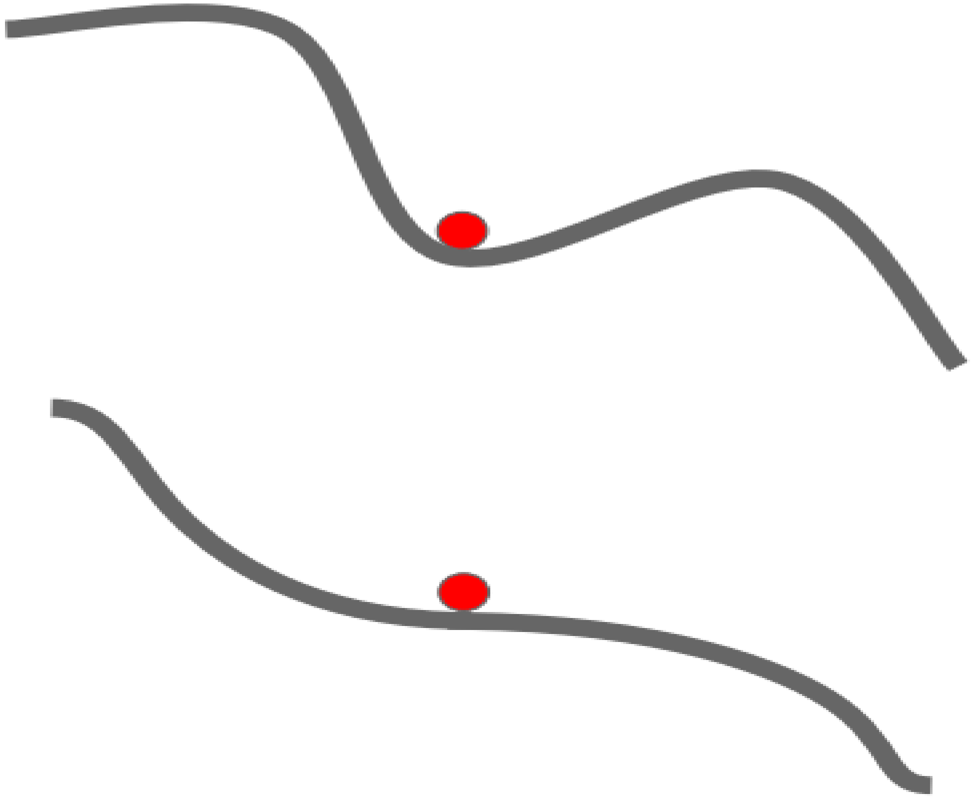


Source: cs231n

SGD: PROBLEMS

What if the loss function has a **local minima** or **saddle point**?

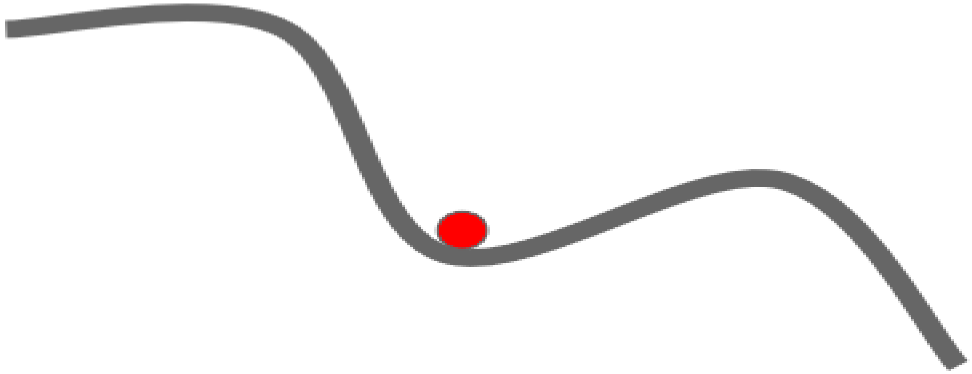
Zero gradient, gradient descent gets stuck



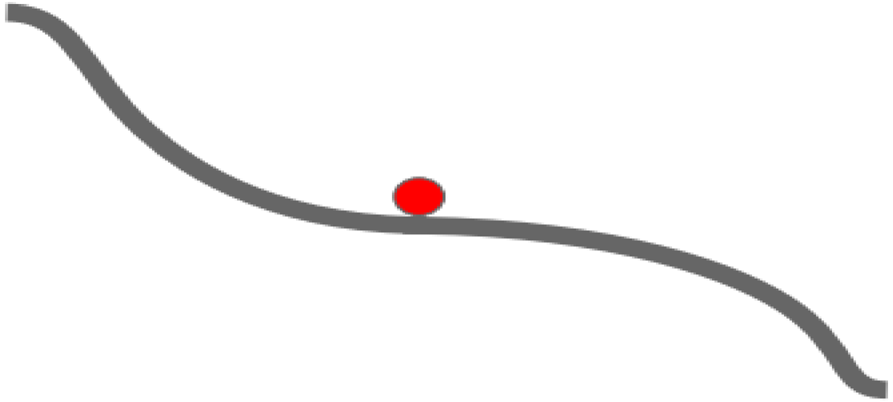
Source: cs231n

SGD: PROBLEMS

What if the loss function has a **local minima** or **saddle point**?



Zero gradient, gradient descent gets stuck



Saddle points much more common in high dimension

Dauphin et al, "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization", NIPS 2014

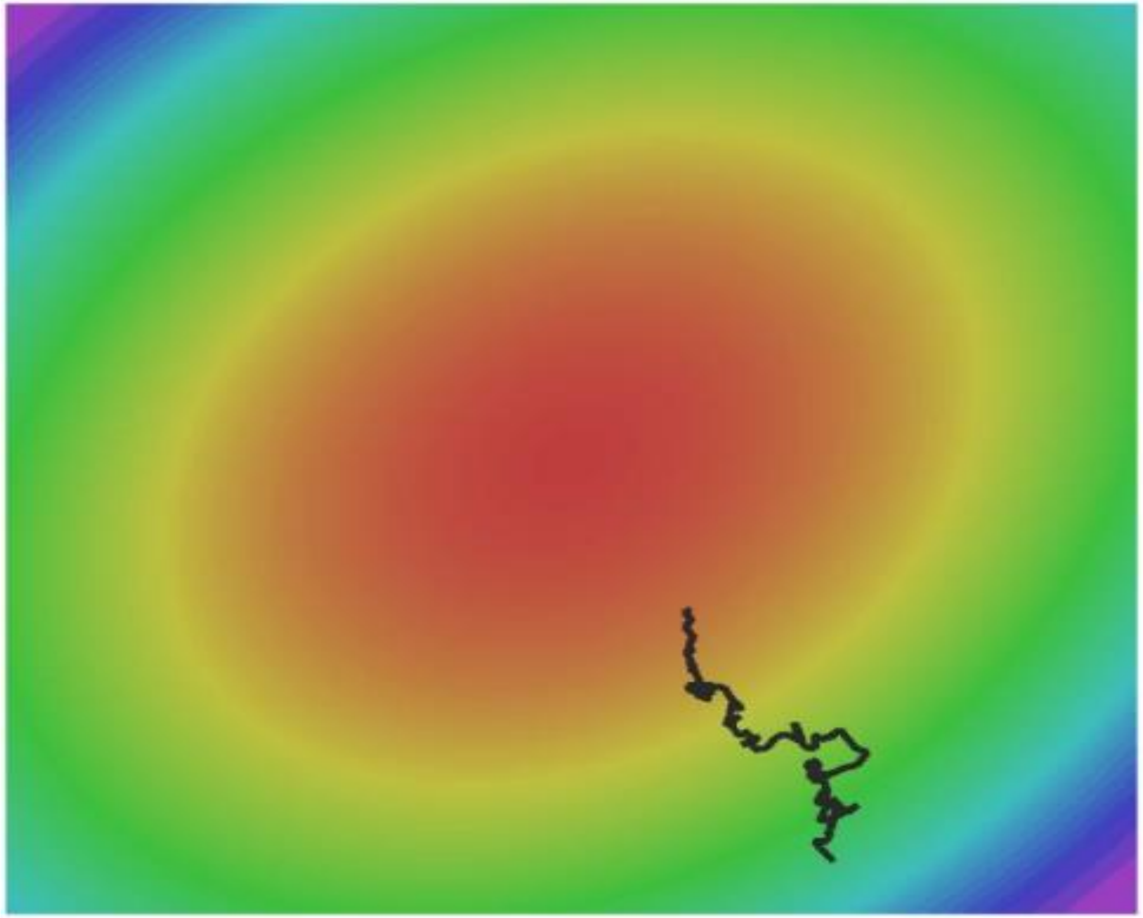
Source: cs231n

SGD: PROBLEMS

Our gradients come from **minibatches** so they can be **noisy!**

$$L(W) = \frac{1}{N} \sum_{i=1}^{N} L_i(x_i, y_i, W)$$

$$\nabla_W L(W) = \frac{1}{N} \sum_{i=1}^{N} \nabla_W L_i(x_i, y_i, W)$$



Source: cs231n

SGD + MOMENTUM

SGD

$$x_{t+1} = x_t - \alpha \nabla f(x_t)$$

```
while True:
  dx = compute_gradient(x)
  x -= learning_rate * dx
```

Source: cs231n

SGD + MOMENTUM

SGD

$$x_{t+1} = x_t - \alpha \nabla f(x_t)$$

```
while True:
  dx = compute_gradient(x)
  x -= learning_rate * dx
```

SGD+Momentum

$$v_{t+1} = \rho v_t + \nabla f(x_t)$$
$$x_{t+1} = x_t - \alpha v_{t+1}$$

Source: cs231n

SGD + MOMENTUM

SGD

$$x_{t+1} = x_t - \alpha \nabla f(x_t)$$

```
while True:
  dx = compute_gradient(x)
  x -= learning_rate * dx
```

SGD+Momentum

$$v_{t+1} = \rho v_t + \nabla f(x_t)$$
$$x_{t+1} = x_t - \alpha v_{t+1}$$

```
vx = 0
while True:
  dx = compute_gradient(x)
  vx = rho * vx + dx
  x -= learning_rate * vx
```

Source: cs231n

SGD + MOMENTUM

SGD

$$x_{t+1} = x_t - \alpha \nabla f(x_t)$$

```
while True:
  dx = compute_gradient(x)
  x -= learning_rate * dx
```

SGD+Momentum

$$v_{t+1} = \rho v_t + \nabla f(x_t)$$
$$x_{t+1} = x_t - \alpha v_{t+1}$$

```
vx = 0
while True:
  dx = compute_gradient(x)
  vx = rho * vx + dx
  x -= learning_rate * vx
```

- Build up "velocity" in any direction that has consistent gradient
- Rho gives "friction"; typically rho=0.9 or 0.99

Source: cs231n

SGD + MOMENTUM

SGD

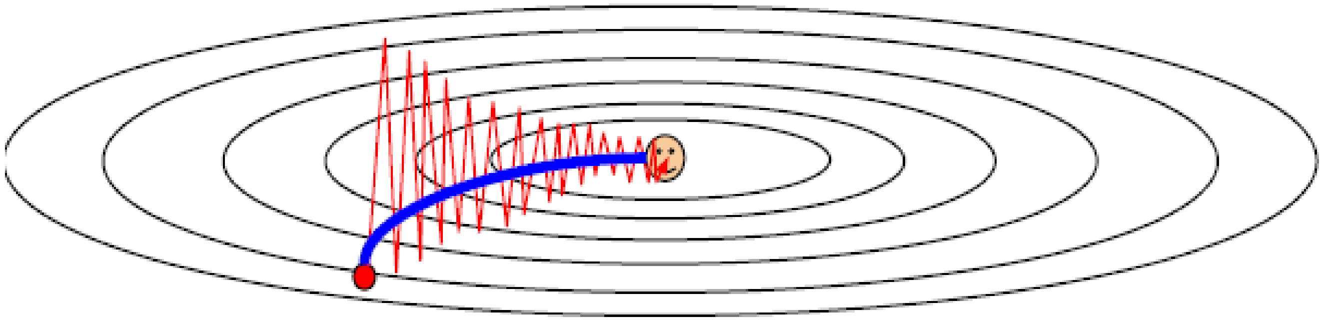
$$x_{t+1} = x_t - \alpha \nabla f(x_t)$$

```
while True:
    dx = compute_gradient(x)
    x -= learning_rate * dx
```

SGD+Momentum

$$v_{t+1} = \rho v_t + \nabla f(x_t)$$
$$x_{t+1} = x_t - \alpha v_{t+1}$$

```
vx = 0
while True:
    dx = compute_gradient(x)
    vx = rho * vx + dx
    x -= learning_rate * vx
```



Source: cs231n

ADAGRAD

```
grad_squared = 0
while True:
  dx = compute_gradient(x)
  grad_squared += dx * dx
  x -= learning_rate * dx / (np.sqrt(grad_squared) + 1e-7)
```

Added element-wise scaling of the gradient based on the historical sum of squares in each dimension

Duchi et al, "Adaptive subgradient methods for online learning and stochastic optimization", JMLR 2011

Source: cs231n

ADAGRAD

```
grad_squared = 0
while True:
  dx = compute_gradient(x)
  grad_squared += dx * dx
  x -= learning_rate * dx / (np.sqrt(grad_squared) + 1e-7)
```

What happens to the step size over long time?

Duchi et al, "Adaptive subgradient methods for online learning and stochastic optimization", JMLR 2011

Source: cs231n

ADAGRAD

```
grad_squared = 0
while True:
  dx = compute_gradient(x)
  grad_squared += dx * dx
  x -= learning_rate * dx / (np.sqrt(grad_squared) + 1e-7)
```

What happens to the step size over long time?

Effective learning rate diminishing problem

Duchi et al, "Adaptive subgradient methods for online learning and stochastic optimization", JMLR 2011

Source: cs231n

RMSPROP

AdaGrad

```
grad_squared = 0
while True:
  dx = compute_gradient(x)
  grad_squared += dx * dx
  x -= learning_rate * dx / (np.sqrt(grad_squared) + 1e-7)
```

RMSProp

```
grad_squared = 0
while True:
  dx = compute_gradient(x)
  grad_squared = decay_rate * grad_squared + (1 - decay_rate) * dx * dx
  x -= learning_rate * dx / (np.sqrt(grad_squared) + 1e-7)
```

Tieleman and Hinton, 2012

Source: cs231n

ADAM

Kingma and Ba, "Adam: A method for stochastic optimization", ICLR 2015

```
first_moment = 0
second_moment = 0
while True:
  dx = compute_gradient(x)
  first_moment = beta1 * first_moment  + (1 - beta1) * dx
  second_moment = beta2 * second_moment + (1 - beta2) * dx * dx
  x -= learning_rate * first_moment / (np.sqrt(second_moment) + 1e-7))
```

Source: cs231n

ADAM

Kingma and Ba, "Adam: A method for stochastic optimization", ICLR 2015

```
first_moment = 0
second_moment = 0
while True:
  dx = compute_gradient(x)
  first_moment = beta1 * first_moment  + (1 - beta1) * dx
  second_moment = beta2 * second_moment + (1 - beta2) * dx * dx
  x -= learning_rate * first_moment / (np.sqrt(second_moment) + 1e-7))
```

Sort of like RMSProp with Momentum

Source: cs231n

ADAM

Kingma and Ba, "Adam: A method for stochastic optimization", ICLR 2015

```
first_moment = 0
second_moment = 0
while True:
  dx = compute_gradient(x)
  first_moment = beta1 * first_moment  + (1 - beta1) * dx
  second_moment = beta2 * second_moment + (1 - beta2) * dx * dx
  x -= learning_rate * first_moment / (np.sqrt(second_moment) + 1e-7))
```

Sort of like RMSProp with Momentum

Problem:

Initially, second_moment=0 and beta2=0.999

After 1st iteration, second_moment -> close to zero

So, very large step for update of x

Source: cs231n

ADAM (WITH BIAS CORRECTION)

Kingma and Ba, "Adam: A method for stochastic optimization", ICLR 2015

```
first_moment = 0
second_moment = 0
for t in range(num_iterations):
  dx = compute_gradient(x)
  first_moment = beta1 * first_moment  + (1 - beta1) * dx
  second_moment = beta2 * second_moment + (1 - beta2) * dx * dx
  first_unbias = first_moment / (1 - beta1 ** t)
  second_unbias = second_moment / (1 - beta2 ** t)
  x -= learning_rate * first_unbias / (np.sqrt(second_unbias) + 1e-7))
```

AdaGrad/ RMSProp

Bias Correction

Momentum

Bias correction for the fact that first and second moment estimates start at zero

Source: cs231n

ADAM (WITH BIAS CORRECTION)

Kingma and Ba, "Adam: A method for stochastic optimization", ICLR 2015

```
first_moment = 0
second_moment = 0
for t in range(num_iterations):
  dx = compute_gradient(x)
  first_moment = beta1 * first_moment  + (1 - beta1) * dx
  second_moment = beta2 * second_moment + (1 - beta2) * dx * dx
  first_unbias = first_moment / (1 - beta1 ** t)
  second_unbias = second_moment / (1 - beta2 ** t)
  x -= learning_rate * first_unbias / (np.sqrt(second_unbias) + 1e-7))
```

AdaGrad/ RMSProp

Bias Correction

Momentum

Bias correction for the fact that first and second moment estimates start at zero

Adam with **beta1 = 0.9**, **beta2 = 0.999**, and **learning_rate = 1e-3 or 5e-4** is a **great starting point** for many models!

Source: cs231n

MAJOR PROBLEM WITH ADAM

- Does not the optimization trajectory information such as short term gradient behavior
- Overshoots the optima
- Oscillates near the optima

RECENT SGD BASED OPTIMIZERS

- diffGrad (IEEE TNNLS 2020) – by us
- AdaBelief (NeurIPS 2020)
- Rectified Adam (RADAM) (ICLR 2020)
- Moment Centralization SGD (CVMI 2022) – by us
- AdaInject (IEEE TAI 2022) – by us
- AdaNorm (WACV 2023) – by us

and many more…. still a challenging problem.

https://pythonawesome.com/a-collection-of-optimizers-for-pytorch/

DIFFGRAD OPTIMIZER

Solves the previously mentioned problems by incorporating the local gradient change as friction in effective learning rate.

High local gradient change → low friction → high learning rate

Small local gradient change → high friction → slow learning rate

S.R. Dubey et al. (2020), diffGrad: An Optimization Method for Convolutional Neural Networks, IEEE Transactions on Neural Network and Learning Systems.

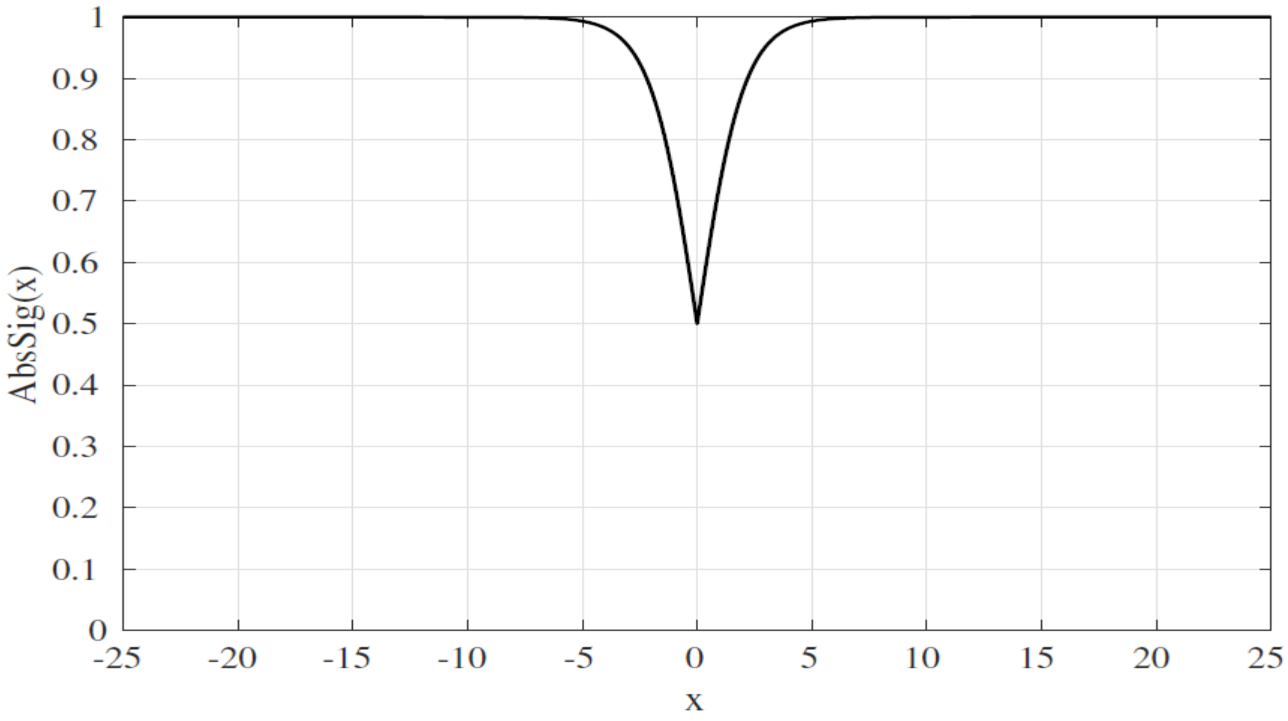
DIFFGRAD OPTIMIZER

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\alpha_t \times \xi_{t,i} \times \hat{m}_{t,i}}{\sqrt{\hat{v}_{t,i}} + \epsilon}$$

$$\xi_{t,i} = AbsSig(\Delta g_{t,i})$$

$$AbsSig(x) = \frac{1}{1 + e^{-|x|}}$$

$$\Delta g_{t,i} = g_{t-1,i} - g_{t,i}$$



S.R. Dubey et al. (2020), diffGrad: An Optimization Method for Convolutional Neural Networks, IEEE Transactions on Neural Network and Learning Systems.

ADABELIEF OPTIMIZER

Algorithm 1: Adam Optimizer

Initialize θ_0, $m_0 \leftarrow 0$, $v_0 \leftarrow 0$, $t \leftarrow 0$
While θ_t not converged

$t \leftarrow t + 1$

$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$

$m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t$

$v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$

Bias Correction

$$\widehat{m_t} \leftarrow \frac{m_t}{1-\beta_1^t}, \widehat{v_t} \leftarrow \frac{v_t}{1-\beta_2^t}$$

Update

$$\theta_t \leftarrow \prod_{\mathcal{F},\sqrt{\widehat{v_t}}} \left(\theta_{t-1} - \frac{\alpha \widehat{m_t}}{\sqrt{\widehat{v_t}}+\epsilon}\right)$$

Algorithm 2: AdaBelief Optimizer

Initialize θ_0, $m_0 \leftarrow 0$, $s_0 \leftarrow 0$, $t \leftarrow 0$
While θ_t not converged

$t \leftarrow t + 1$

$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$

$m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t$

$s_t \leftarrow \beta_2 s_{t-1} + (1 - \beta_2)(g_t - m_t)^2 + \epsilon$

Bias Correction

$$\widehat{m_t} \leftarrow \frac{m_t}{1-\beta_1^t}, \widehat{s_t} \leftarrow \frac{s_t}{1-\beta_2^t}$$

Update

$$\theta_t \leftarrow \prod_{\mathcal{F},\sqrt{\widehat{s_t}}} \left(\theta_{t-1} - \frac{\alpha \widehat{m_t}}{\sqrt{\widehat{s_t}}+\epsilon}\right)$$

Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. *Advances in neural information processing systems*, 2020.

AdaInject Optimizer

Algorithm 1: Adam Optimizer

Initialize: $\theta_0, m_0 \leftarrow 0, v_0 \leftarrow 0, t \leftarrow 0$
Hyperparameters: α, β_1, β_2
While θ_t not converged
$t \leftarrow t+1$
$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$
$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g_t$
$v_t \leftarrow \beta_2 \cdot v_{t-1} + (1-\beta_2) \cdot g_t^2$
Bias Correction
$\widehat{m_t} \leftarrow m_t/(1-\beta_1^t)$, $\widehat{v_t} \leftarrow v_t/(1-\beta_2^t)$
Update
$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \widehat{m_t}/(\sqrt{\widehat{v_t}} + \epsilon)$

Algorithm 2: AdamInject (i.e., Adam + AdaInject) Optimizer

Initialize: $\theta_0, s_0 \leftarrow 0, v_0 \leftarrow 0, t \leftarrow 0$
Hyperparameters: $\alpha, \beta_1, \beta_2, k$
While θ_t not converged
$t \leftarrow t+1$
$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$
If t = 1
$s_t \leftarrow \beta_1 \cdot s_{t-1} + (1-\beta_1) \cdot g_t$
Else
$\Delta\theta \leftarrow \theta_{t-2} - \theta_{t-1}$
$s_t \leftarrow \beta_1 \cdot s_{t-1} + (1-\beta_1) \cdot (g_t + \Delta\theta \cdot g_t^2)/k$
$v_t \leftarrow \beta_2 \cdot v_{t-1} + (1-\beta_2) \cdot g_t^2$
Bias Correction
$\widehat{s_t} \leftarrow s_t/(1-\beta_1^t)$, $\widehat{v_t} \leftarrow v_t/(1-\beta_2^t)$
Update
$\theta_t \leftarrow \theta_{t-1} - \alpha \widehat{s_t}/(\sqrt{\widehat{v_t}} + \epsilon)$

AdaInject: Injection Based Adaptive Gradient Descent Optimizers for Convolutional Neural Networks. *IEEE TAI 2022*.

AdaNorm Optimizer

Algorithm 1: Adam Optimizer

Initialize: $\boldsymbol{\theta}_0, \boldsymbol{m}_0 \leftarrow 0, \boldsymbol{v}_0 \leftarrow 0, t \leftarrow 0$
Hyperparameters: α, β_1, β_2
While $\boldsymbol{\theta}_t$ not converged
$t \leftarrow t + 1$
$\boldsymbol{g}_t \leftarrow \nabla_\theta f_t(\boldsymbol{\theta}_{t-1})$
$\boldsymbol{m}_t \leftarrow \beta_1 \boldsymbol{m}_{t-1} + (1 - \beta_1)\boldsymbol{g}_t$
$\boldsymbol{v}_t \leftarrow \beta_2 \boldsymbol{v}_{t-1} + (1 - \beta_2)\boldsymbol{g}_t^2$
Bias Correction
$\widehat{\boldsymbol{m}_t} \leftarrow \boldsymbol{m}_t/(1 - \beta_1^t)$, $\widehat{\boldsymbol{v}_t} \leftarrow \boldsymbol{v}_t/(1 - \beta_2^t)$
Update
$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \alpha\widehat{\boldsymbol{m}_t}/(\sqrt{\widehat{\boldsymbol{v}_t}} + \epsilon)$

Algorithm 2: AdamNorm Optimizer

Initialize: $\boldsymbol{\theta}_0, \boldsymbol{m}_0 \leftarrow 0, \boldsymbol{v}_0 \leftarrow 0, e_0 \leftarrow 0, t \leftarrow 0$
Hyperparameters: $\alpha, \beta_1, \beta_2, \gamma$
While $\boldsymbol{\theta}_t$ not converged
$t \leftarrow t + 1$
$\boldsymbol{g}_t \leftarrow \nabla_\theta f_t(\boldsymbol{\theta}_{t-1})$
$g_{norm} \leftarrow L_2 Norm(\boldsymbol{g}_t)$
$e_t = \gamma e_{t-1} + (1 - \gamma) g_{norm}$
$\boldsymbol{s}_t = \boldsymbol{g}_t$
If $e_t > g_{norm}$
$\boldsymbol{s}_t = (e_t/g_{norm})\boldsymbol{g}_t$
$\boldsymbol{m}_t \leftarrow \beta_1 \boldsymbol{m}_{t-1} + (1 - \beta_1)\boldsymbol{s}_t$
$\boldsymbol{v}_t \leftarrow \beta_2 \boldsymbol{v}_{t-1} + (1 - \beta_2)\boldsymbol{g}_t^2$
Bias Correction
$\widehat{\boldsymbol{m}_t} \leftarrow \boldsymbol{m}_t/(1 - \beta_1^t)$, $\widehat{\boldsymbol{v}_t} \leftarrow \boldsymbol{v}_t/(1 - \beta_2^t)$
Update
$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \alpha\widehat{\boldsymbol{m}_t}/(\sqrt{\widehat{\boldsymbol{v}_t}} + \epsilon)$

AdaNorm: Adaptive Gradient Norm Correction based Optimizer for CNNs. *WACV* 2023.

ADANORM OPTIMIZER

Table 1. Classification results in terms of accuracy (%) on CIFAR10, CIFAR100 and TinyImageNet datasets using Adam, diffGrad, Radam and AdaBelief without and with the proposed AdaNorm technique. The value of γ is set to 0.95 in this experiment and results are computed as an average over three independent runs.

CNN Models	Classification accuracy (%) using different optimizers without and with AdaNorm							
	Adam		diffGrad		Radam		AdaBelief	
	Adam	AdamNorm	diffGrad	diffGradNorm	Radam	RadamNorm	AdaBelief	AdaBeliefNorm
Results on CIFAR10 Dataset								
VGG16	92.55	**92.83** (↑ 0.30)	92.76	**92.87** (↑ 0.12)	92.94	**93.14** (↑ 0.22)	92.71	**92.81** (↑ 0.11)
ResNet18	93.54	**93.78** (↑ 0.26)	93.49	**93.98** (↑ 0.52)	93.82	**93.89** (↑ 0.07)	93.63	**93.66** (↑ 0.03)
ResNet50	93.83	**94.01** (↑ 0.19)	93.81	**94.23** (↑ 0.45)	94.14	**94.21** (↑ 0.07)	94.1	**94.16** (↑ 0.06)
Results on CIFAR100 Dataset								
VGG16	67.29	**69.15** (↑ 2.76)	68.19	**68.31** (↑ 0.18)	70.69	**70.77** (↑ 0.11)	68.92	**69.24** (↑ 0.46)
ResNet18	71.09	**73.11** (↑ 2.84)	73.5	**73.64** (↑ 0.19)	73.22	**73.34** (↑ 0.16)	72.72	**73.31** (↑ 0.81)
ResNet50	71.88	**75.53** (↑ 5.08)	75.06	**75.49** (↑ 0.57)	74.95	**75.39** (↑ 0.59)	**75.53**	75.49 (↓ 0.05)
Results on TinyImageNet Dataset								
VGG16	41.93	**44.67** (↑ 6.53)	42.91	**43.49** (↑ 1.35)	43.84	**45.02** (↑ 2.69)	44.23	**44.79** (↑ 1.27)
ResNet18	47.73	**49.57** (↑ 3.86)	49.34	**49.80** (↑ 0.93)	48.73	**50.50** (↑ 3.63)	49.25	**49.99** (↑ 1.50)
ResNet50	48.98	**54.44** (↑ 11.15)	51.32	**53.75** (↑ 4.73)	51.63	**52.87** (↑ 2.40)	53.57	**54.44** (↑ 1.62)

AdaNorm: Adaptive Gradient Norm Correction based Optimizer for CNNs. *WACV* 2023.

ADANORM OPTIMIZER

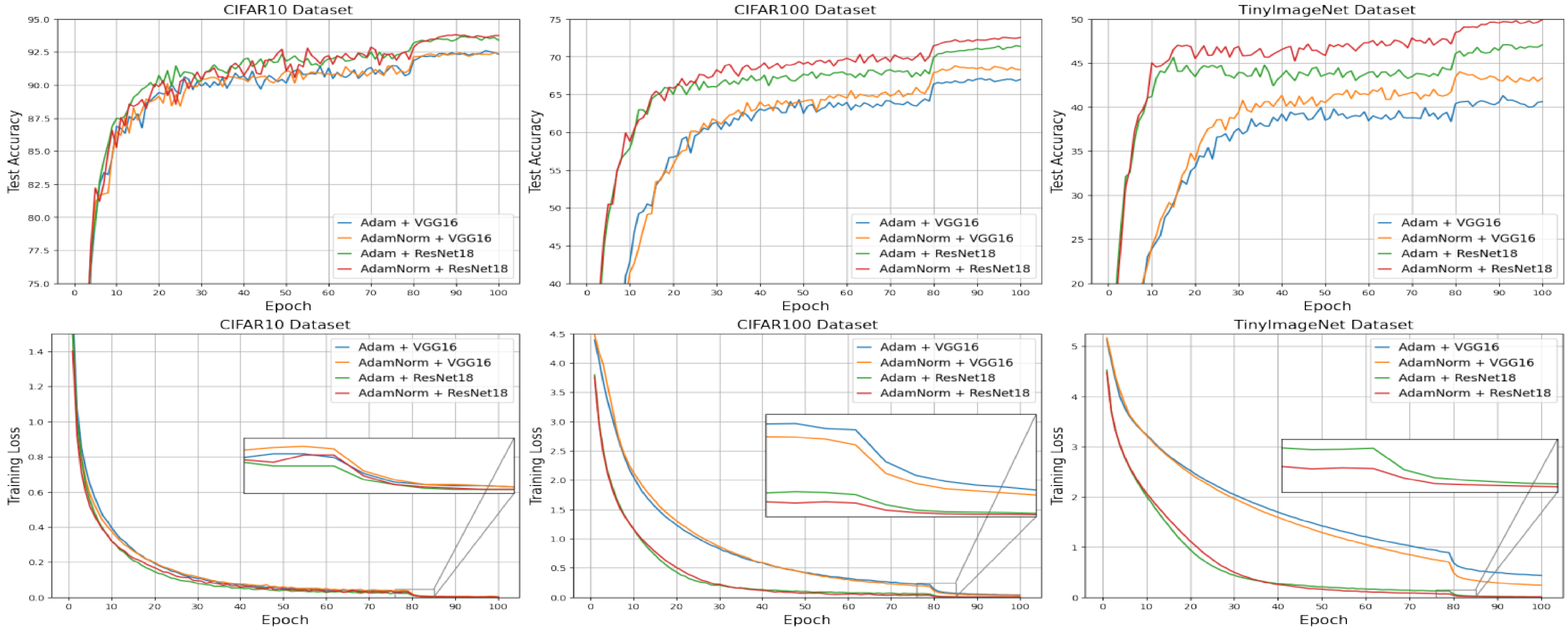


Figure 2. Test Accuracy (top row) and Training Loss (bottom row) vs Epoch plots using the Adam and AdamNorm optimizers for VGG16 and ResNet18 models on CIFAR10 (left), CIFAR100 (middle) and TinyImageNet (right) datasets. The value of γ is 0.95 in AdamNorm in this experiment. (Best viewed in color)

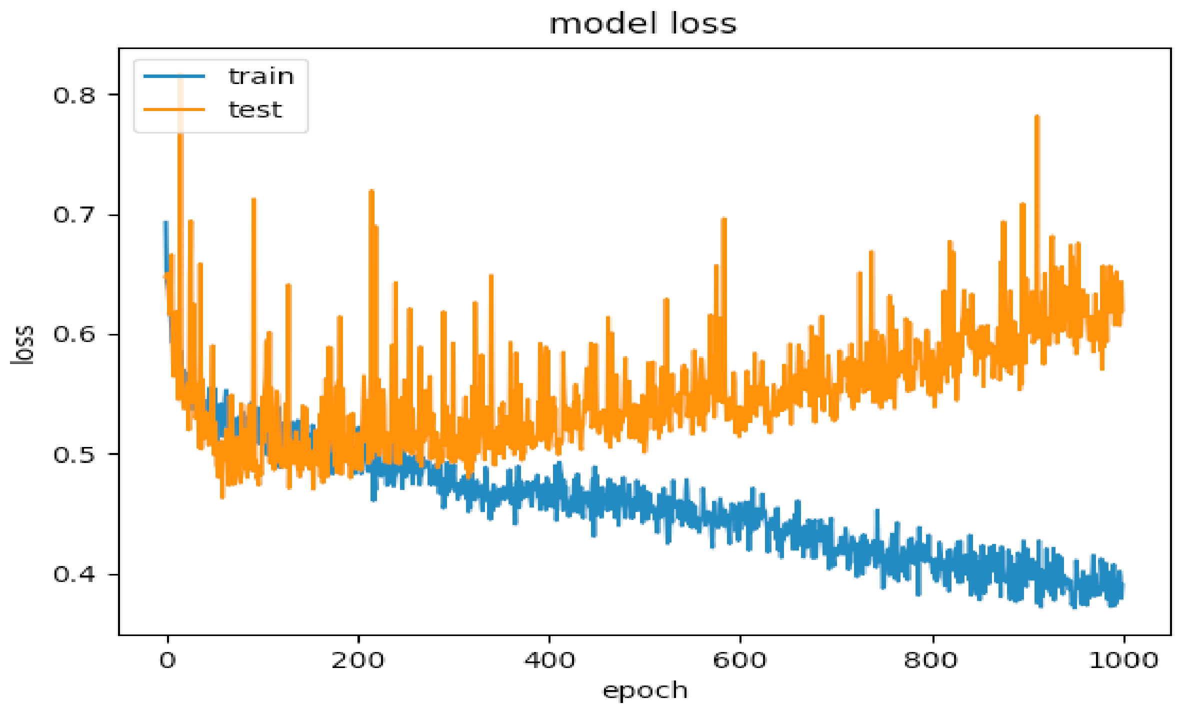
AdaNorm: Adaptive Gradient Norm Correction based Optimizer for CNNs. *WACV* 2023.

OPTIMIZER

In Practice:

- **Adam** is a good default choice in most cases
    - Try out diffGrad, RADAM, AdaBelief, AdaInject, and AdaNorm

More Optimizer: http://ruder.io/optimizing-gradient-descent/

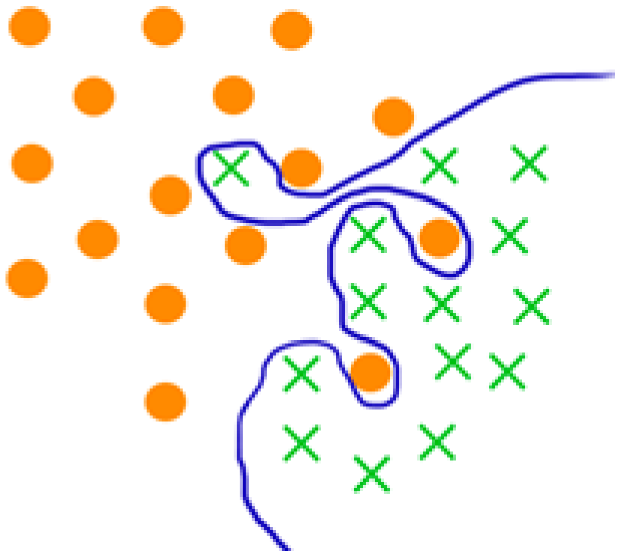


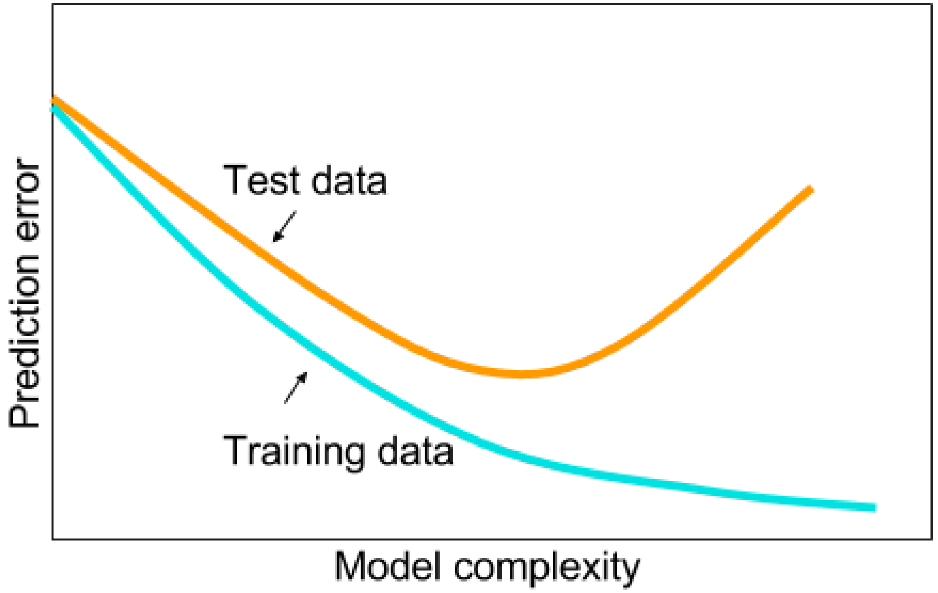
Regularization

Image Source: https://stackoverflow.com/questions/44909134/how-to-avoid-overfitting-on-a-simple-feed-forward-network/44985765

REGULARIZATION

- Techniques for controlling the capacity of a neural network to prevent overfitting





REGULARIZATION

$$\underbrace{L(W) = \frac{1}{N}\sum_{i=1}^{N} L_i(f(x_i, W), y_i)}$$

Data loss: Model predictions should match training data

Source: cs231n

REGULARIZATION

λ = regularization strength (hyperparameter)

$$L(W) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} L_i(f(x_i, W), y_i)} + \underbrace{\lambda R(W)}$$

Data loss: Model predictions should match training data

Regularization: Prevent the model from doing *too* well on training data

Source: cs231n

REGULARIZATION

λ = regularization strength (hyperparameter)

$$L(W) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} L_i(f(x_i, W), y_i)}_{} + \underbrace{\lambda R(W)}_{}$$

Data loss: Model predictions should match training data

Regularization: Prevent the model from doing *too* well on training data

Simple examples

<u>L2 regularization:</u> $R(W) = \sum_k \sum_l W_{k,l}^2$

L1 regularization: $R(W) = \sum_k \sum_l |W_{k,l}|$

Elastic net (L1 + L2): $R(W) = \sum_k \sum_l \beta W_{k,l}^2 + |W_{k,l}|$

Source: cs231n

REGULARIZATION

λ = regularization strength (hyperparameter)

$$L(W) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} L_i(f(x_i, W), y_i)}_{} + \underbrace{\lambda R(W)}_{}$$

Data loss: Model predictions should match training data

Regularization: Prevent the model from doing *too* well on training data

Why regularize?

- Express preferences over weights
- Make the model *simple* so it works on test data
- Improve optimization by adding curvature

Source: cs231n

REGULARIZATION

$$x = [1,1,1,1]$$

$$w_1 = [1,0,0,0]$$

$$w_2 = [0.25,0.25,0.25,0.25]$$

Source: cs231n

REGULARIZATION

$$x = [1,1,1,1]$$

$$w_1 = [1,0,0,0]$$

$$w_2 = [0.25,0.25,0.25,0.25]$$

$$w_1 \, . x = w_2 \, . x = 1$$

Source: cs231n

REGULARIZATION

$$x = [1,1,1,1]$$

$$w_1 = [1,0,0,0]$$

$$w_2 = [0.25,0.25,0.25,0.25]$$

$$w_1 \, . x = w_2 \, . x = 1$$

Which W to consider?

Source: cs231n

REGULARIZATION

$$x = [1,1,1,1]$$

$$w_1 = [1,0,0,0]$$

$$w_2 = [0.25,0.25,0.25,0.25]$$

$$w_1 \, . x = w_2 \, . x = 1$$

L2 Regularization

$$R(W) = \sum_k \sum_l W_{k,l}^2$$

Source: cs231n

REGULARIZATION

$$x = [1,1,1,1]$$

$$w_1 = [1,0,0,0]$$

$$\boxed{w_2 = [0.25,0.25,0.25,0.25]}$$

L2 regularization likes to "spread out" the weights

$$w_1 \,.\, x = w_2 \,.\, x = 1$$

L2 Regularization

$$R(W) = \sum_k \sum_l W_{k,l}^2$$

Source: cs231n

OTHER TYPES OF REGULARIZATION

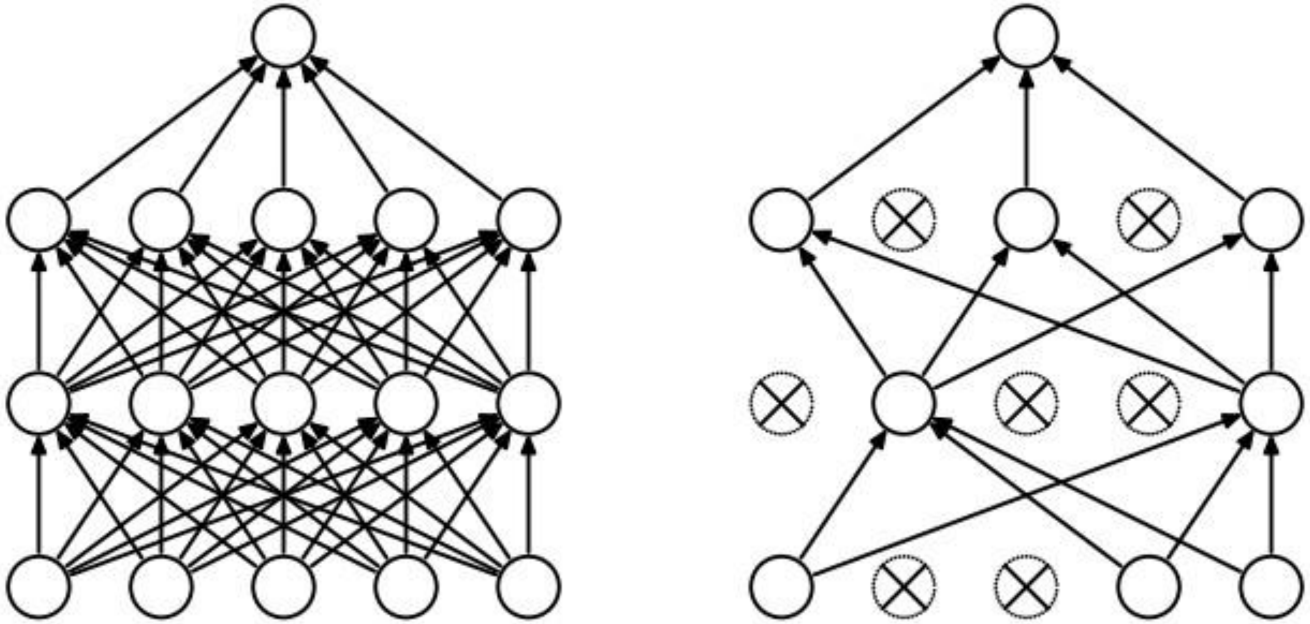
- Dropout
- Dropconnect
- Batch Normalization
- Data Augmentation

Dropout

DROPOUT

In each forward pass, randomly set some neurons to zero

Probability of dropping is a hyperparameter; 0.5 is common

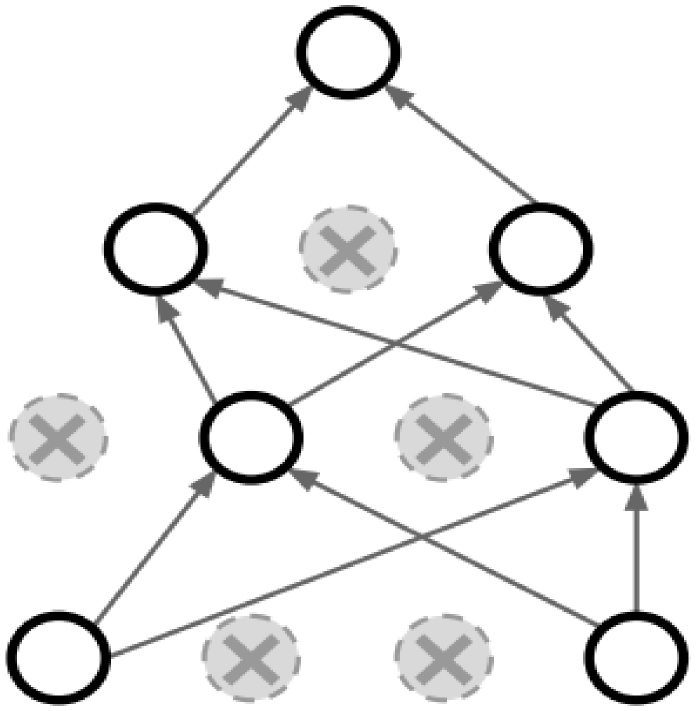


(a) Standard Neural Net (b) After applying dropout.

Dropout: A simple way to prevent neural networks from overfitting [Srivastava JMLR 2014] Source: cs231n

DROPOUT

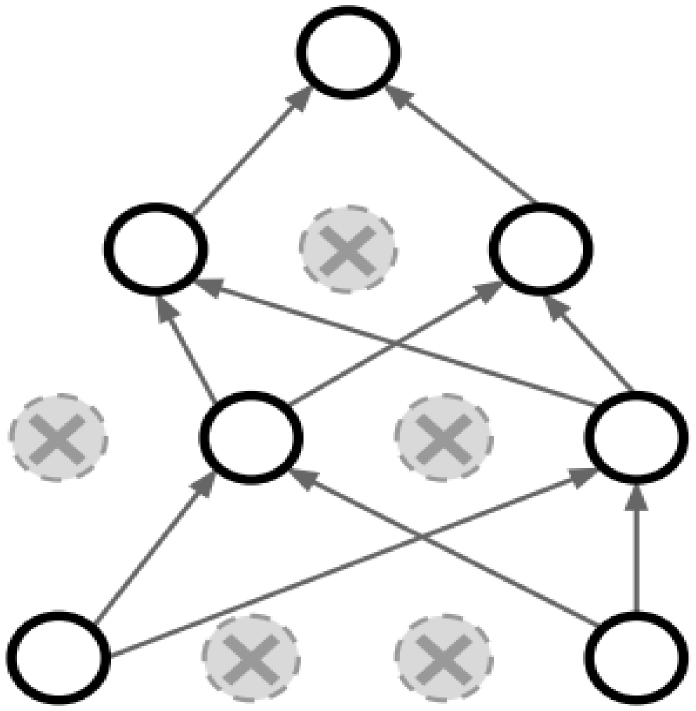
How can this possibly be a good idea?



Source: cs231n

DROPOUT

How can this possibly be a good idea?

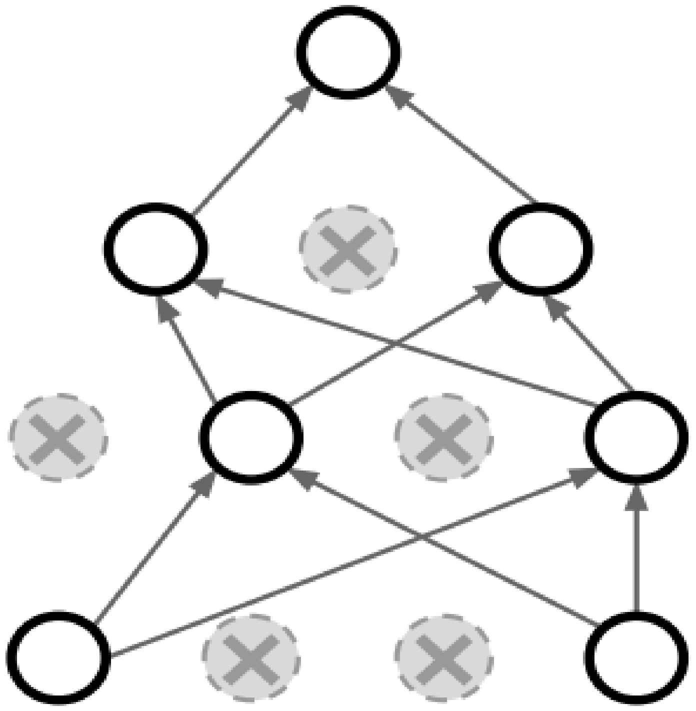


- Intuitions
  - Prevent "co-adaptation" of units, increase robustness to noise
  - Train *implicit ensemble*

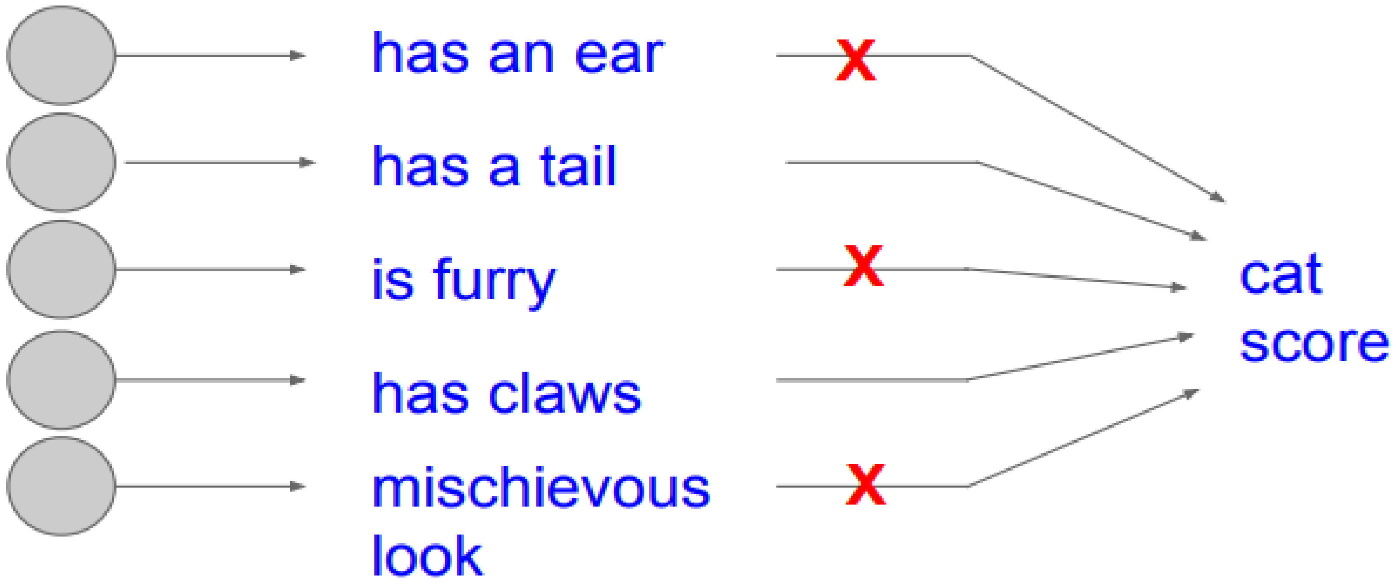
Source: cs231n

DROPOUT

How can this possibly be a good idea?



Forces the network to have a redundant representation;
Prevents co-adaptation of features



Source: cs231n

DROPOUT: TEST TIME

```python
def predict(X):
  # ensembled forward pass
  H1 = np.maximum(0, np.dot(W1, X) + b1) * p # NOTE: scale the activations
  H2 = np.maximum(0, np.dot(W2, H1) + b2) * p # NOTE: scale the activations
  out = np.dot(W3, H2) + b3
```

At test time all neurons are active always
=> We must scale the activations so that for each neuron:
<u>output at test time</u> = <u>expected output at training time</u>

More common: "Inverted dropout"

Source: cs231n

DROPOUT: MORE COMMON: "INVERTED DROPOUT"

We drop and scale at train time and don't do anything at test time.

```
p = 0.5 # probability of keeping a unit active. higher = less dropout

def train_step(X):
  # forward pass for example 3-layer neural network
  H1 = np.maximum(0, np.dot(W1, X) + b1)
  U1 = (np.random.rand(*H1.shape) < p) / p # first dropout mask. Notice /p!
  H1 *= U1 # drop!
  H2 = np.maximum(0, np.dot(W2, H1) + b2)
  U2 = (np.random.rand(*H2.shape) < p) / p # second dropout mask. Notice /p!
  H2 *= U2 # drop!
  out = np.dot(W3, H2) + b3

  # backward pass: compute gradients... (not shown)
  # perform parameter update... (not shown)

def predict(X):
  # ensembled forward pass
  H1 = np.maximum(0, np.dot(W1, X) + b1) # no scaling necessary
  H2 = np.maximum(0, np.dot(W2, H1) + b2)
  out = np.dot(W3, H2) + b3
```

test time is unchanged!

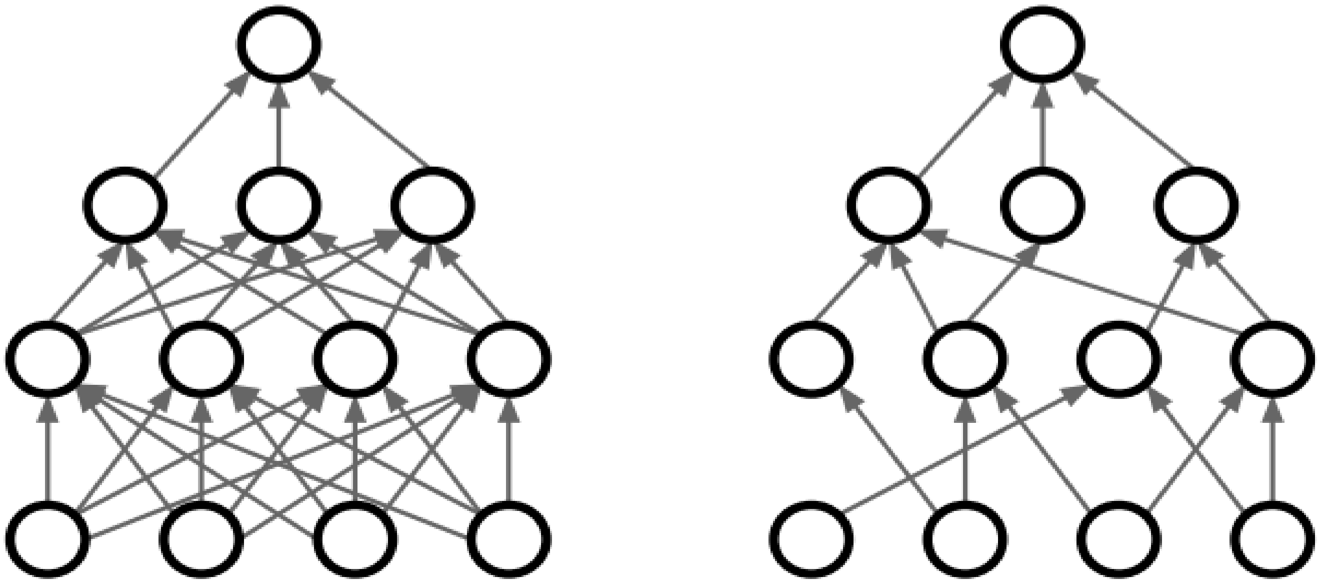
Source: cs231n

CURRENT STATUS OF DROPOUT

- Against
  - Slows down convergence
  - Made redundant by batch normalization or possibly even clashes with it
  - Unnecessary for larger datasets or with sufficient data augmentation
- In favor
  - Can still help in certain situations: e.g., used in Wide Residual Networks
  - Helpful in RNNs

DROPCONNECT

Dropping some connections



Wan et al, "Regularization of Neural Networks using DropConnect", ICML 2013

Source: cs231n

Batch Normalization

BATCH NORMALIZATION

"We want zero-mean unit-variance activations? lets make them so."

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [Ioffe and Szegedy 2015]

Source: cs231n

BATCH NORMALIZATION

"We want zero-mean unit-variance activations? lets make them so."

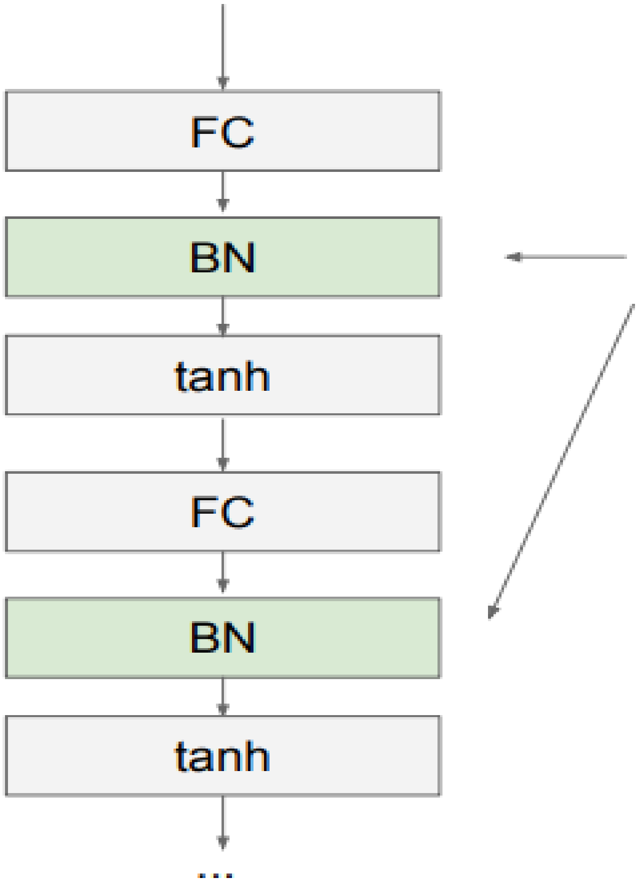
Consider a batch of activations at some layer.
To make each dimension zero-mean unit-variance, apply:

$$\widehat{x}^{(k)} = \frac{x^{(k)} - \mathrm{E}[x^{(k)}]}{\sqrt{\mathrm{Var}[x^{(k)}]}}$$

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [Ioffe and Szegedy 2015]

Source: cs231n

BATCH NORMALIZATION

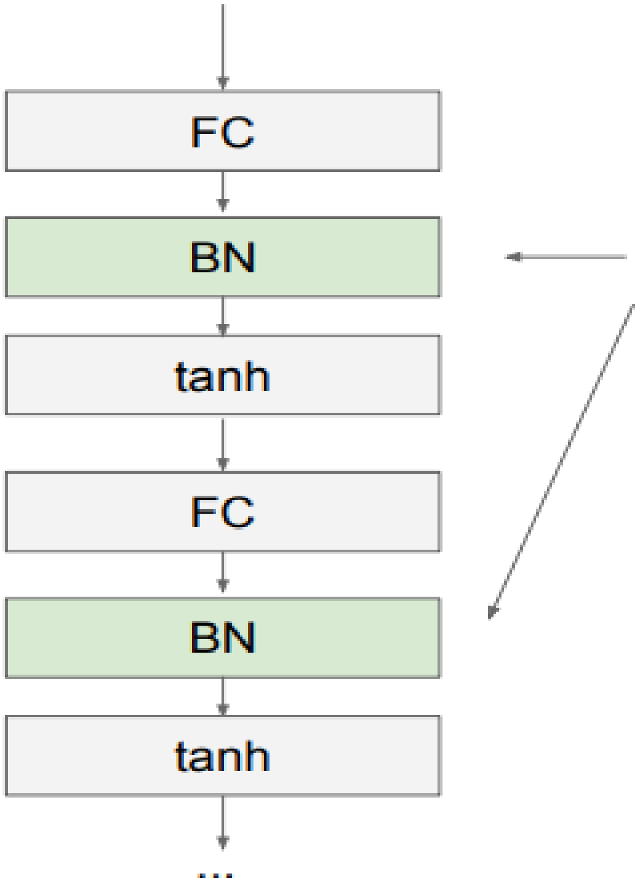


Usually inserted after Fully Connected or Convolutional layers, and before nonlinearity.

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [Ioffe and Szegedy 2015]

Source: cs231n

BATCH NORMALIZATION



Usually inserted after Fully Connected or Convolutional layers, and before nonlinearity.

Problem:
do we necessarily want a
zero-mean unit-variance input?

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [Ioffe and Szegedy 2015]

Source: cs231n

BATCH NORMALIZATION

Normalize:

$$\widehat{x}^{(k)} = \frac{x^{(k)} - \mathrm{E}[x^{(k)}]}{\sqrt{\mathrm{Var}[x^{(k)}]}}$$

And then allow the network to squash
the range if it wants to:

$$y^{(k)} = \gamma^{(k)}\widehat{x}^{(k)} + \beta^{(k)}$$

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [Ioffe and Szegedy 2015]

Source: cs231n

BATCH NORMALIZATION

Normalize:

$$\widehat{x}^{(k)} = \frac{x^{(k)} - \mathrm{E}[x^{(k)}]}{\sqrt{\mathrm{Var}[x^{(k)}]}}$$

And then allow the network to squash
the range if it wants to:

$$y^{(k)} = \gamma^{(k)}\widehat{x}^{(k)} + \beta^{(k)}$$

Note, the network can learn:

$$\gamma^{(k)} = \sqrt{\mathrm{Var}[x^{(k)}]}$$

$$\beta^{(k)} = \mathrm{E}[x^{(k)}]$$

to recover the identity mapping.

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [Ioffe and Szegedy 2015]

Source: cs231n

BATCH NORMALIZATION

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [Ioffe and Szegedy 2015]

Source: cs231n

BATCH NORMALIZATION

Note: at test time BatchNorm layer functions differently:

The mean/std are not computed based on the batch.

Instead, a single fixed empirical mean of activations during training is used.

(e.g. can be estimated during training with running averages)

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [Ioffe and Szegedy 2015]

Source: cs231n

BATCH NORMALIZATION

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \text{BN}_{\gamma,\beta}(x_i)\}$

At test time (usually):

$$\cancel{\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i}$$ // ~~mini-batch~~ mean
training set

$$\cancel{\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2}$$ // ~~mini-batch~~ variance
training set

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}$$ // normalize

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \text{BN}_{\gamma,\beta}(x_i)$$ // scale and shift

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [Ioffe and Szegedy 2015]

Source: cs231n

BATCH NORMALIZATION

Benefits

- Improves gradient flow through the network
- Allows higher learning rates and Accelerates convergence of training
- Reduces the strong dependence on initialization
- Acts as a form of regularization

Pitfalls

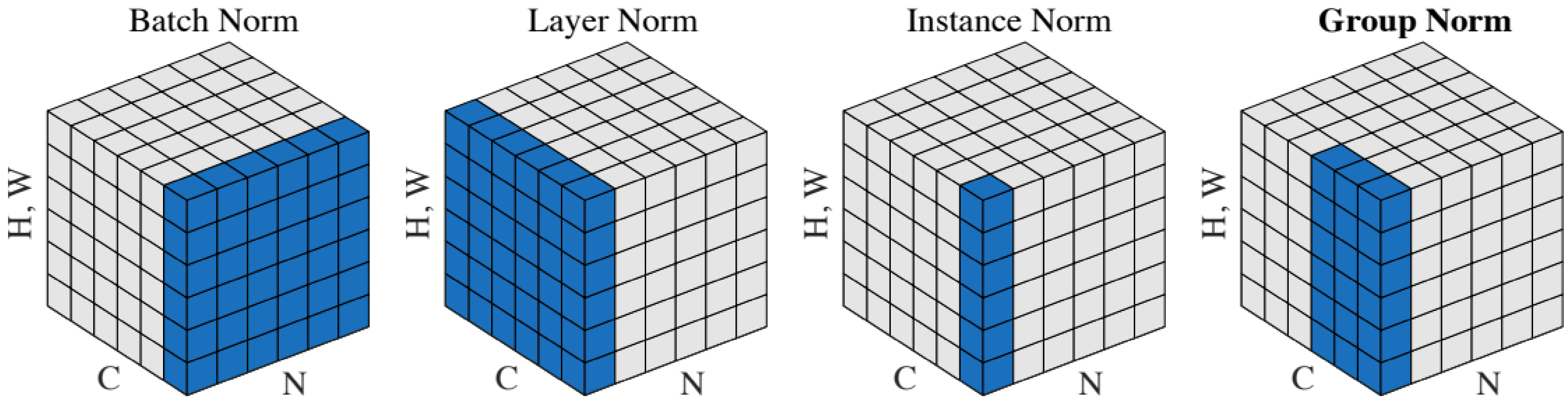
- Behavior depends on composition of mini-batches, can lead to hard-to-catch bugs if there is a mismatch between training and test regime (example)
- Doesn't work well for small mini-batch sizes
- Cannot be used in recurrent models

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [Ioffe and Szegedy 2015]

Source: cs231n

OTHER TYPES OF NORMALIZATION

- Layer normalization (Ba et al., 2016)
- Instance normalization (Ulyanov et al., 2017)
- Group normalization (Wu and He, 2018)
- Weight normalization (Salimans et al., 2016)



Y. Wu and K. He, Group Normalization, ECCV 2018

BATCH NORMALIZATION: RECENT TRENDS

Layer Normalization:

Ba, Kiros, and Hinton, "Layer Normalization", arXiv 2016

Instance Normalization:

Ulyanov et al, Improved Texture Networks: Maximizing Quality and Diversity in Feed-forward Stylization and Texture Synthesis, CVPR 2017

Group Normalization:

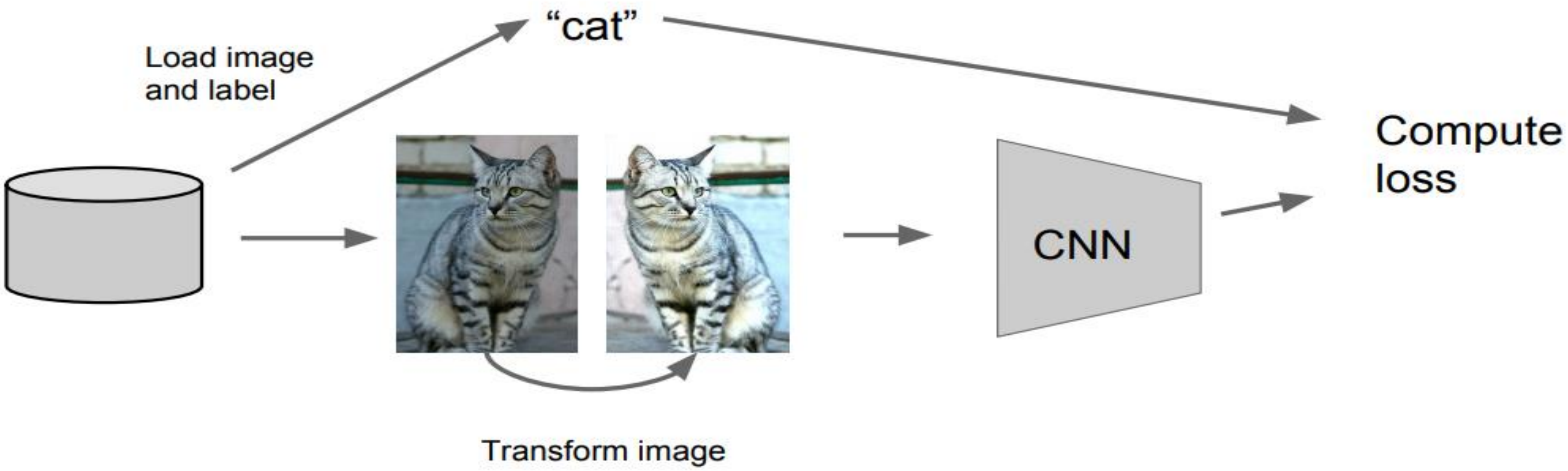
Wu and He, "Group Normalization", arXiv 2018
(Appeared 3/22/2018)

Decorrelated Normalization:

Huang et al, "Decorrelated Batch Normalization", arXiv 2018
(Appeared 4/23/2018)

Data Augmentation

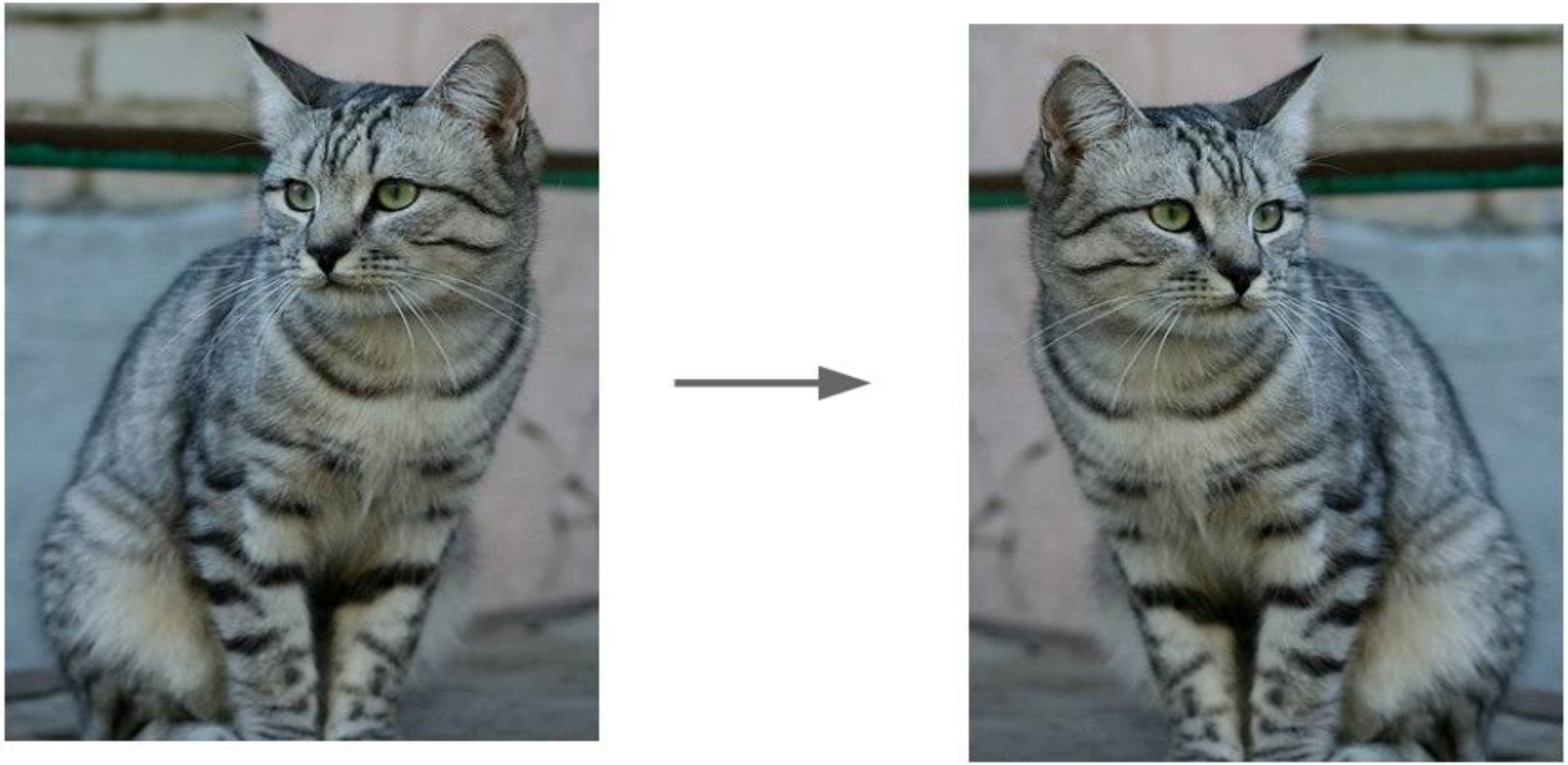
DATA AUGMENTATION (JITTERING)



Source: cs231n

DATA AUGMENTATION (JITTERING)

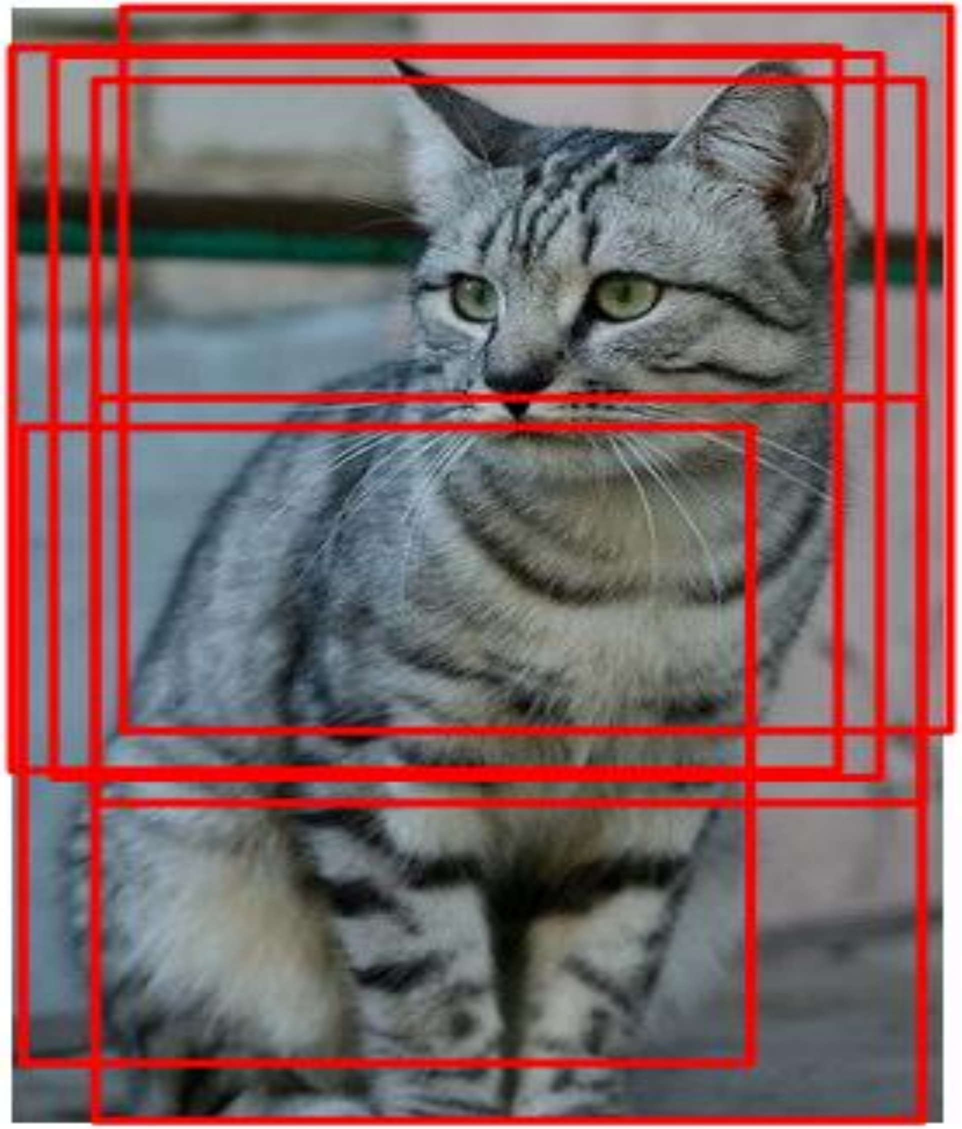
Horizontal Flips



Source: cs231n

DATA AUGMENTATION (JITTERING)

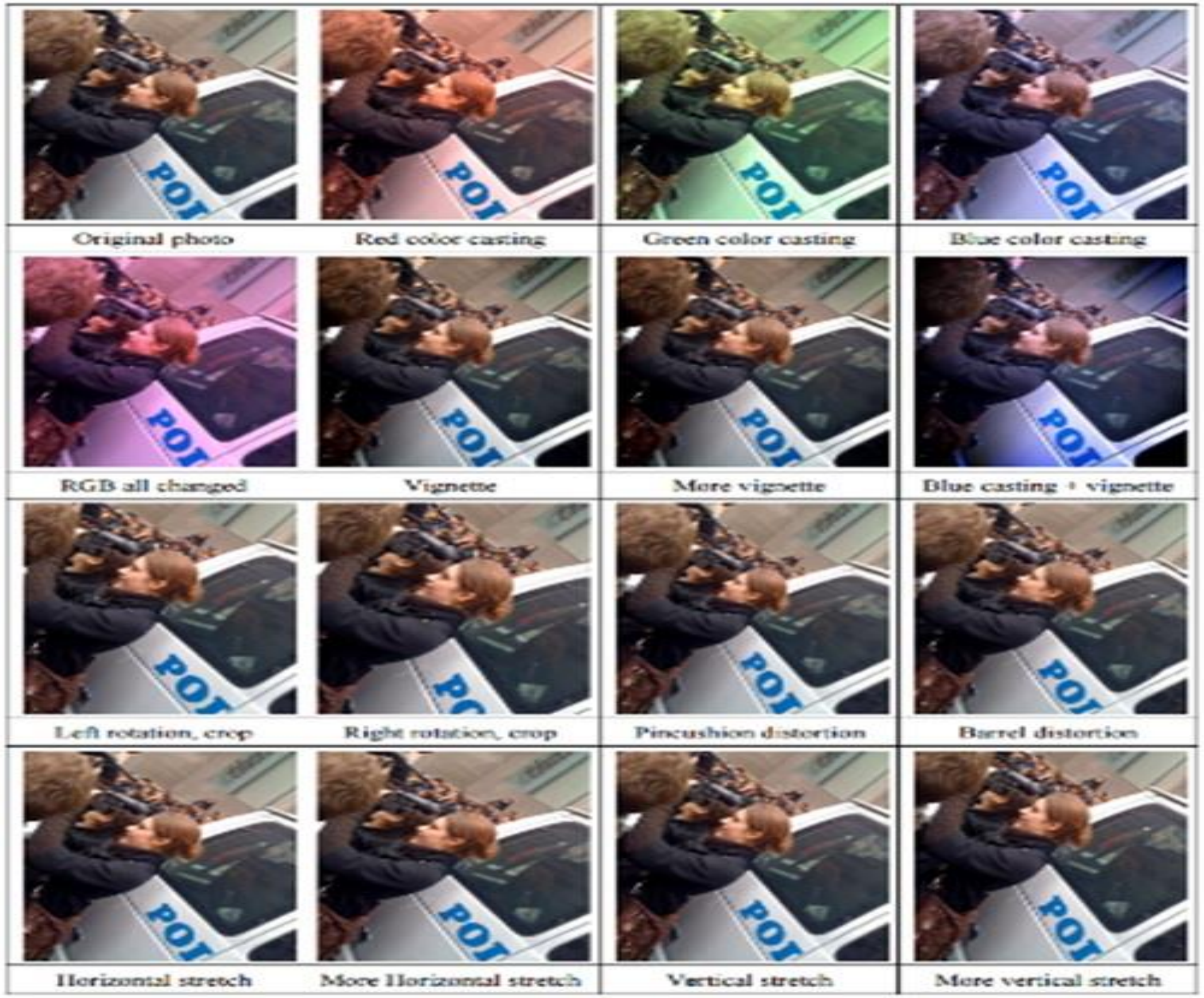
Random crops and scales



Source: cs231n

DATA AUGMENTATION (JITTERING)

- Create *virtual* training samples
- Get creative for your problem!
  - Horizontal flip
  - Random crop
  - Color casting
  - Randomize contrast
  - Randomize brightness
  - Geometric distortion
  - Rotation
  - Photometric changes

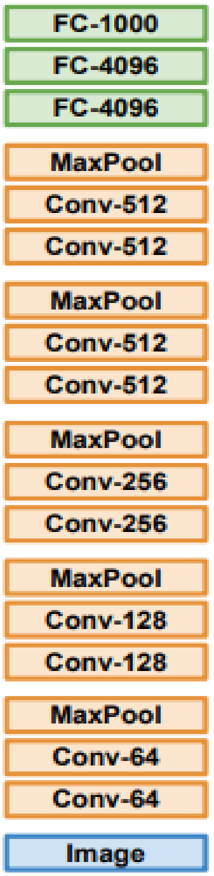


Deep Image [Wu et al. 2015]

Transfer Learning

TRANSFER LEARNING WITH CNNS

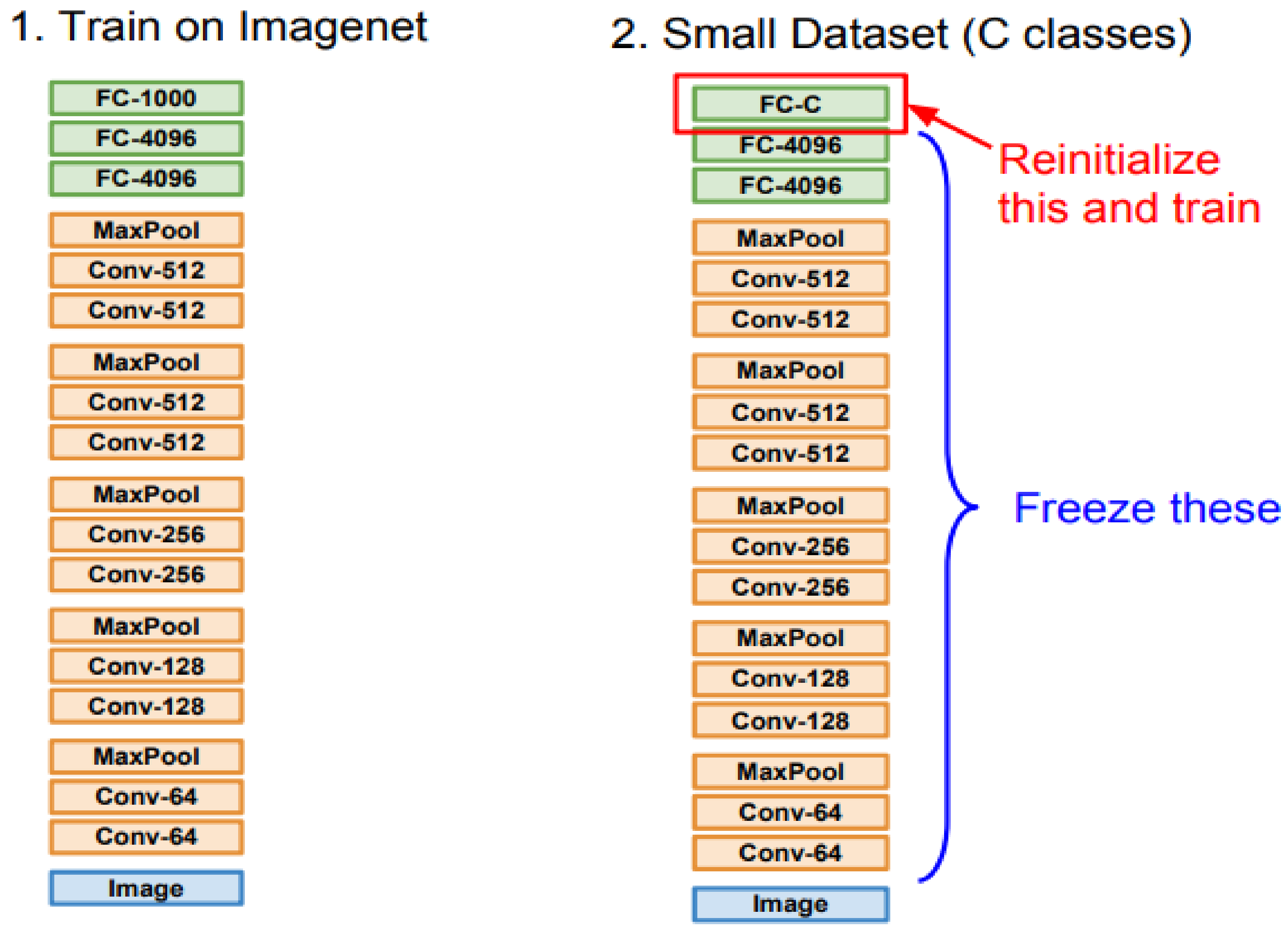
1. Train on Imagenet



Donahue et al, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition", ICML 2014
Razavian et al, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition", CVPR Workshops 2014

Source: cs231n

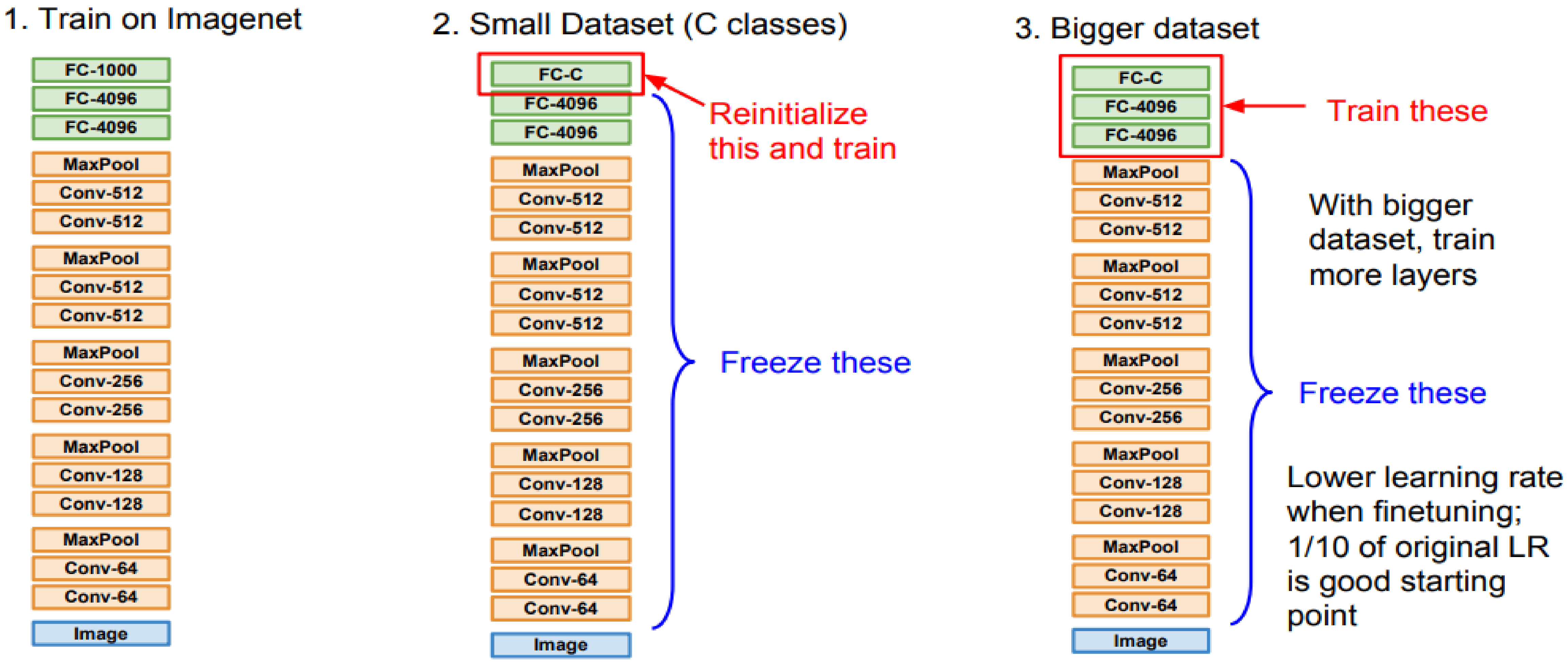
TRANSFER LEARNING WITH CNNS



Donahue et al, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition", ICML 2014
Razavian et al, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition", CVPR Workshops 2014

Source: cs231n

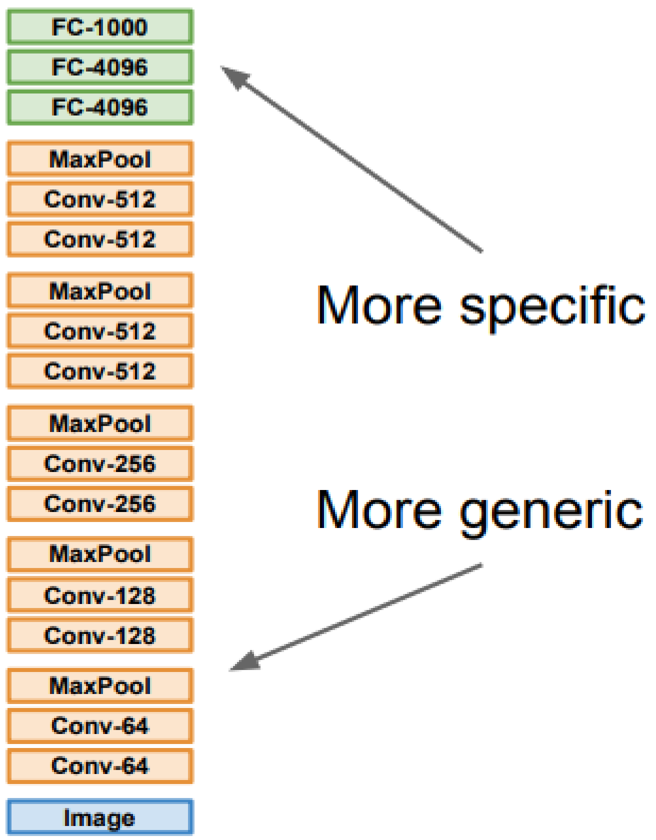
TRANSFER LEARNING WITH CNNS



Donahue et al, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition", ICML 2014
Razavian et al, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition", CVPR Workshops 2014

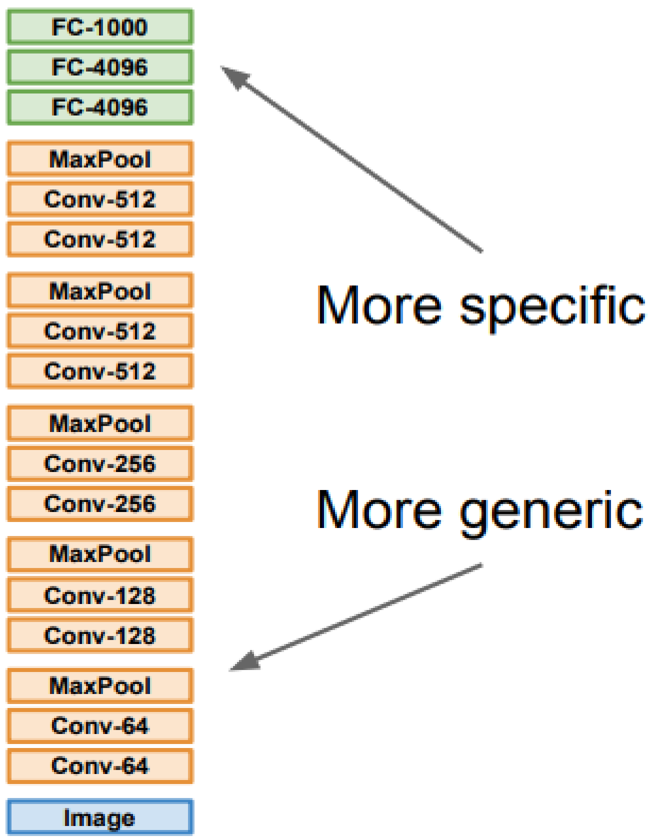
Source: cs231n

TRANSFER LEARNING WITH CNNS



Source: cs231n

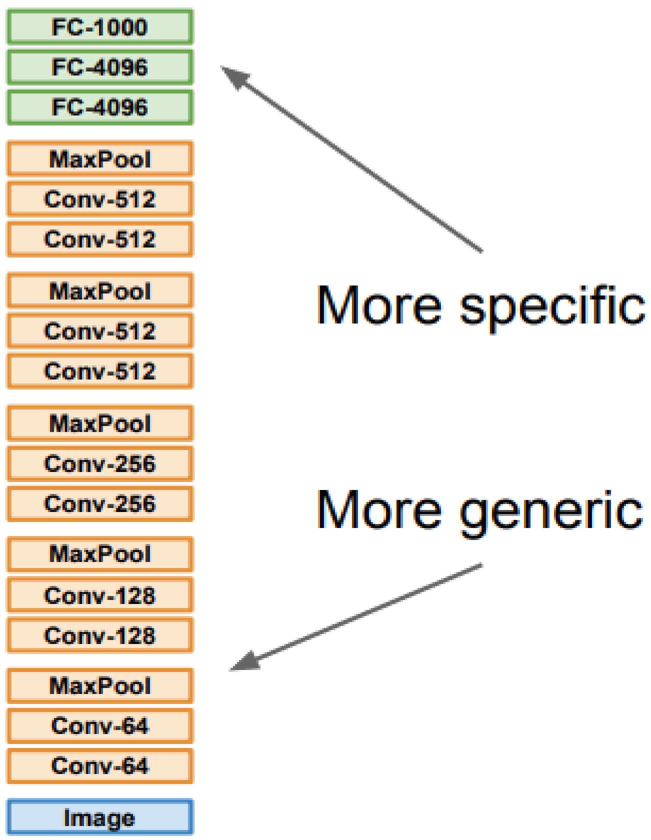
TRANSFER LEARNING WITH CNNS



	very similar dataset	very different dataset
very little data		
quite a lot of data		

Source: cs231n

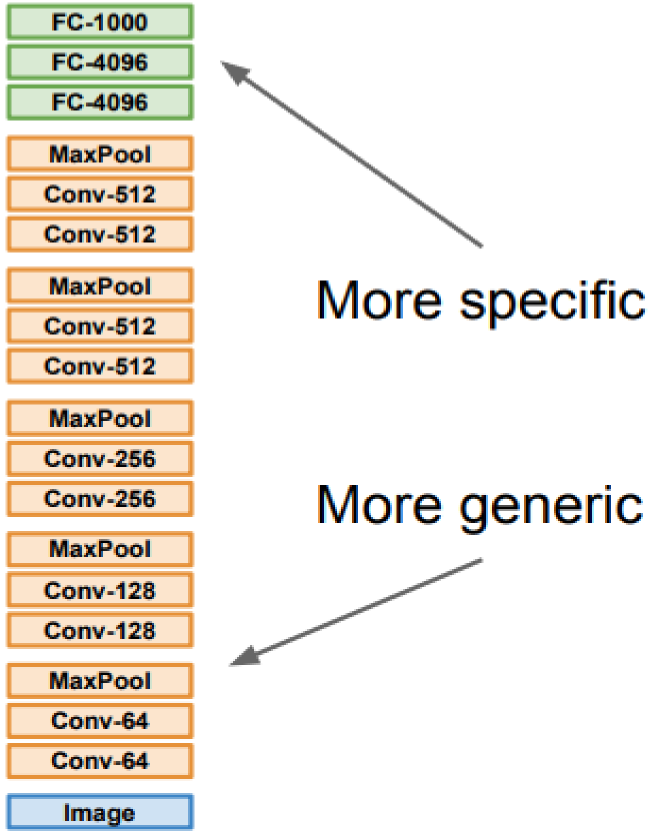
TRANSFER LEARNING WITH CNNS



	very similar dataset	**very different dataset**
very little data	Use Linear Classifier on top layer	
quite a lot of data		

Source: cs231n

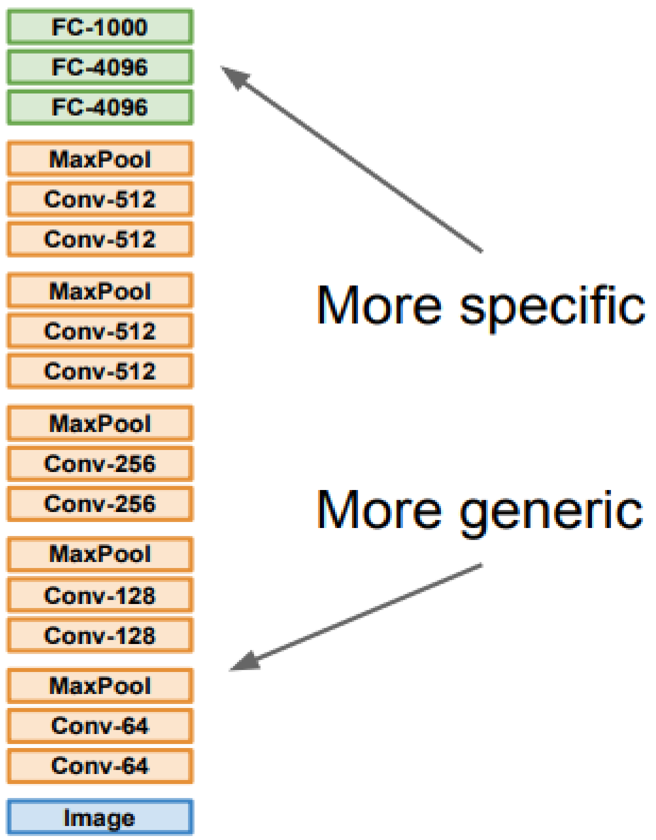
TRANSFER LEARNING WITH CNNS



	very similar dataset	very different dataset
very little data	Use Linear Classifier on top layer	
quite a lot of data	Finetune a few layers	

Source: cs231n

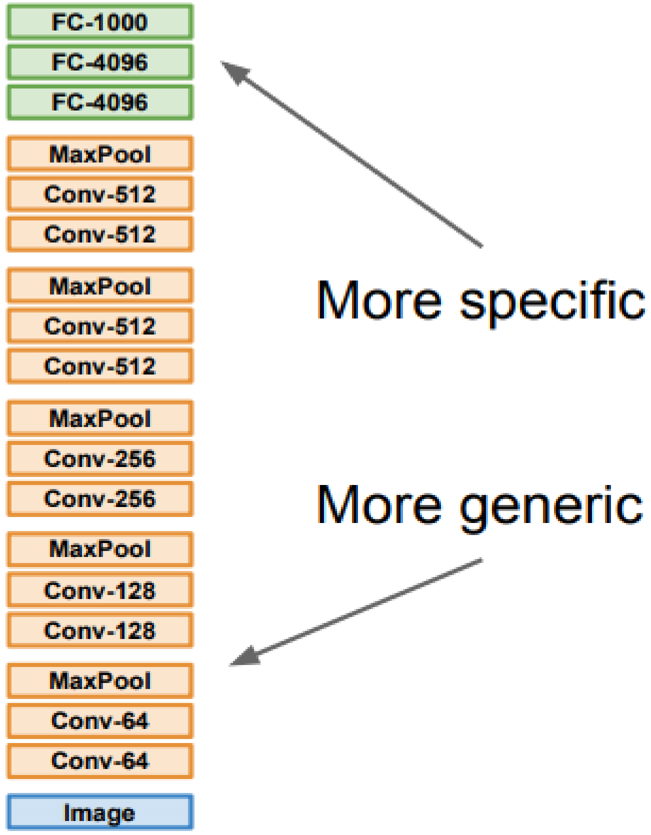
TRANSFER LEARNING WITH CNNS



	very similar dataset	very different dataset
very little data	Use Linear Classifier on top layer	
quite a lot of data	Finetune a few layers	Finetune a larger number of layers

Source: cs231n

TRANSFER LEARNING WITH CNNS



	very similar dataset	**very different dataset**
very little data	Use Linear Classifier on top layer	You're in trouble… Try linear classifier from different stages
quite a lot of data	Finetune a few layers	Finetune a larger number of layers

Source: cs231n

TRANSFER LEARNING WITH CNNS

Takeaway for your projects and beyond:

Have some dataset of interest but it has < ~1M images?

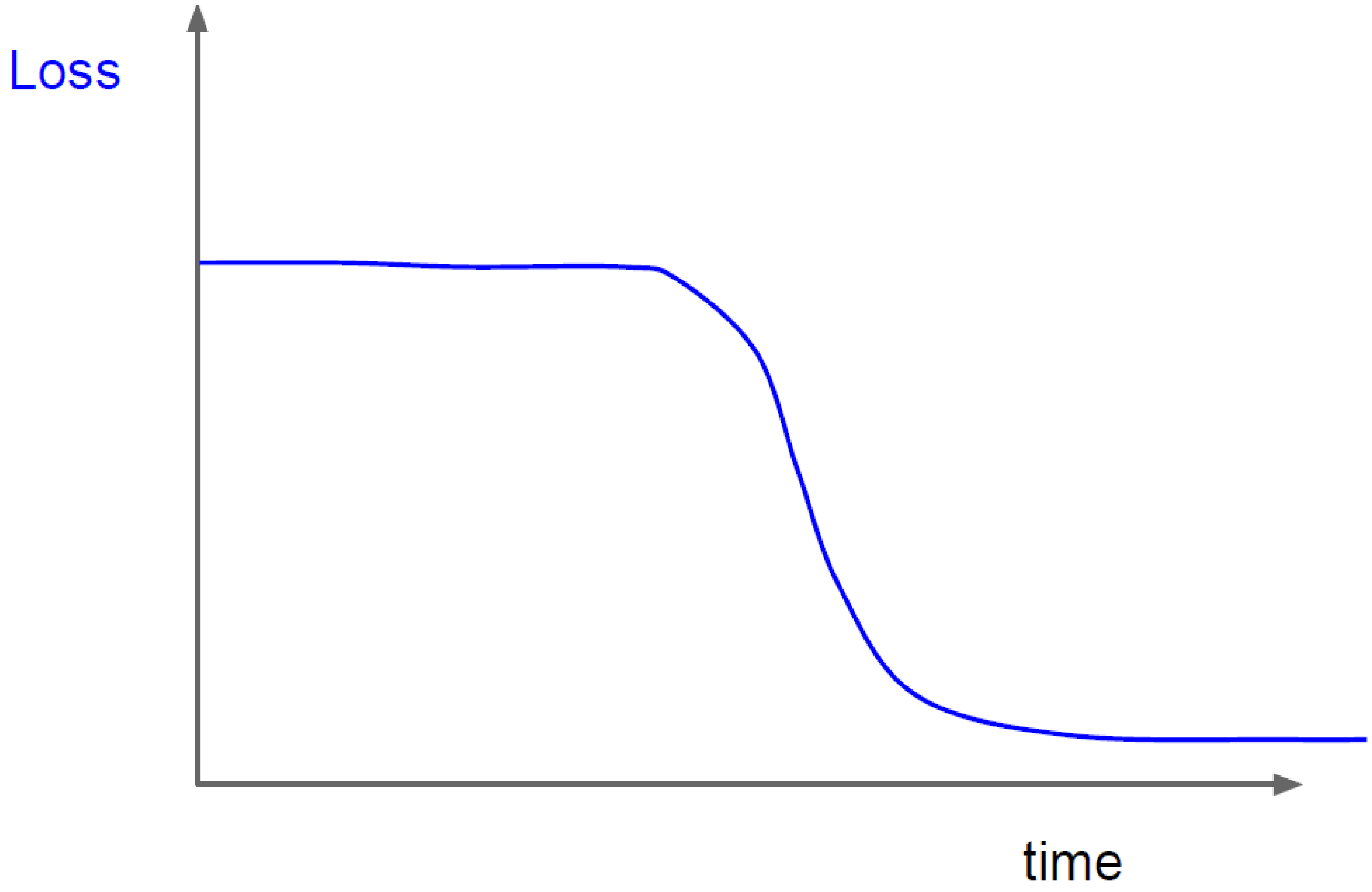
1. Find a very large dataset that has similar data, train a big ConvNet there
2. Transfer learn to your dataset

Deep learning frameworks provide a "Model Zoo" of pretrained models so you don't need to train your own

Caffe: https://github.com/BVLC/caffe/wiki/Model-Zoo
TensorFlow: https://github.com/tensorflow/models
PyTorch: https://github.com/pytorch/vision
Matconvnet: http://www.vlfeat.org/matconvnet/pretrained/

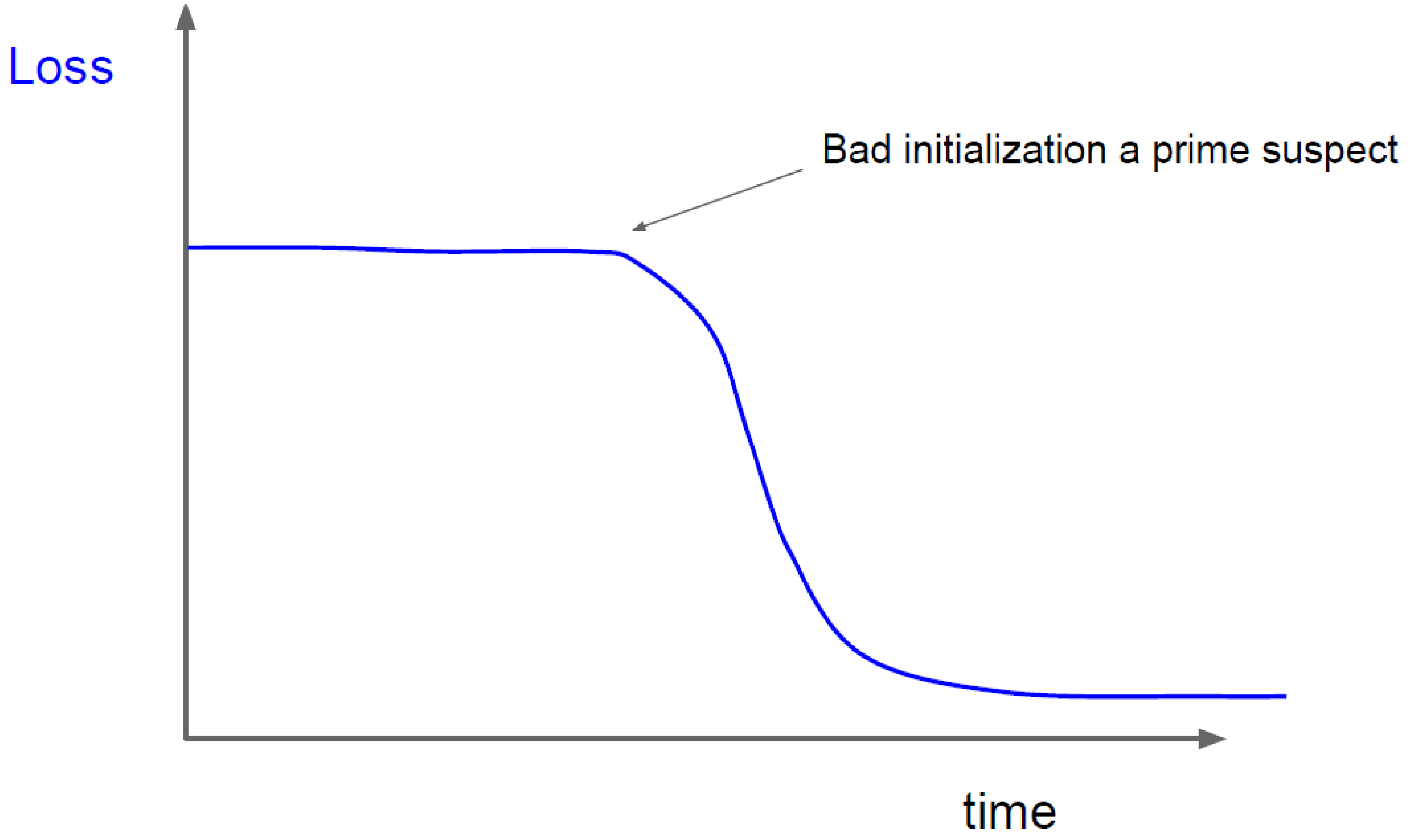
Source: cs231n

MONITOR AND VISUALIZE THE LOSS CURVE



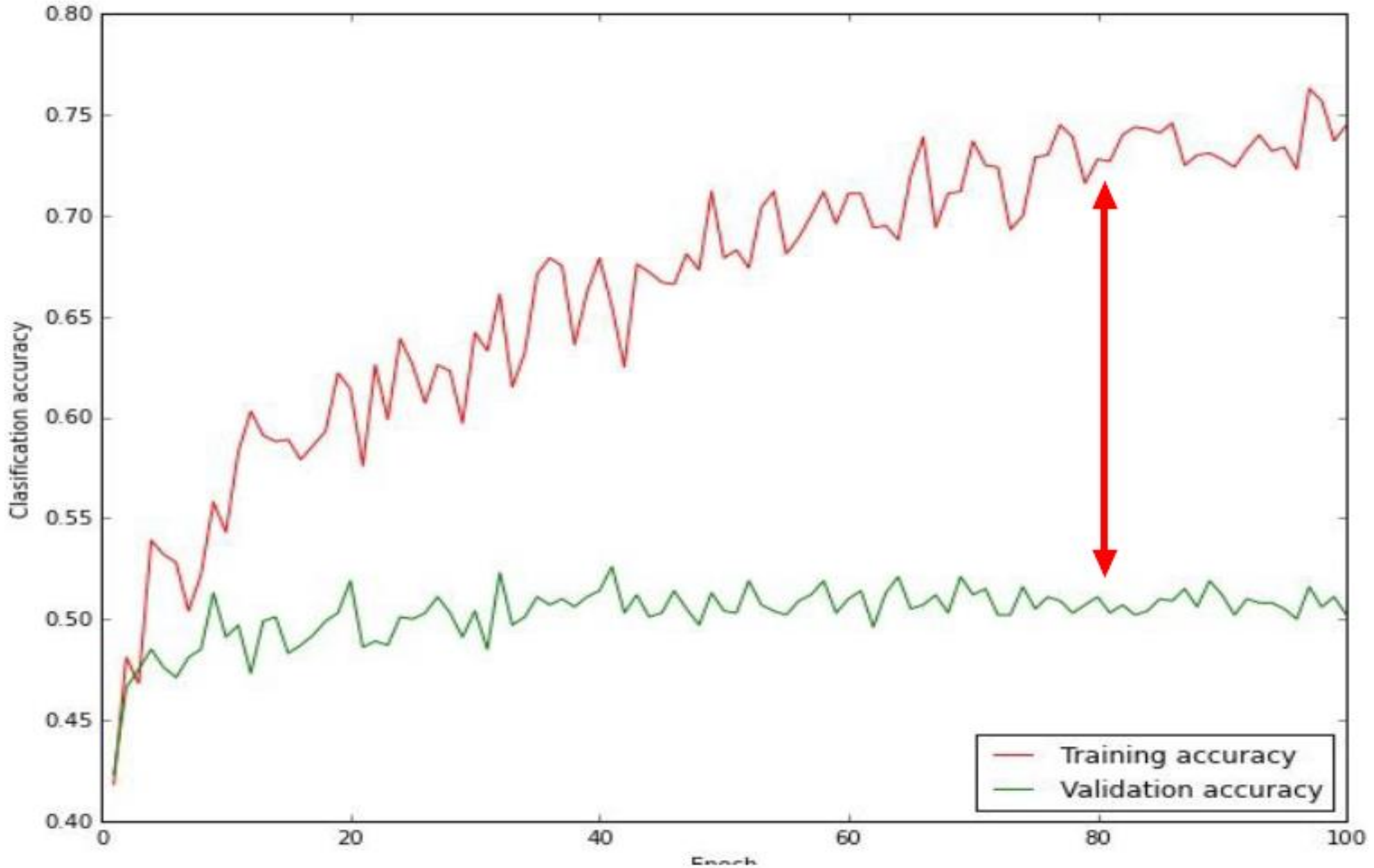
Source: cs231n

MONITOR AND VISUALIZE THE LOSS CURVE



Source: cs231n

MONITOR AND VISUALIZE THE LOSS CURVE

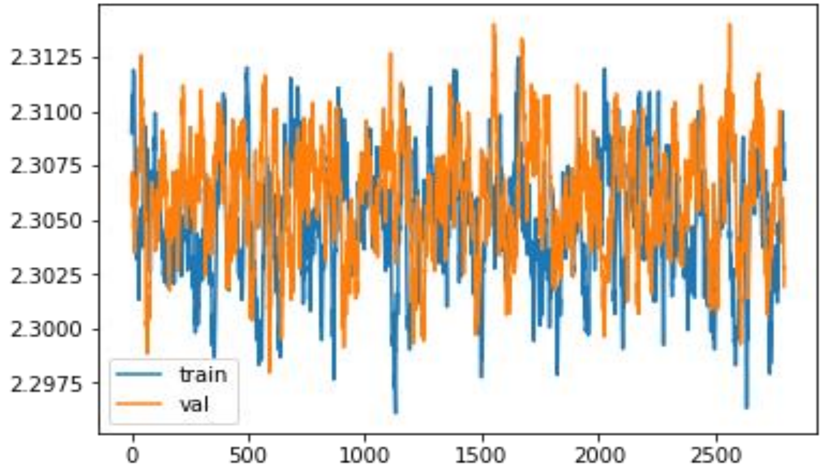


big gap = overfitting
=> increase
regularization strength?

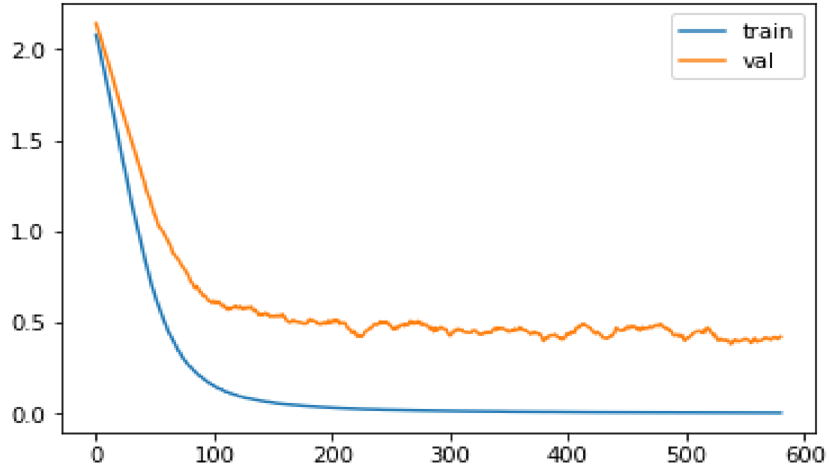
no gap
=> increase model
capacity?

Source: cs231n

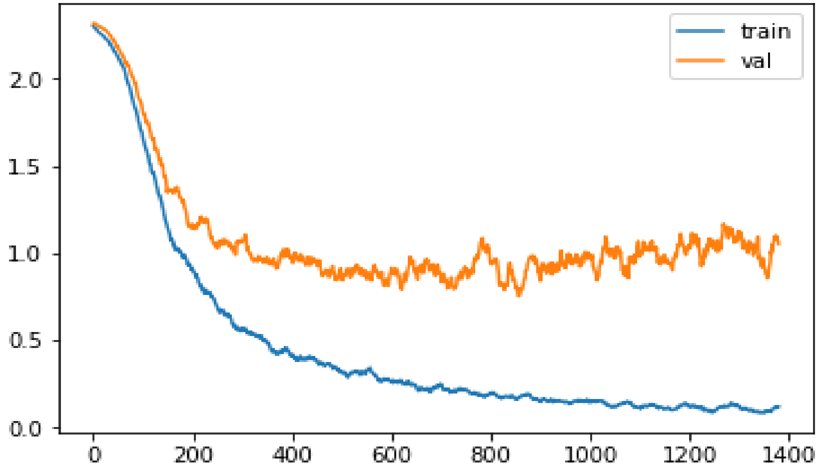
MONITOR AND VISUALIZE THE LOSS CURVE



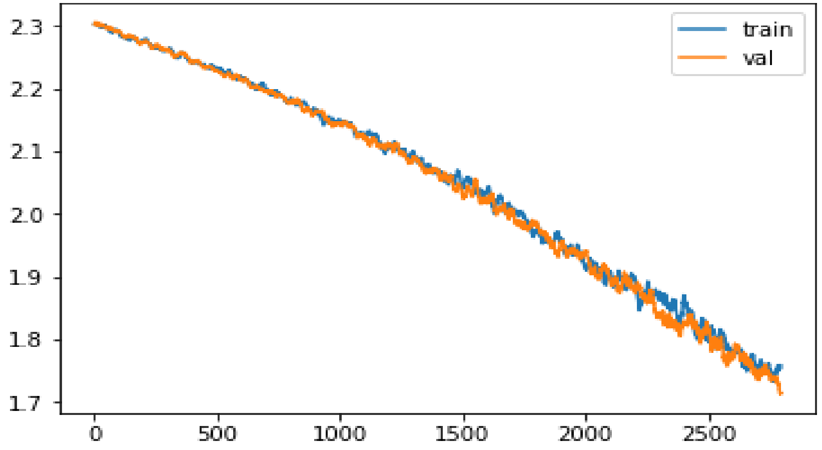
Not learning: gradients not applied to weights



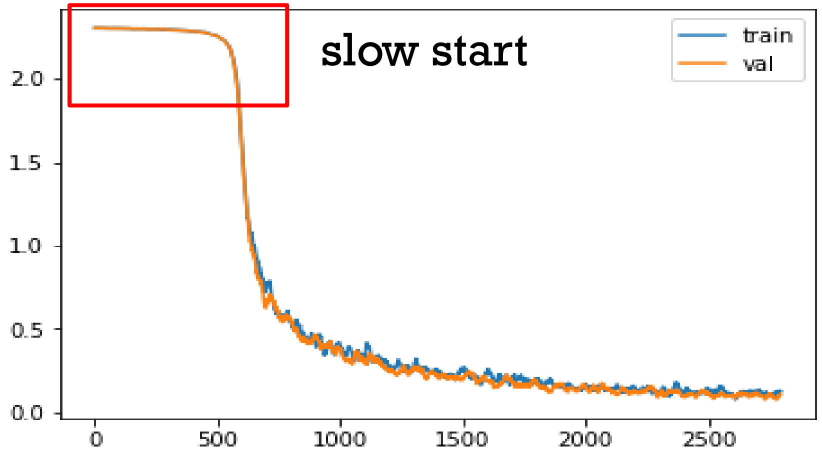
Overfit: model too large/dataset too small



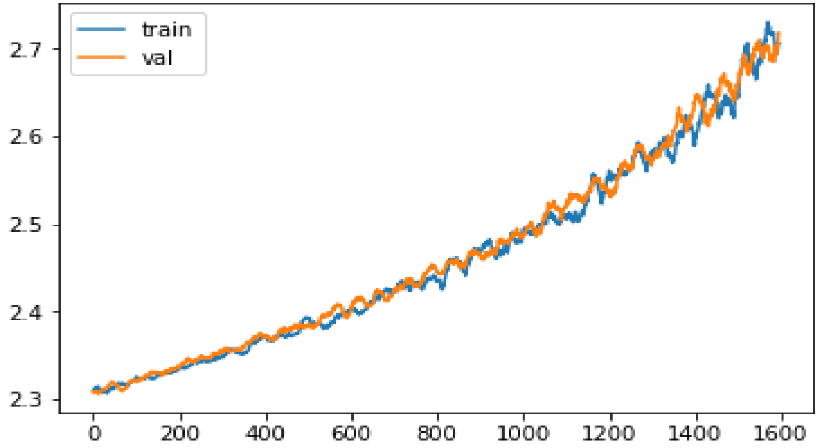
More extreme case of overfitting



Not converged yet: need longer training



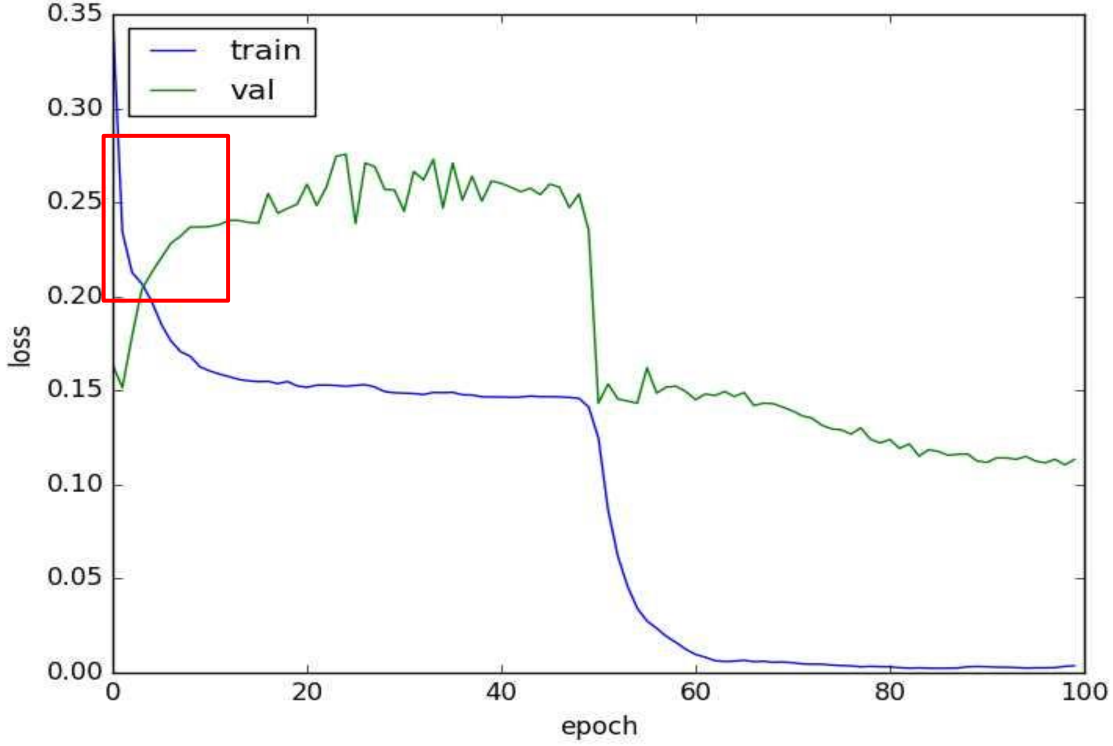
Slow start: initialization weights too small



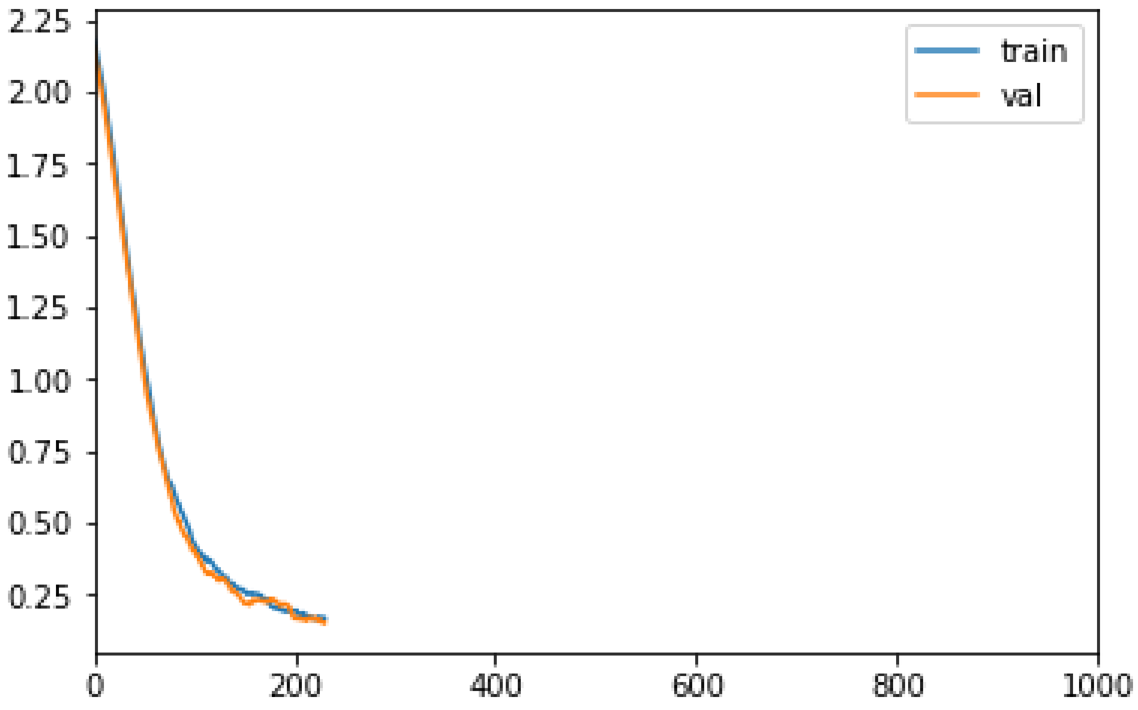
Applied the negative of gradients

Source: cs231n

MONITOR AND VISUALIZE THE LOSS CURVE



Problem: val set too small, statistics not meaningful

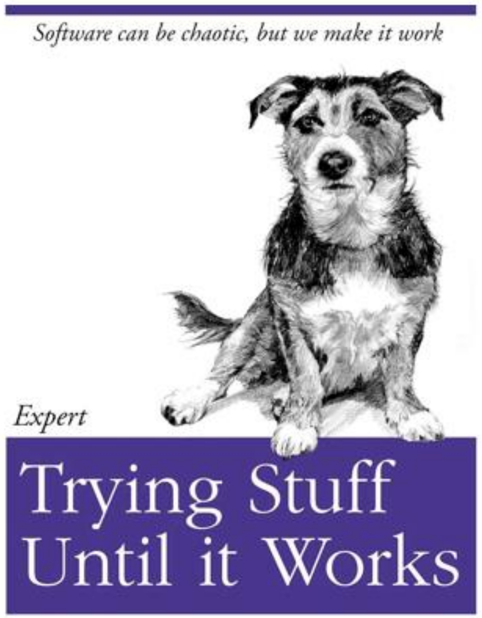


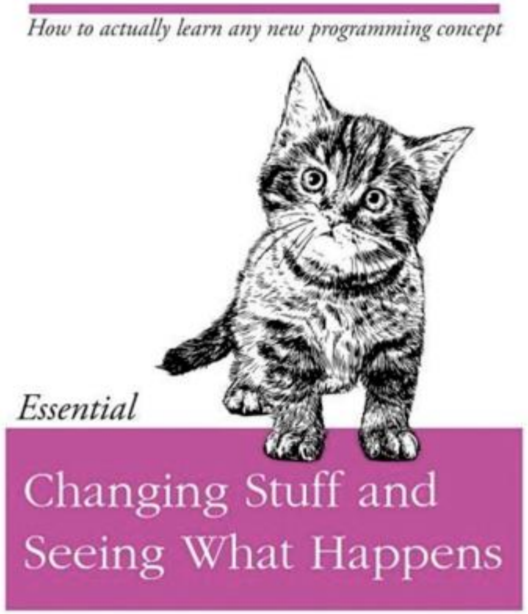
Get nans in the loss after a number of iterations: caused by high learning rate and numerical instability in models

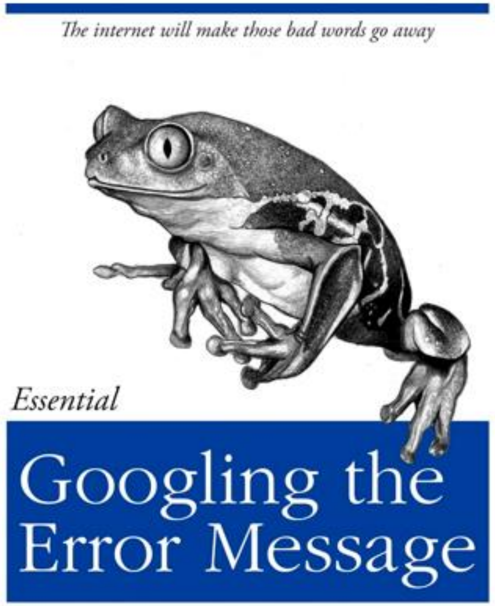
Source: cs231n

ATTEMPT AT A CONCLUSION

- Training neural networks is still a black art
- Process requires close “babysitting”
- For many techniques, the reasons why, when, and whether they work are in active dispute – read everything but don’t trust anything
- It all comes down to (principled) trial and error
- Further reading: A. Karpathy, A recipe for training neural networks







ACKNOWLEDGEMENT

Thanks to the following courses and corresponding researchers for making their teaching/research material online

- Deep Learning, Stanford University
- Introduction to Deep Learning, University of Illinois at Urbana-Champaign
- Introduction to Deep Learning, Carnegie Mellon University
- Convolutional Neural Networks for Visual Recognition, Stanford University
- Natural Language Processing with Deep Learning, Stanford University
- And Many More ……

Thank You