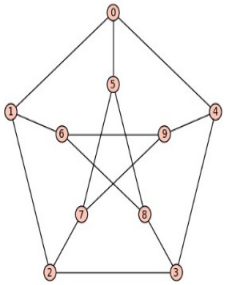
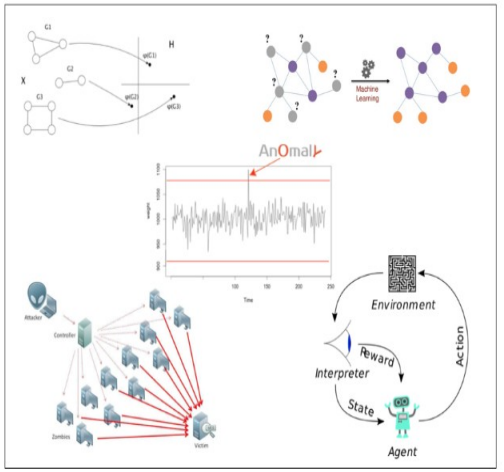
Training Stability

Pierre Pereira
Université Côte d'Azur / Inria Coati
pierre.pereira@inria.fr





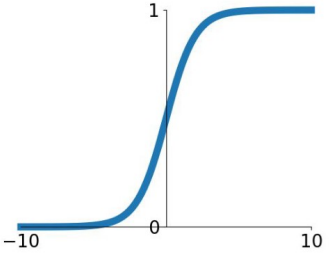
$$
\begin{aligned}
\min \quad & \sum_{e \in \mathcal{E}} y_e \\
\text{s.t.} \quad & \sum_{a \in A_i^+(u)} f_a^i - \sum_{a \in A_i^-(u)} f_a^i = \begin{cases} |V_i| - 1 & \text{if } u = s_i \\ -1 & \text{if } u \neq s_i \end{cases} && \forall u \in V_i,\ V_i \in C \\
& f_a^i \leq |V_i| \cdot x_a, && \forall V_i \in C, a \in A \\
& x_{(u,v)} \leq y_{uv}, && \forall uv \in \mathcal{E} \\
& x_{(v,u)} \leq y_{uv}, && \forall uv \in \mathcal{E}
\end{aligned}
$$

Sources multiples, principalement
Roger Grosse, Toronto

Last time: Activation Functions

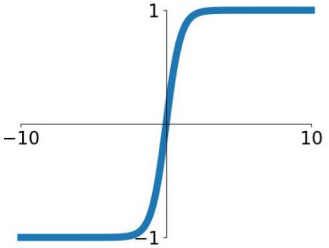
Sigmoid

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



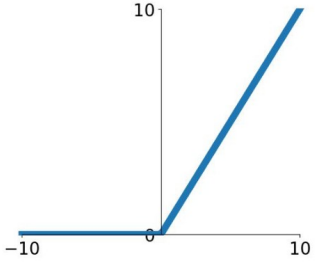
tanh

$$\tanh(x)$$



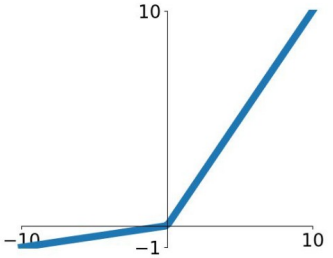
ReLU

$$\max(0, x)$$



Leaky ReLU

$$\max(0.1x, x)$$

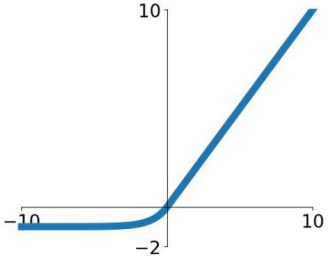


Maxout

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

ELU

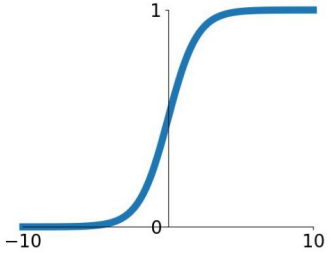
$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



Last time: Activation Functions

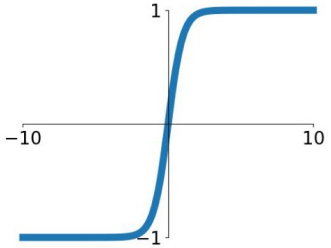
Sigmoid

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



tanh

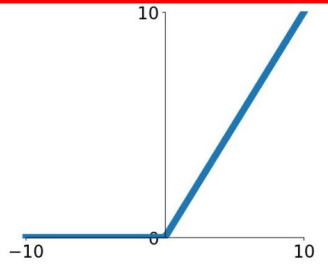
$$\tanh(x)$$



ReLU

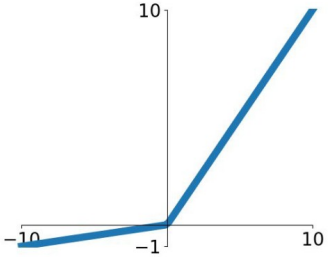
$$\max(0, x)$$

Good default choice



Leaky ReLU

$$\max(0.1x, x)$$

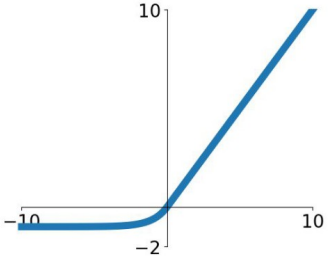


Maxout

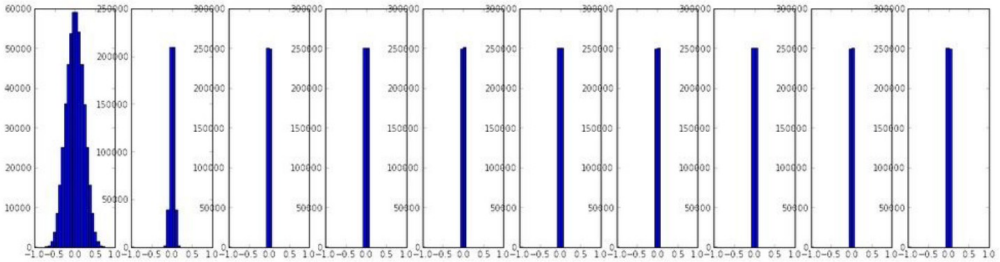
$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

ELU

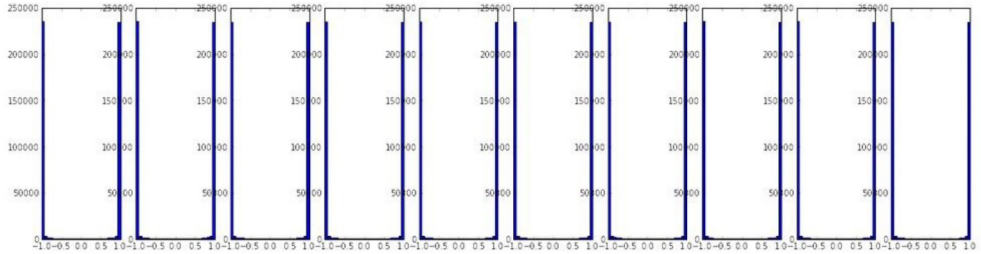
$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



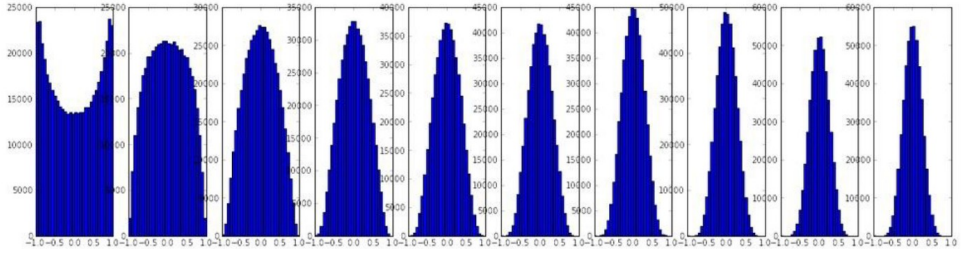
Last time: Weight Initialization



Initialization too small:
Activations go to zero, gradients also zero,
No learning

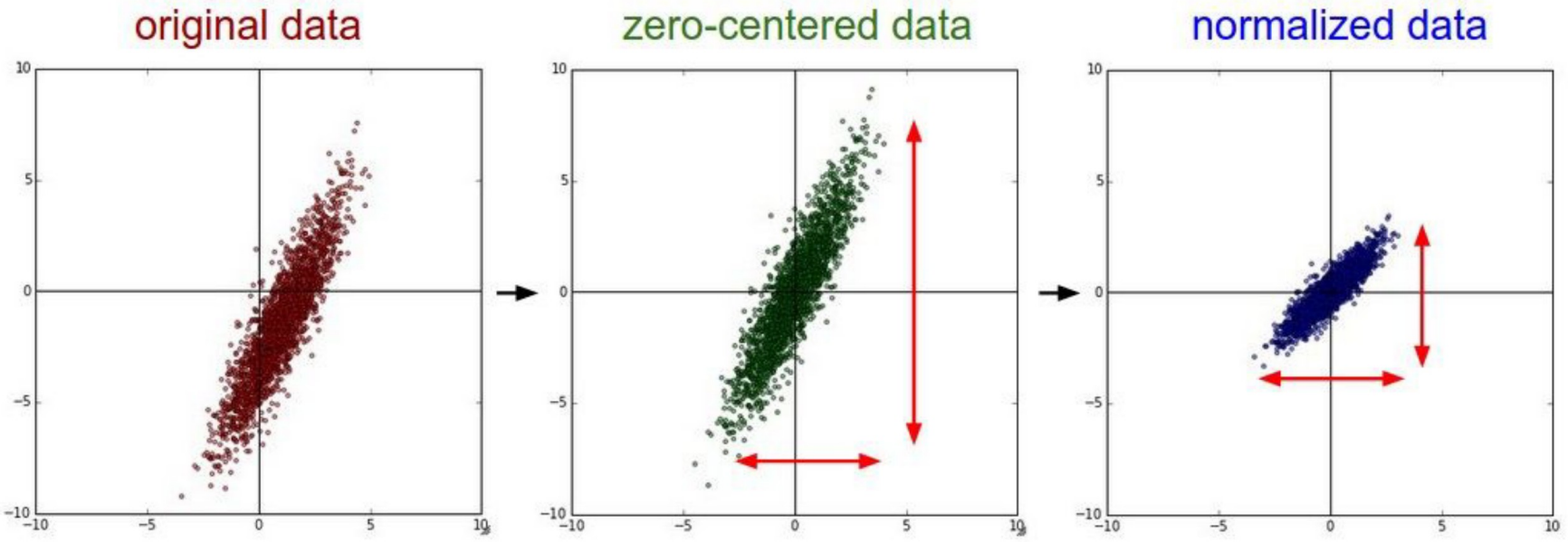


Initialization too big:
Activations saturate (for tanh),
Gradients zero, no learning



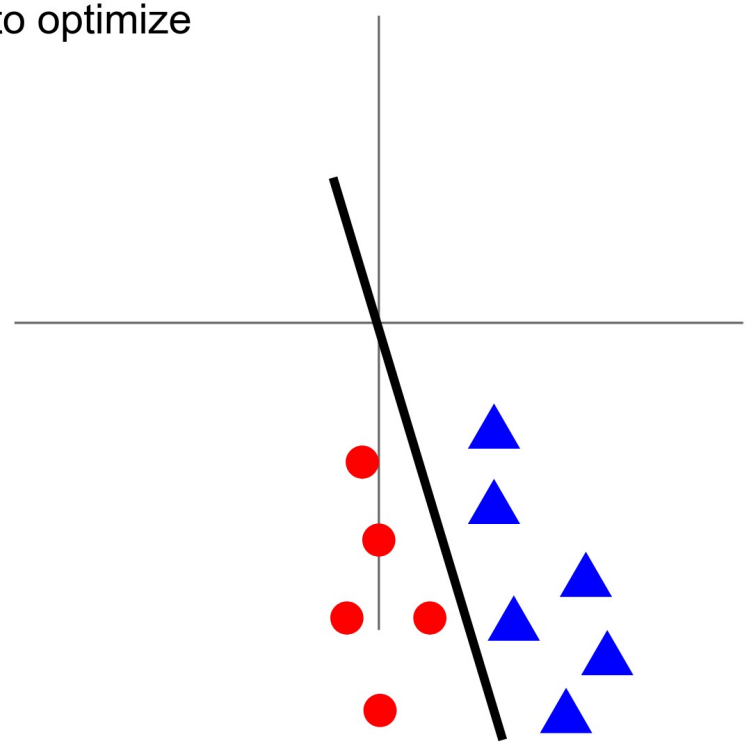
Initialization just right:
Nice distribution of activations at all layers,
Learning proceeds nicely

Last time: Data Preprocessing

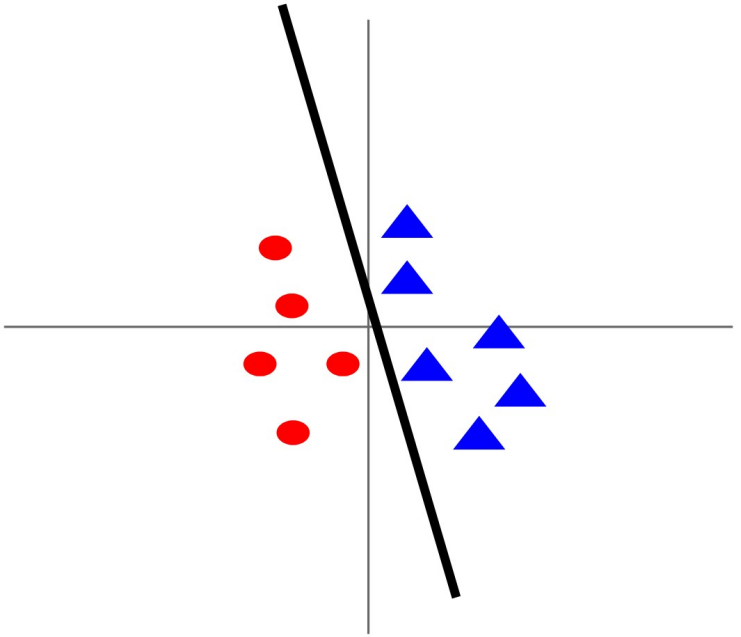


Last time: Data Preprocessing

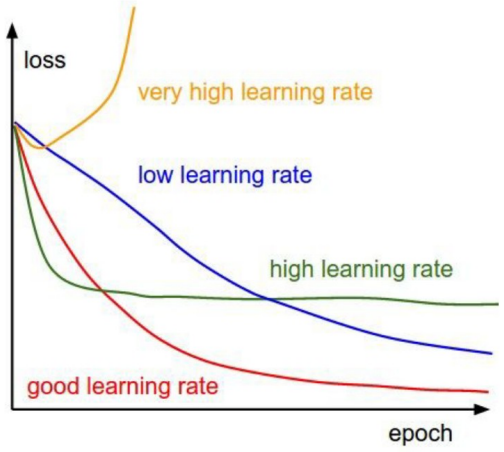
Before normalization: classification loss very sensitive to changes in weight matrix; hard to optimize

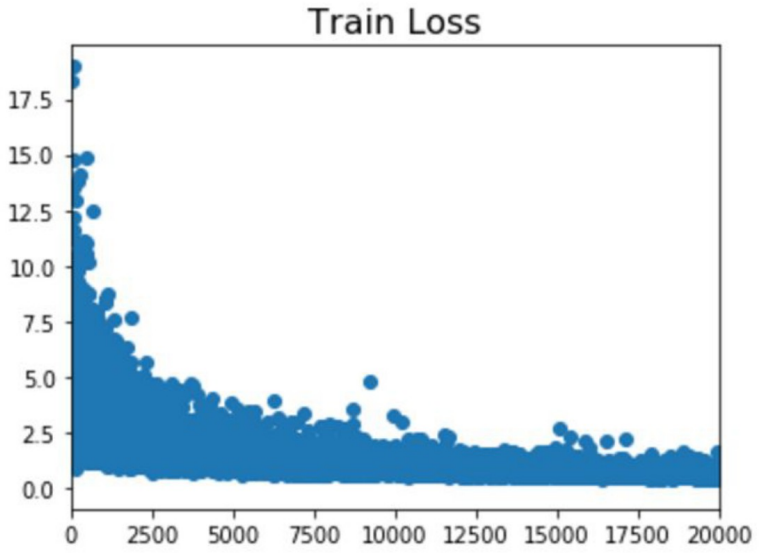


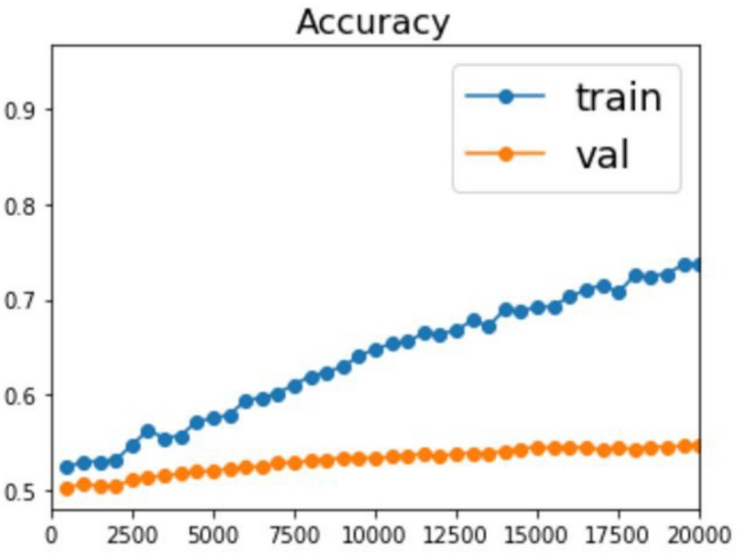
After normalization: less sensitive to small changes in weights; easier to optimize



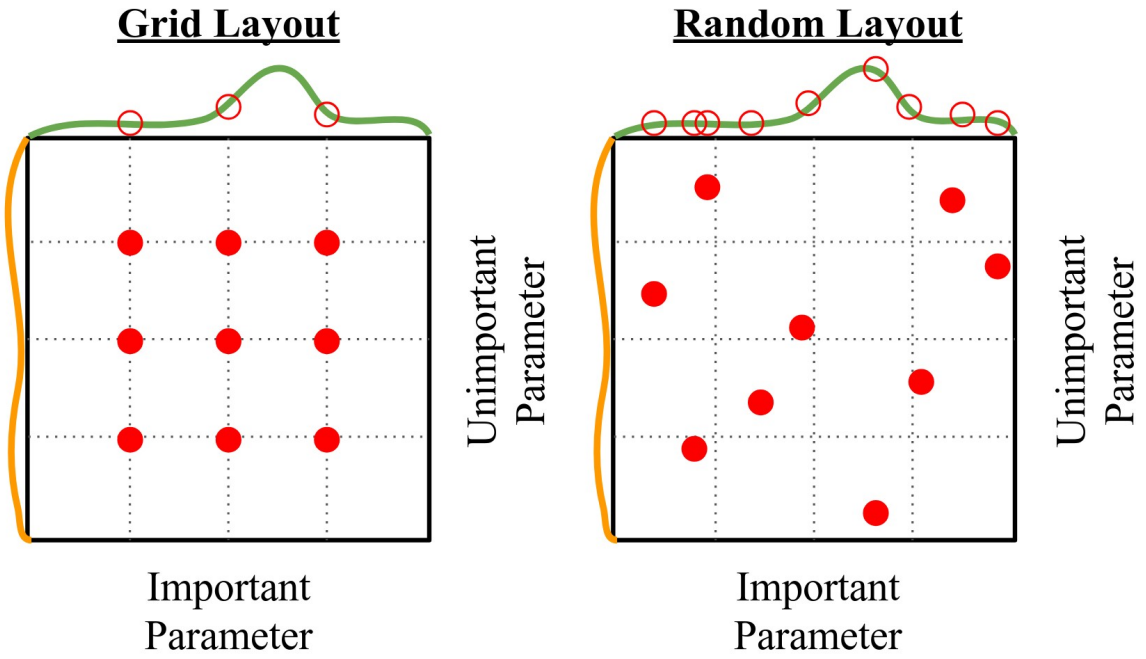
Last time: Babysitting Learning







Last time: Hyperparameter Search



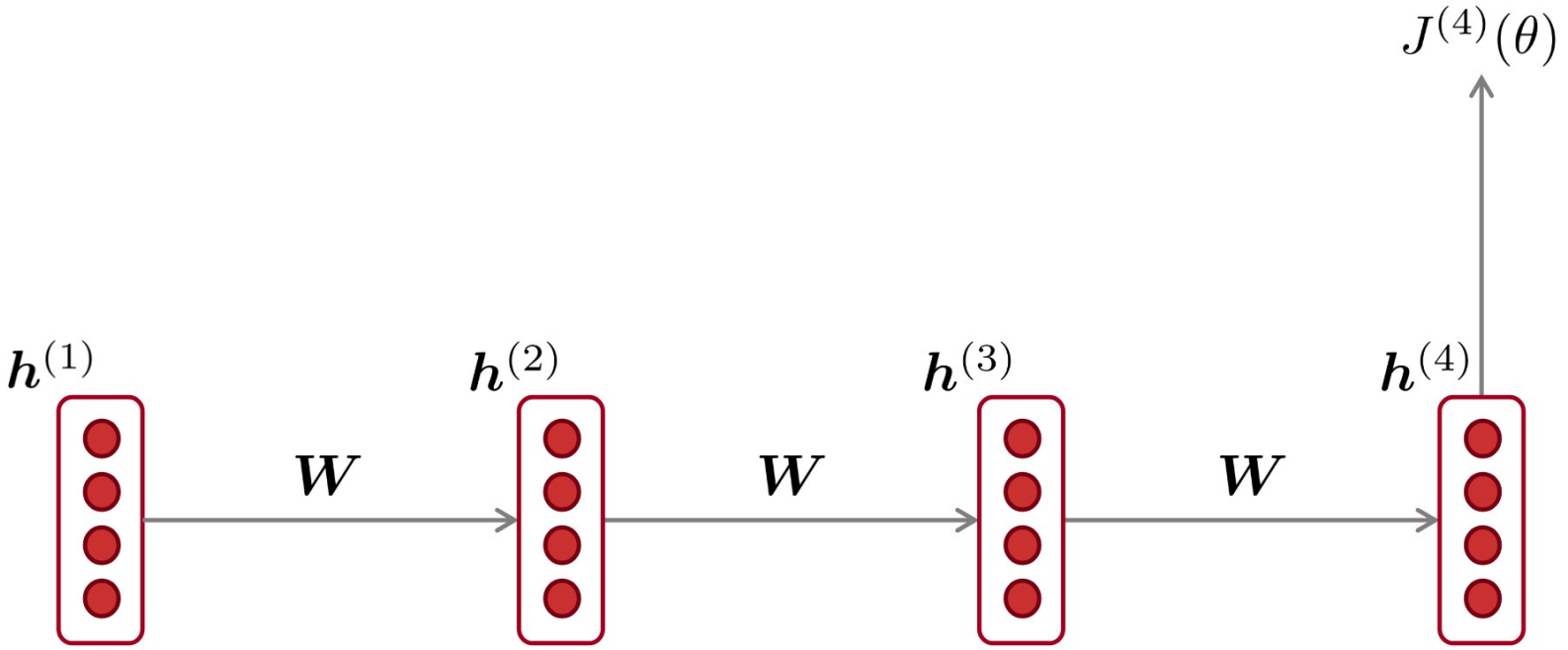
Coarse to fine search

```
val_acc: 0.412000, lr: 1.405206e-04, reg: 4.793564e-01, (1 / 100)
val_acc: 0.214000, lr: 7.231888e-06, reg: 2.321281e-04, (2 / 100)
val_acc: 0.208000, lr: 2.119571e-06, reg: 8.011857e+01, (3 / 100)
val_acc: 0.196000, lr: 1.551131e-05, reg: 4.374936e-05, (4 / 100)
val_acc: 0.079000, lr: 1.753300e-05, reg: 1.200424e+03, (5 / 100)
val_acc: 0.223000, lr: 4.215128e-05, reg: 4.196174e+01, (6 / 100)
val_acc: 0.441000, lr: 1.750259e-04, reg: 2.110807e-04, (7 / 100)
val_acc: 0.241000, lr: 6.749231e-05, reg: 4.226413e+01, (8 / 100)
val_acc: 0.482000, lr: 4.296863e-04, reg: 6.642555e-01, (9 / 100)
val_acc: 0.079000, lr: 5.401602e-06, reg: 1.599828e+04, (10 / 100)
val_acc: 0.154000, lr: 1.618508e-06, reg: 4.925252e-01, (11 / 100)
```

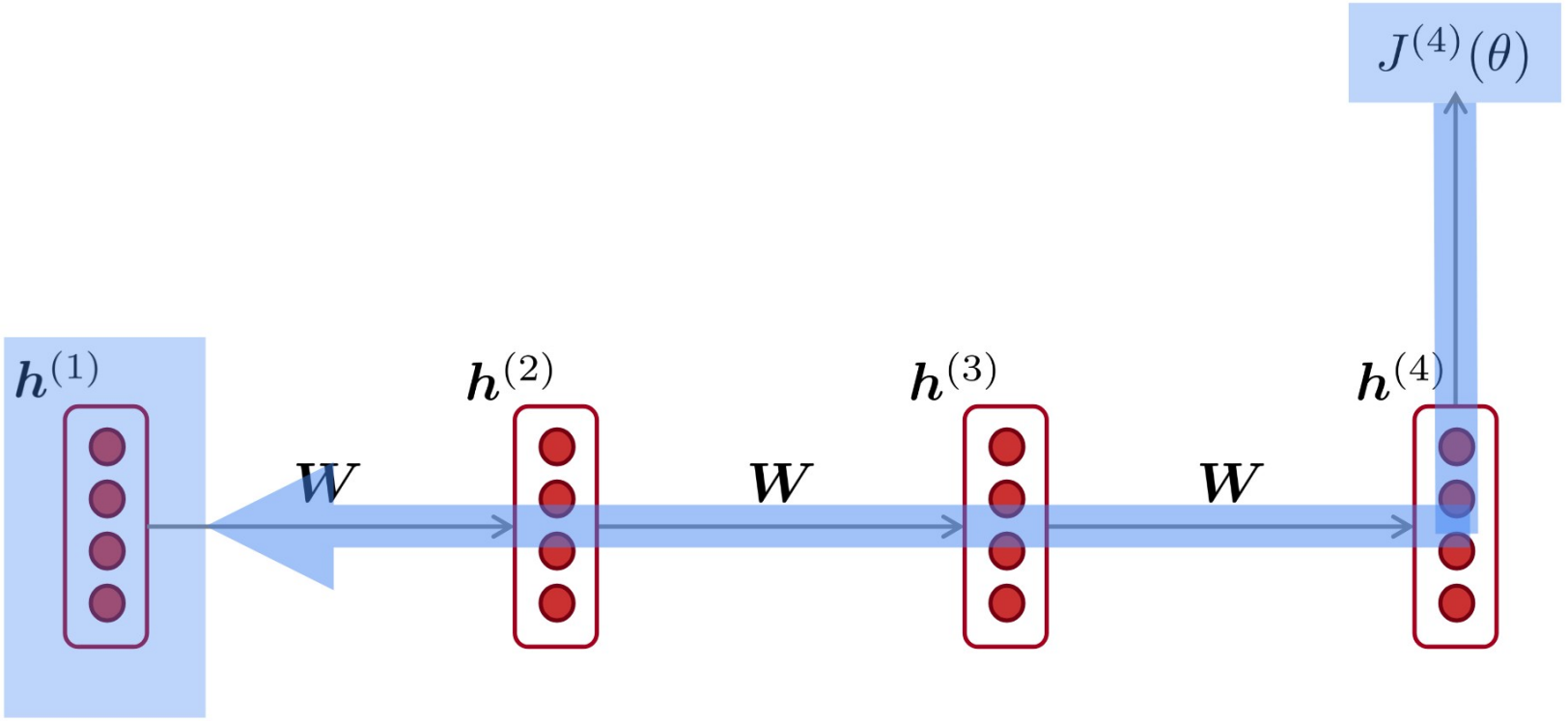
```
val_acc: 0.527000, lr: 5.340517e-04, reg: 4.097824e-01, (0 / 100)
val_acc: 0.492000, lr: 2.279484e-04, reg: 9.991345e-04, (1 / 100)
val_acc: 0.512000, lr: 8.680827e-04, reg: 1.349727e-02, (2 / 100)
val_acc: 0.461000, lr: 1.028377e-04, reg: 1.220193e-02, (3 / 100)
val_acc: 0.460000, lr: 1.113730e-04, reg: 5.244309e-02, (4 / 100)
val_acc: 0.498000, lr: 9.477776e-04, reg: 2.001293e-03, (5 / 100)
val_acc: 0.469000, lr: 1.484369e-04, reg: 4.328313e-01, (6 / 100)
val_acc: 0.522000, lr: 5.586261e-04, reg: 2.312685e-04, (7 / 100)
val_acc: 0.530000, lr: 5.808183e-04, reg: 8.259964e-02, (8 / 100)
val_acc: 0.489000, lr: 1.979168e-04, reg: 1.010889e-04, (9 / 100)
val_acc: 0.490000, lr: 2.036031e-04, reg: 2.406271e-03, (10 / 100)
val_acc: 0.475000, lr: 2.021162e-04, reg: 2.287807e-01, (11 / 100)
val_acc: 0.460000, lr: 1.135527e-04, reg: 3.905040e-02, (12 / 100)
val_acc: 0.515000, lr: 6.947668e-04, reg: 1.562808e-02, (13 / 100)
val_acc: 0.531000, lr: 9.471549e-04, reg: 1.433895e-03, (14 / 100)
val_acc: 0.509000, lr: 3.140888e-04, reg: 2.857518e-01, (15 / 100)
val_acc: 0.514000, lr: 6.438349e-04, reg: 3.033781e-01, (16 / 100)
val_acc: 0.502000, lr: 3.921784e-04, reg: 2.707126e-04, (17 / 100)
val_acc: 0.509000, lr: 9.752279e-04, reg: 2.850865e-03, (18 / 100)
val_acc: 0.500000, lr: 2.412048e-04, reg: 4.997821e-04, (19 / 100)
val_acc: 0.466000, lr: 1.319314e-04, reg: 1.189915e-02, (20 / 100)
val_acc: 0.516000, lr: 8.039527e-04, reg: 1.528291e-02, (21 / 100)
```

Vanishing and Exploding Gradients

Vanishing gradient intuition

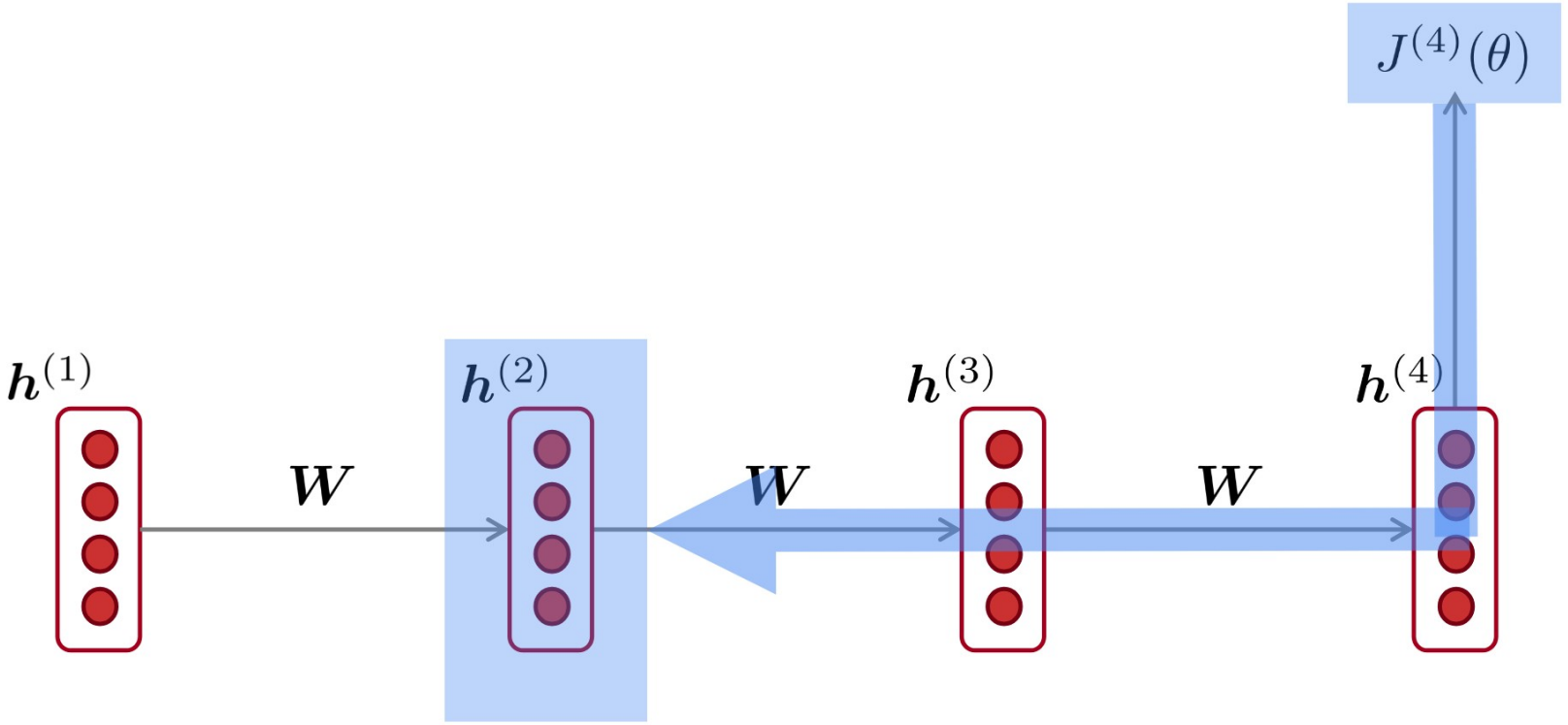


Vanishing gradient intuition



$$\frac{\partial J^{(4)}}{\partial h^{(1)}} = \ ?$$

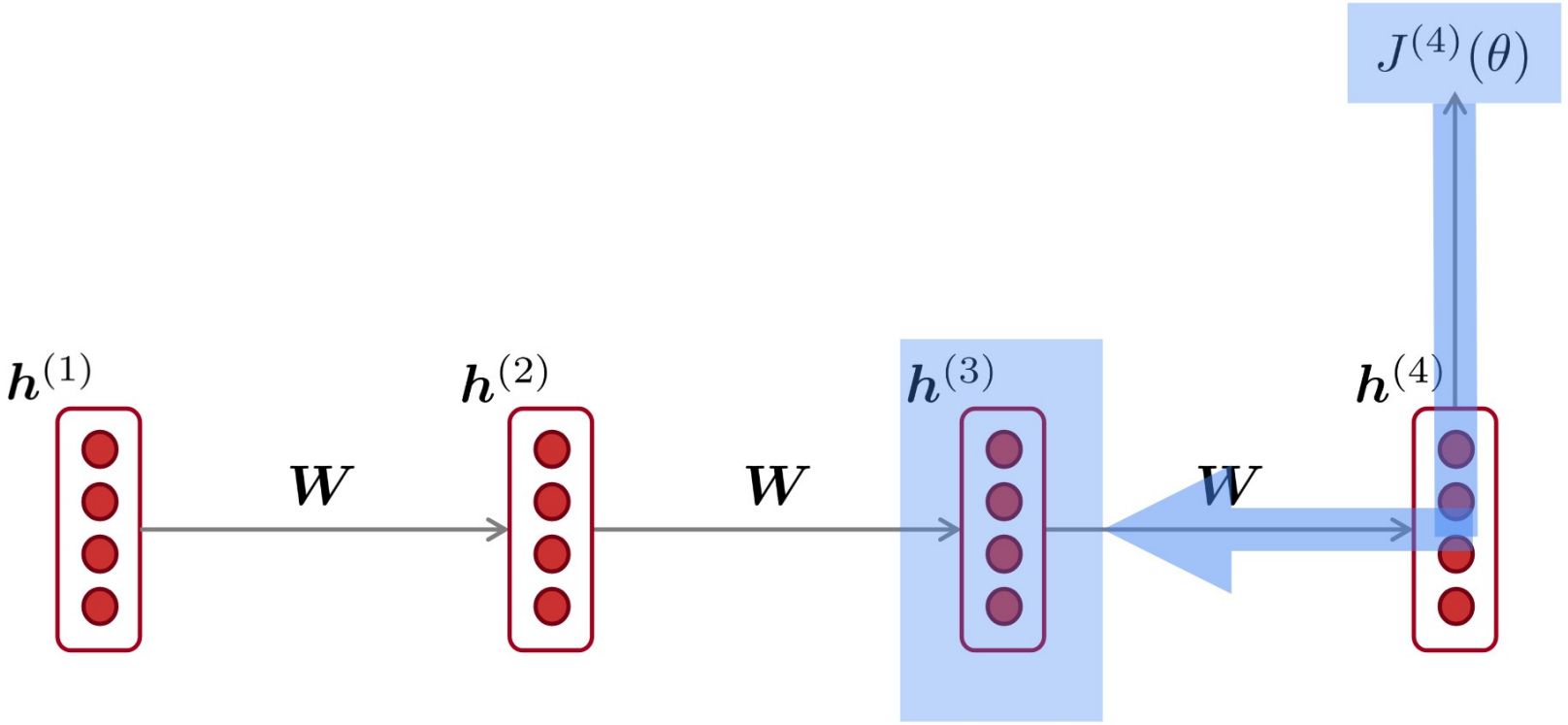
Vanishing gradient intuition



$$\frac{\partial J^{(4)}}{\partial h^{(1)}} = \frac{\partial h^{(2)}}{\partial h^{(1)}} \times \frac{\partial J^{(4)}}{\partial h^{(2)}}$$

chain rule!

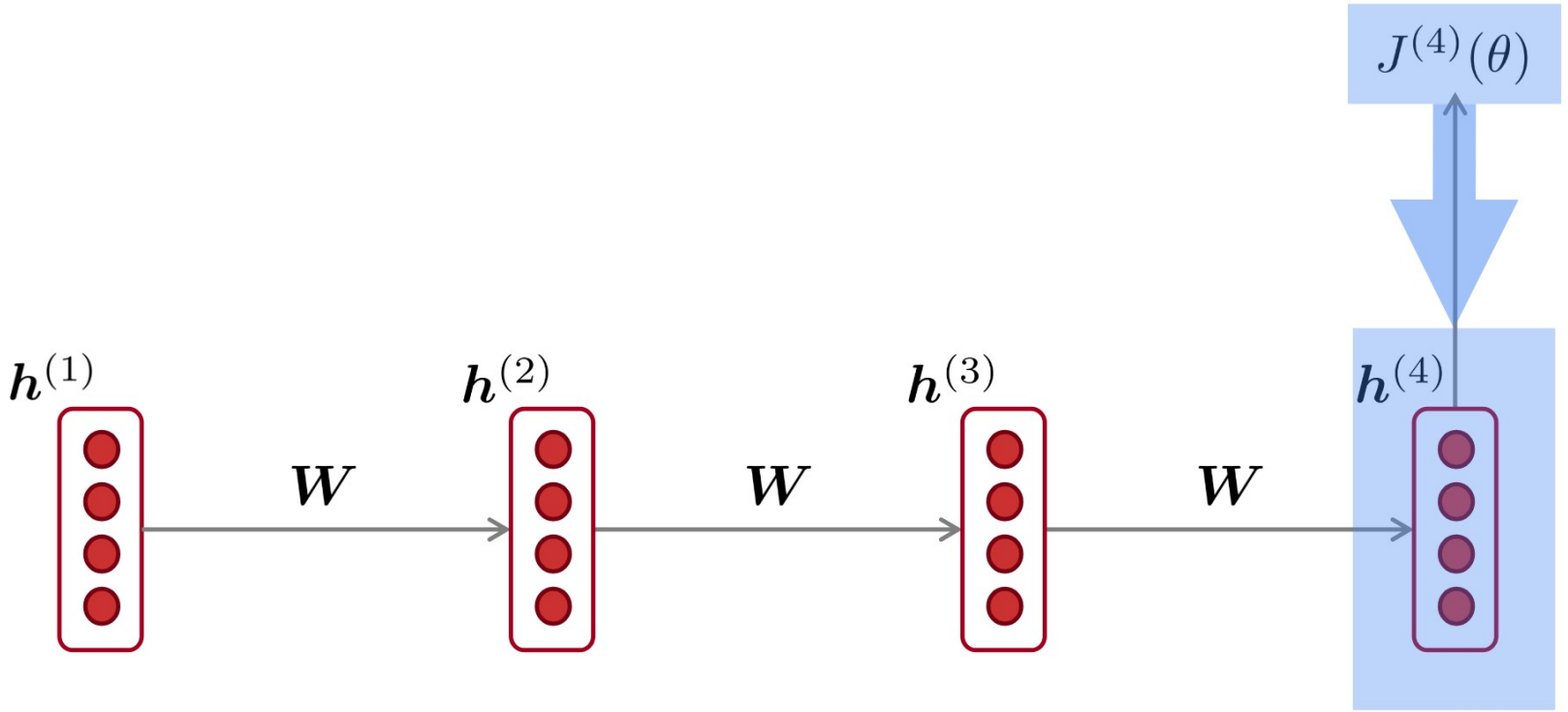
Vanishing gradient intuition



$$\frac{\partial J^{(4)}}{\partial h^{(1)}} = \frac{\partial h^{(2)}}{\partial h^{(1)}} \times \frac{\partial h^{(3)}}{\partial h^{(2)}} \times \frac{\partial J^{(4)}}{\partial h^{(3)}}$$

chain rule!

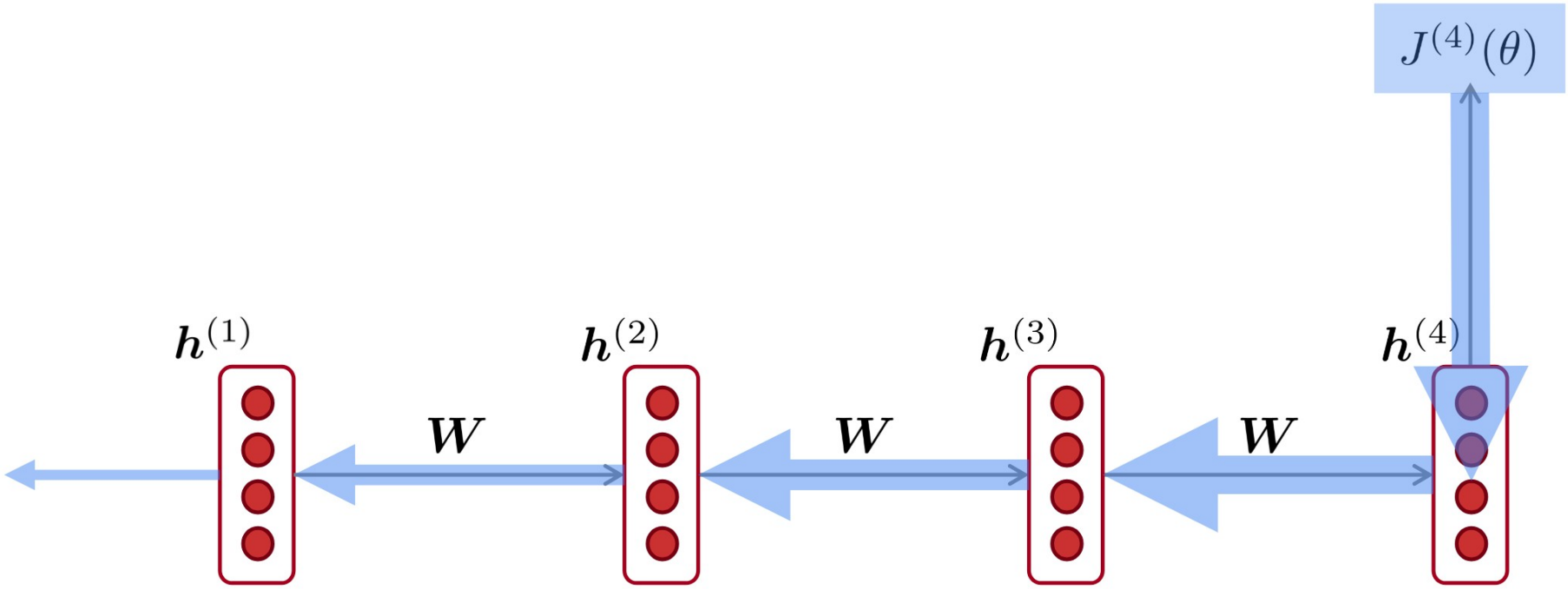
Vanishing gradient intuition



$$\frac{\partial J^{(4)}}{\partial h^{(1)}} = \frac{\partial h^{(2)}}{\partial h^{(1)}} \times \frac{\partial h^{(3)}}{\partial h^{(2)}} \times \frac{\partial h^{(4)}}{\partial h^{(3)}} \times \frac{\partial J^{(4)}}{\partial h^{(4)}}$$

chain rule!

Vanishing gradient intuition

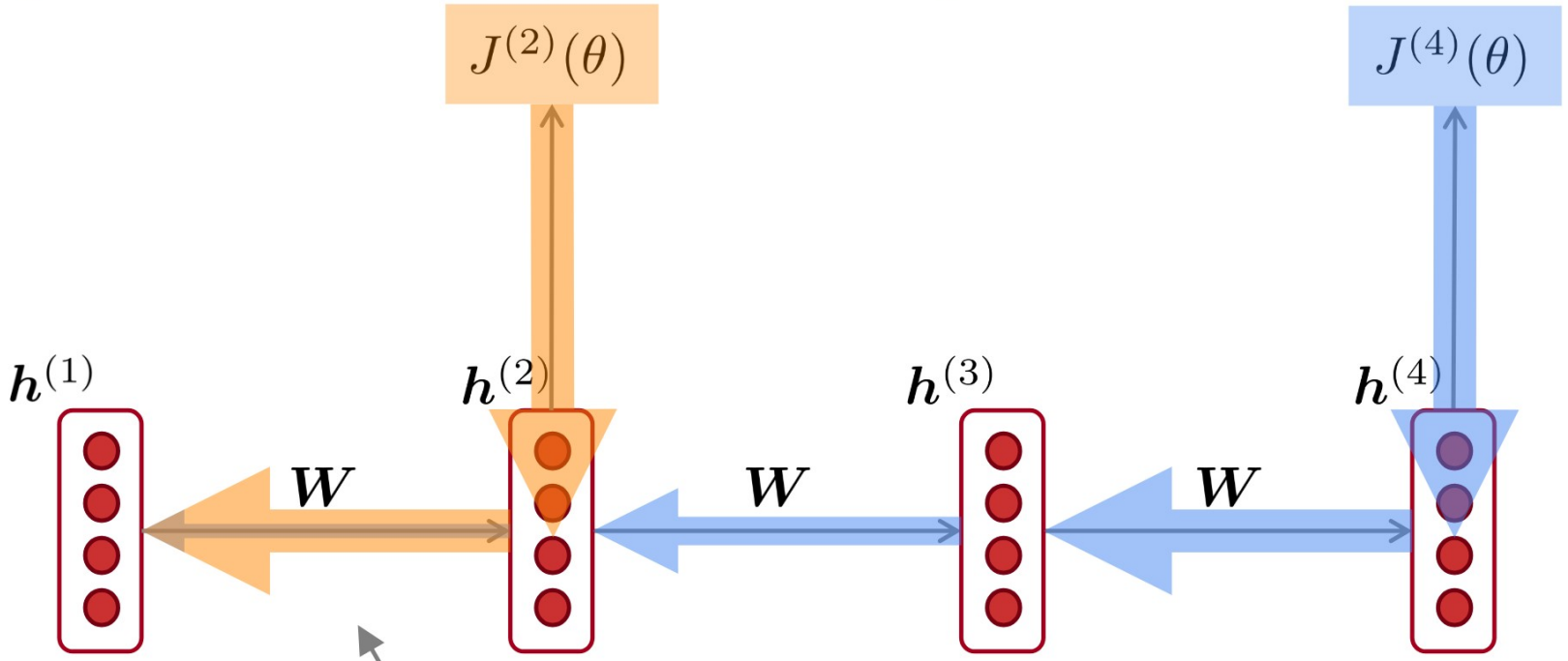


$$\frac{\partial J^{(4)}}{\partial h^{(1)}} = \frac{\partial h^{(2)}}{\partial h^{(1)}} \times \frac{\partial h^{(3)}}{\partial h^{(2)}} \times \frac{\partial h^{(4)}}{\partial h^{(3)}} \times \frac{\partial J^{(4)}}{\partial h^{(4)}}$$

What happens if these are small?

Vanishing gradient problem:
When these are small, the gradient signal gets smaller and smaller as it backpropagates further

Why is vanishing gradient a problem?



Gradient signal from faraway is lost because it's much smaller than gradient signal from close-by.

So model weights are updated only with respect to near effects, not long-term effects.

Why is exploding gradient a problem?

- If the gradient becomes too big, then the SGD update step becomes too big:

$$\theta^{new} = \theta^{old} - \overbrace{\alpha}^{\text{learning rate}} \underbrace{\nabla_\theta J(\theta)}_{\text{gradient}}$$

- This can cause bad updates: we take too large a step and reach a bad parameter configuration (with large loss)

- In the worst case, this will result in Inf or NaN in your network (then you have to restart training from an earlier checkpoint)

Gradient Clipping

Gradient clipping: solution for exploding gradient

- Gradient clipping: if the norm of the gradient is greater than some threshold, scale it down before applying SGD update

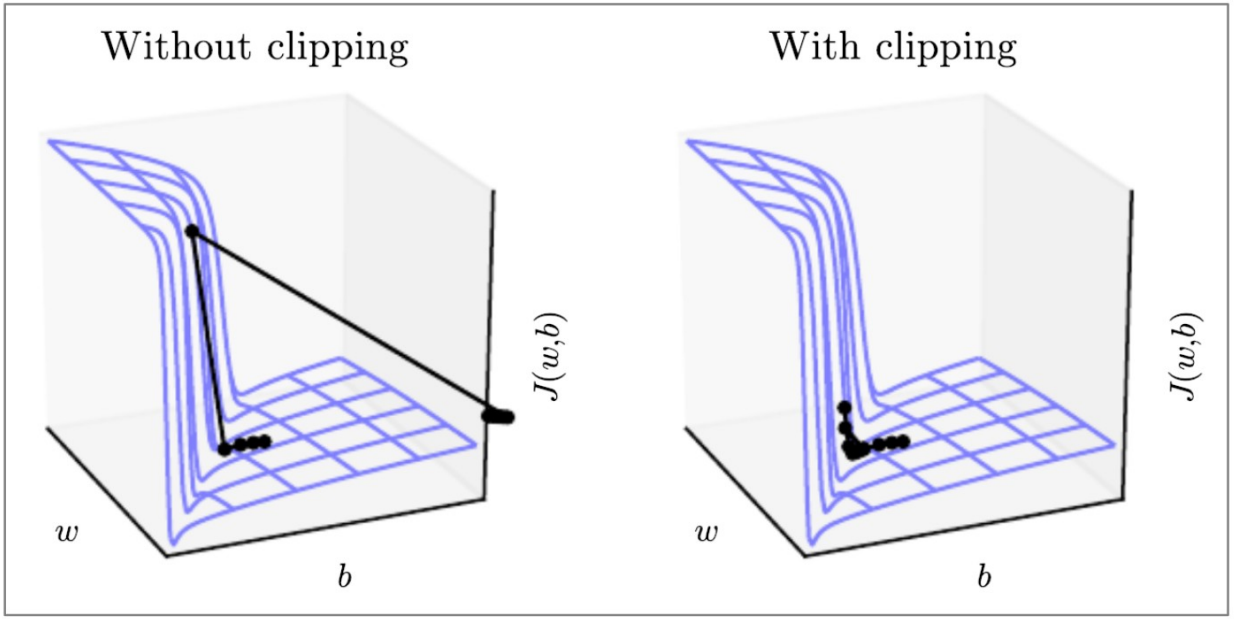
Algorithm 1 Pseudo-code for norm clipping

$\hat{\mathbf{g}} \leftarrow \frac{\partial \mathcal{E}}{\partial \theta}$
if $\|\hat{\mathbf{g}}\| \geq threshold$ **then**
  $\hat{\mathbf{g}} \leftarrow \frac{threshold}{\|\hat{\mathbf{g}}\|}\hat{\mathbf{g}}$
end if

- Intuition: take a step in the same direction, but a smaller step

Source: "On the difficulty of training recurrent neural networks", Pascanu et al, 2013. http://proceedings.mlr.press/v28/pascanu13.pdf

Gradient clipping: solution for exploding gradient



- This shows the loss surface of a simple RNN (hidden state is a scalar not a vector)
- The "cliff" is dangerous because it has steep gradient
- On the left, gradient descent takes two very big steps due to steep gradient, resulting in climbing the cliff then shooting off to the right (both bad updates)
- On the right, gradient clipping reduces the size of those steps, so effect is less drastic

Source: "Deep Learning", Goodfellow, Bengio and Courville, 2016. Chapter 10.11.1. https://www.deeplearningbook.org/contents/rnn.html

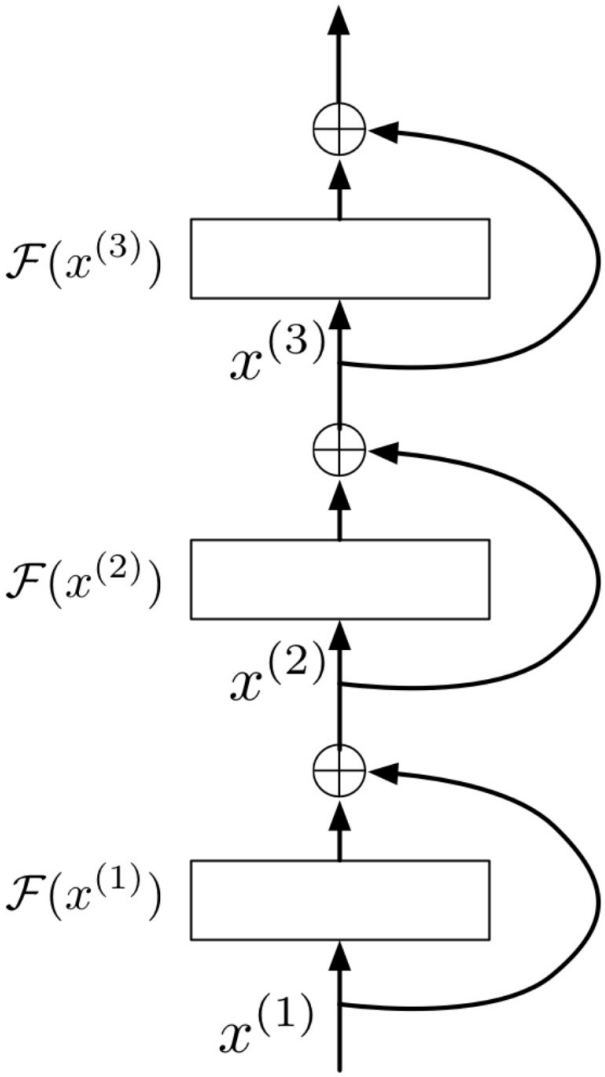
Residual Connections

Deep Residual Networks

- We can string together a bunch of residual units.
- What happens if we set the parameters such that $\mathcal{F}(\mathbf{x}^{(\ell)}) = 0$ in every layer?
  - Then it passes $\mathbf{x}^{(1)}$ straight through unmodified!
  - This means it's easy for the network to represent the identity function.
- Backprop:

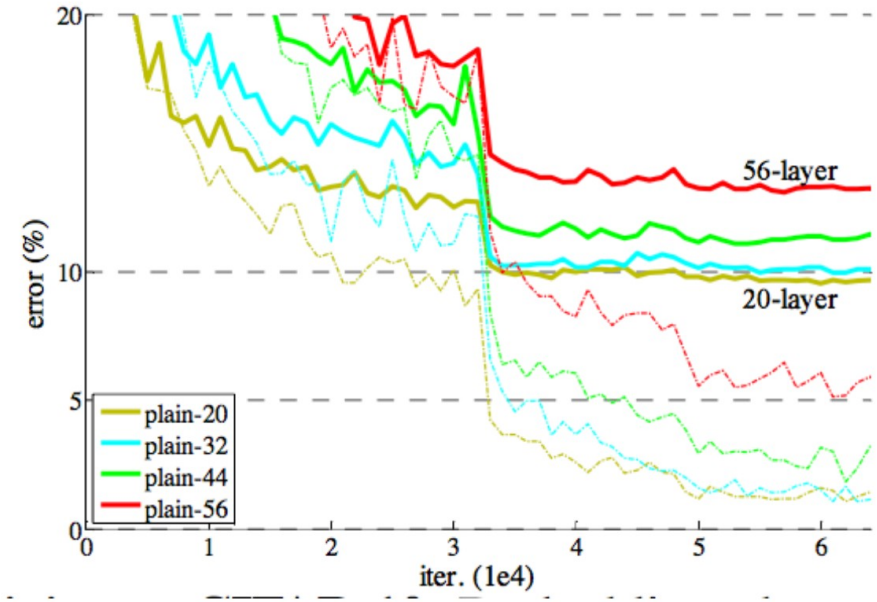
$$\begin{aligned}\overline{\mathbf{x}^{(\ell)}} &= \overline{\mathbf{x}^{(\ell+1)}} + \overline{\mathbf{x}^{(\ell+1)}} \frac{\partial \mathcal{F}}{\partial \mathbf{x}} \\ &= \overline{\mathbf{x}^{(\ell+1)}} \left(\mathbf{I} + \frac{\partial \mathcal{F}}{\partial \mathbf{x}} \right)\end{aligned}$$

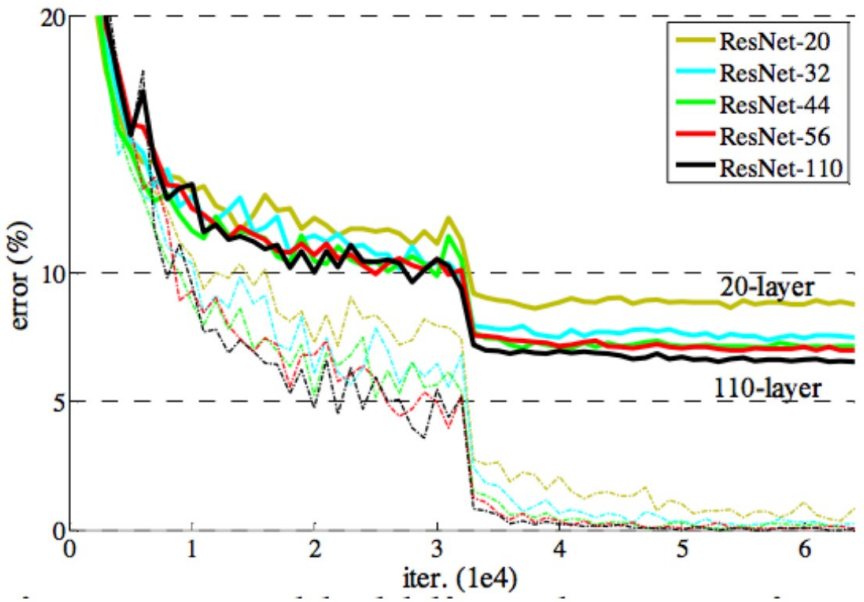
- This means the derivatives don't vanish.



Deep Residual Networks

- Deep Residual Networks (ResNets) consist of many layers of residual units.
- For vision tasks, the $\mathcal{F}$ functions are usually 2- or 3-layer conv nets.
- Performance on CIFAR-10, a small object recognition dataset:





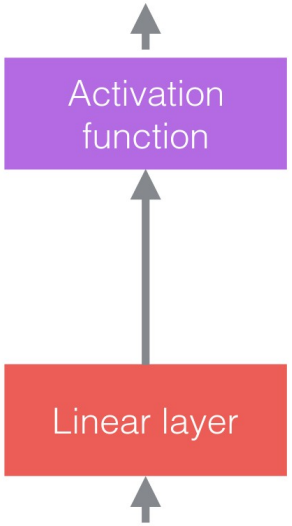
- For a regular convnet, performance declines with depth, but for a ResNet, it keeps improving.

Deep Residual Networks

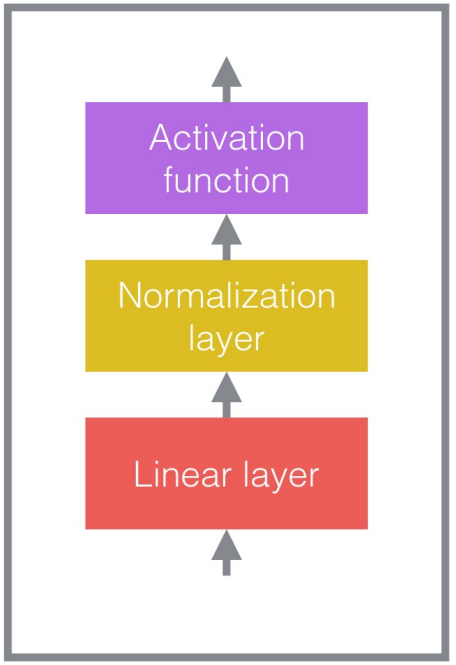
- A 152-layer ResNet achieved 4.49% top-5 error on Image Net. An ensemble of them achieved 3.57%.
- Previous state-of-the-art: 6.6% (GoogLeNet)
- Humans: 5.1%
- They were able to train ResNets with more than 1000 layers, but classification performance leveled off by 150.
- What are all these layers doing? We don't have a clear answer, but the idea that they're computing increasingly abstract features is starting to sound fishy...

Normalization Layers

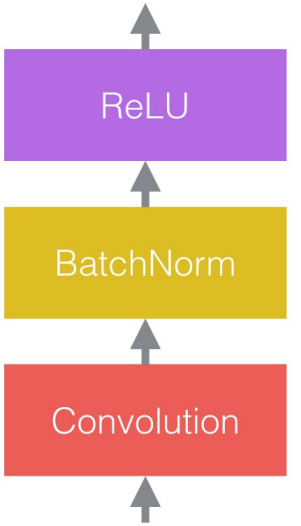
Normalization layers



They are added in-between existing layers of a neural network

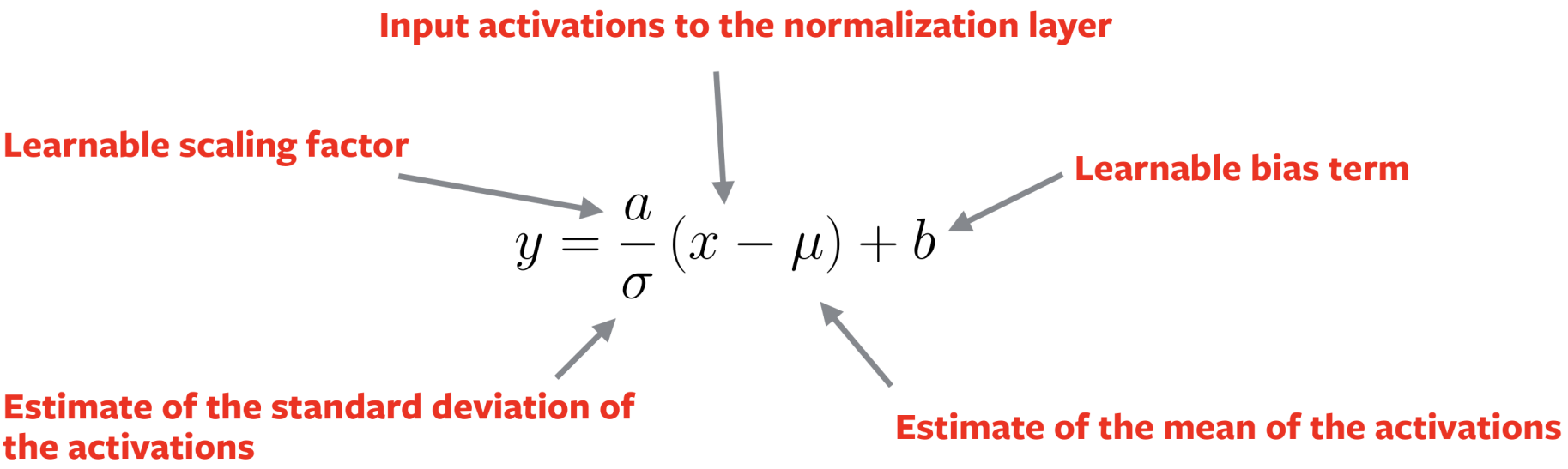


Typical image classification structure



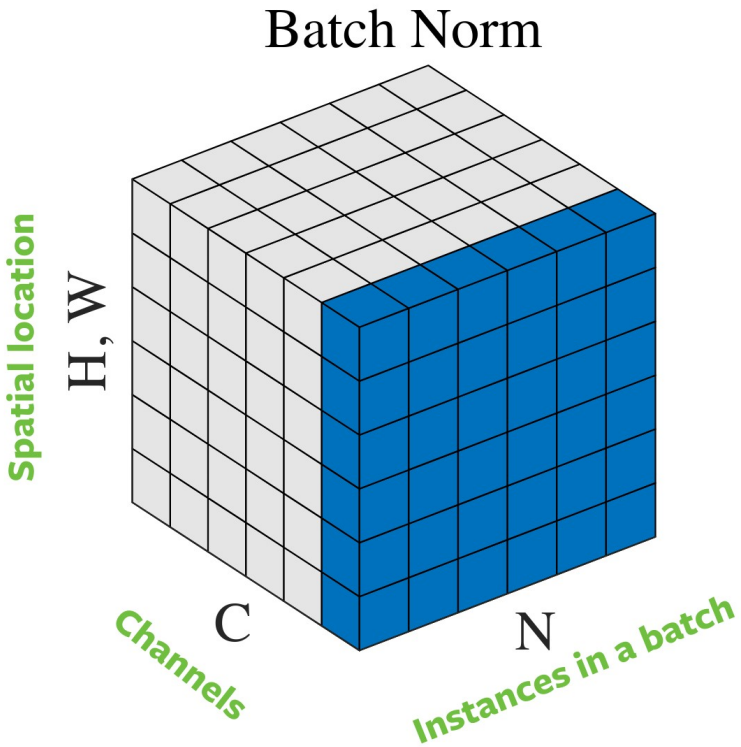
What does it mean to normalize?

Most normalization layers **whiten** activations

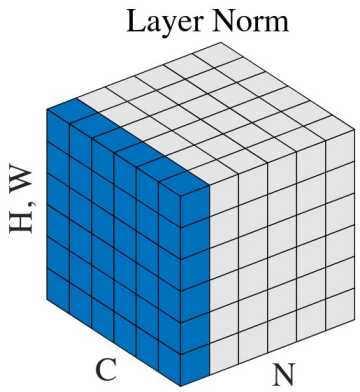


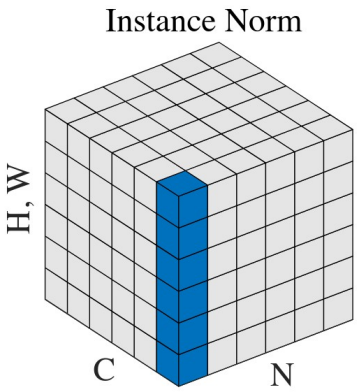
The extra a & b parameters keep the representation power of the network the same

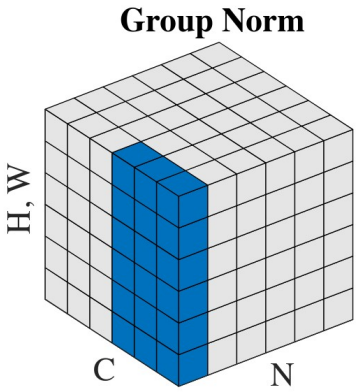
But how do we estimate the mean and standard deviation? Does it differ across channels, spatial location, instances?



Normalize across H,W,N
i.e. across the **batch**, and
Within each separate image channel







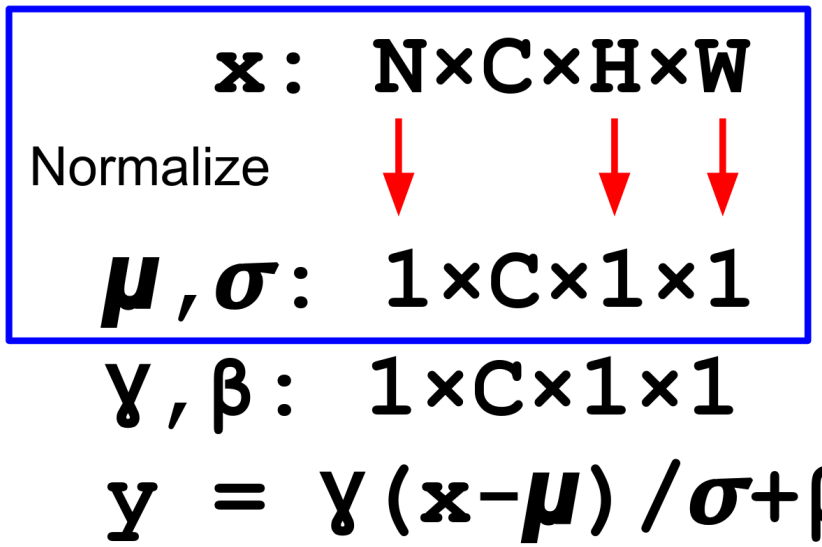
Normalize within
H,W and a **group** of
channels

In practice group/batch norm work well for computer vision problems, and instance/layer norm are heavily used for language problems.

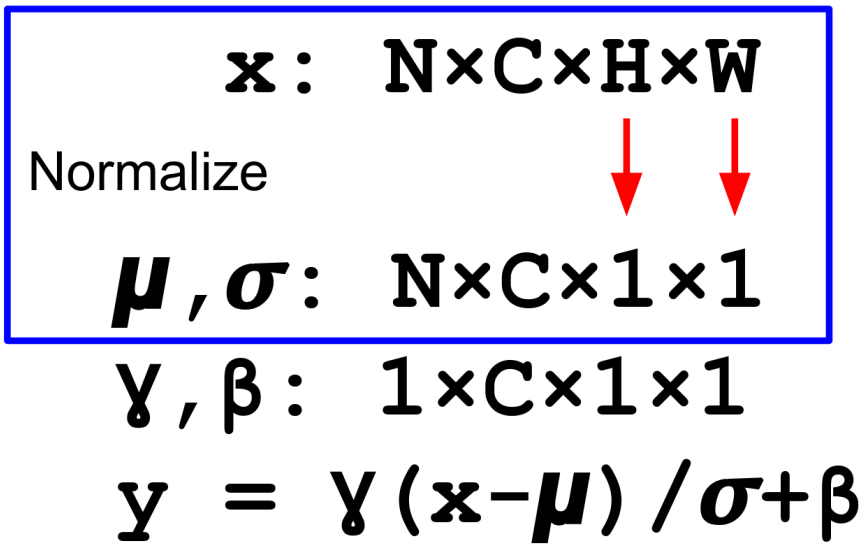
IMG SOURCE: https://arxiv.org/pdf/1803.08494.pdf

Instance Normalization

Batch Normalization for convolutional networks



Instance Normalization for convolutional networks
Same behavior at train / test!



Why does normalization help?

This is still disputed

The original paper said that it “reduces internal covariate shift” whatever that means

As usual, we are using something we don’t fully understand.

But, it’s clearly a combination of a number of factors:

- Networks with normalization layers are easier to optimize, allowing for the use of larger learning rates. (Normalization has a **optimization** effect)
- The mean/std estimates are noisy, this extra “noise” results in better generalization in some cases (Normalization has an **regularization** effect)
- Normalization reduces sensitivity to weight initialization