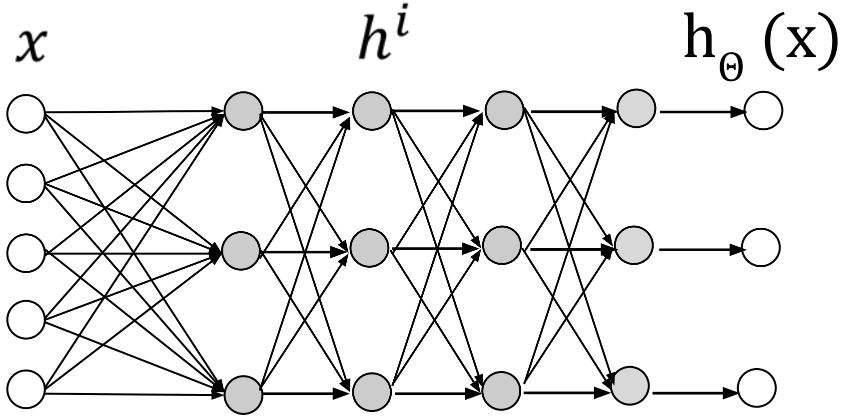
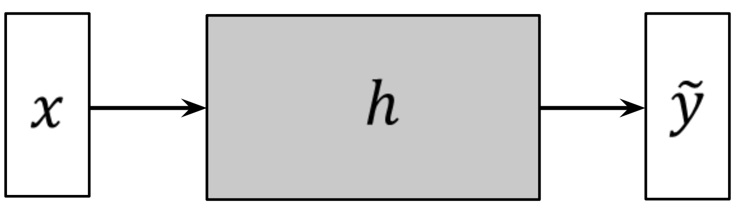
Announcements

- Quiz-3 out today.
- Pset-3 due 03/27

Neural networks: recap





Learn parameters via **gradient descent**

$$\min_{\Theta} J(\Theta)$$

Backpropagation efficiently computes cost (forward pass) and gradient (backward pass)

$$\frac{\partial}{\partial \Theta_{ij}^{(l)}} J(\Theta)$$

Gradient Descent

- Start somewhere = random initialization.
- Compute slope = compute gradient of the cost function wrt parameters.
- Take a step towards steepest direction
- Repeat, for certain steps, stopping criteria.

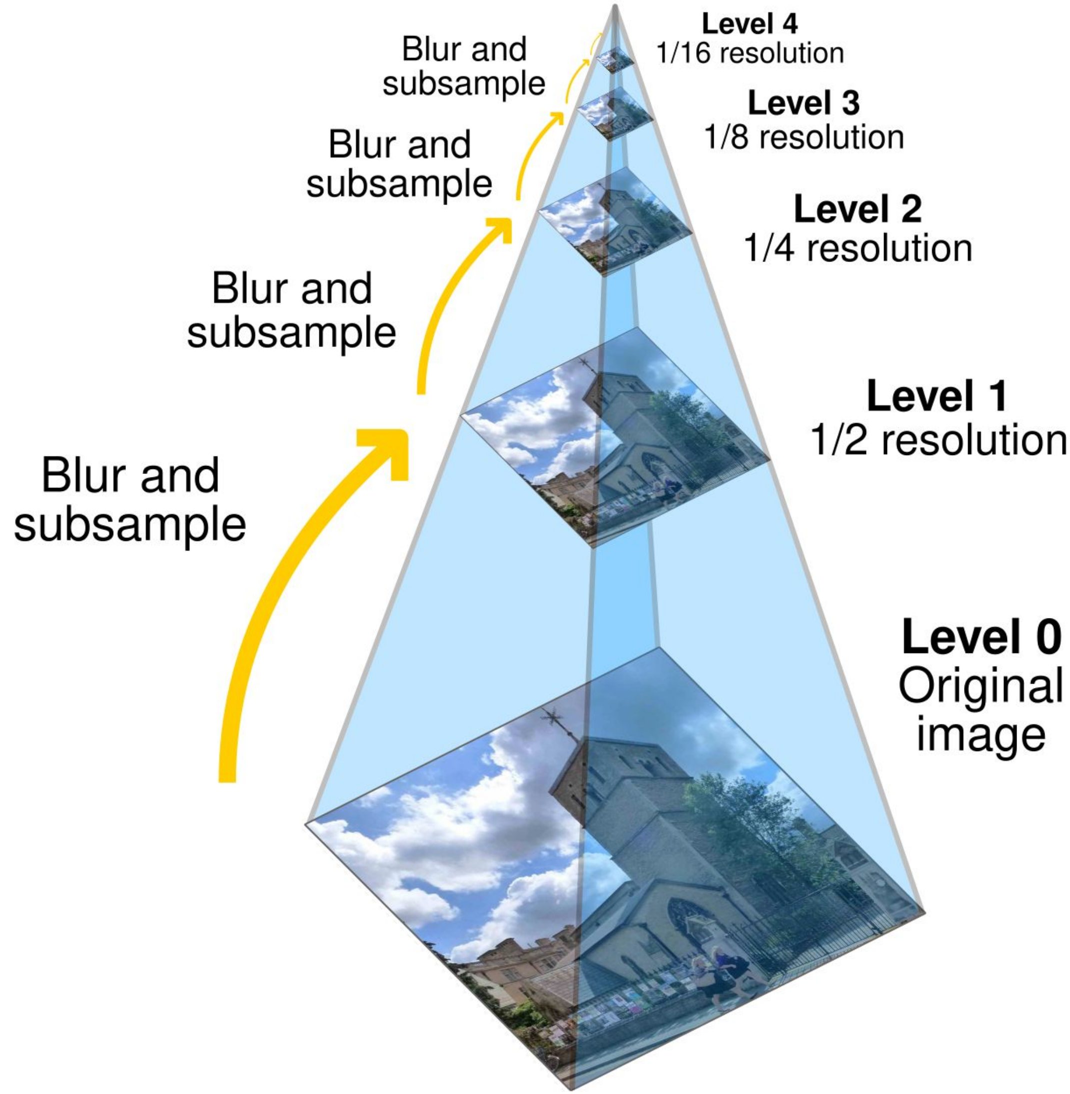
Question from last class

Fig. 1: Teaser figures, from three models and our approach MoCE, showcase a classic example of Latent Concept Misalignment (LC-Mis) in this study: a tea cup of iced coke. Here, a glass cup, an unfamiliar object, substitutes the anticipated tea cup. We denote the iced coke as Concept $\mathcal{A}$, the tea cup as Concept $\mathcal{B}$, and introduce a latent Concept $\mathcal{C}$ —the glass. This combination of $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ forms our investigative focus.

Prompt: “A tea cup of iced coke”

Lost in Translation: Latent Concept Misalignment in Text-to-Image Diffusion Models

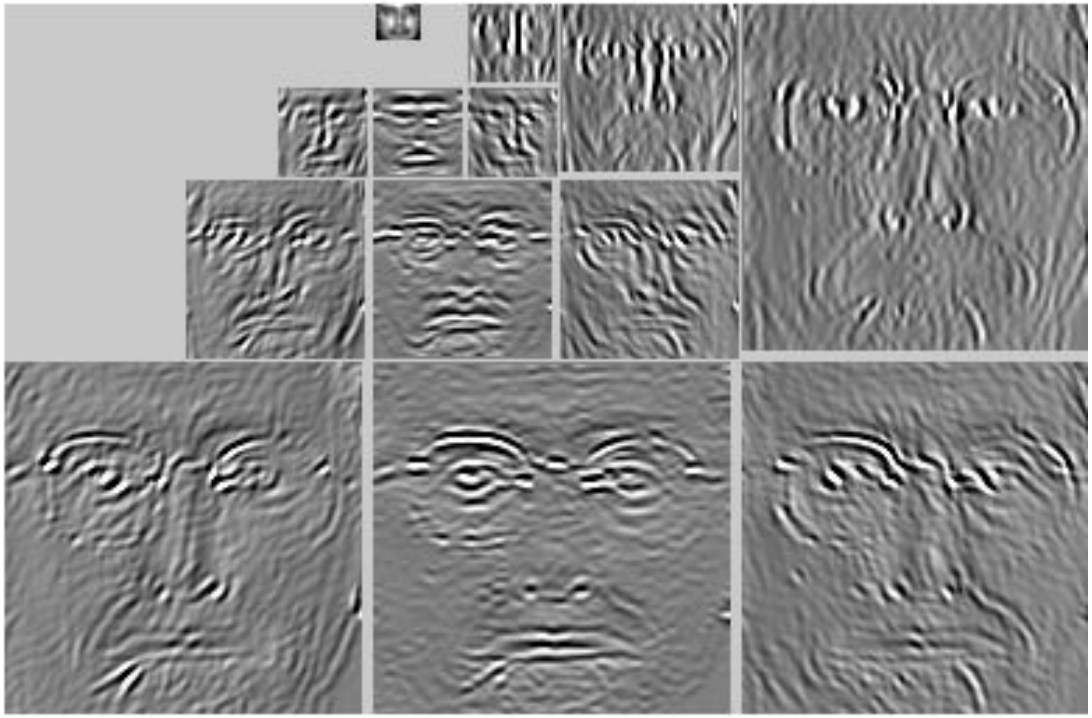
Steerable pyramids



Apply spatial filters (eg: edge detector) at every level!

Steerable pyramid

a



b

Today

- Network regularization
- Data Augmentation
- Convolutional Neural Networks

Today

- **Network regularization**
- Data Augmentation
- Convolutional Neural Networks

What are some regularization techniques we have learnt so far? Select all that apply

Regularizing Neural Nets (L2 Loss)

Recall from linear regression:

- We can regularize the model by minimizing the square of the weights.
- We can do the same with neural nets!

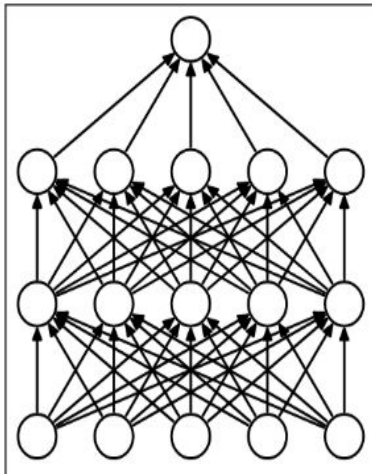
$$\boxed{\text{Error}} + \boxed{\text{Regularizer}}$$

$$\boxed{\frac{1}{N}(y - X\beta)^T(y - X\beta)} + \boxed{\lambda\beta^T\beta}$$

Regularizing Neural Nets (Dropout)

Issue: Some “neurons” might depend only on a handful of “neurons” from the last layer.

- We want diversity!

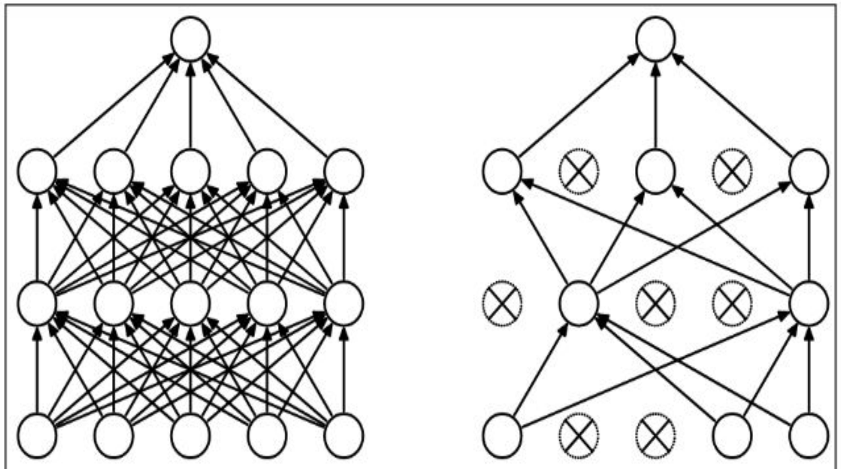


(a) Standard Neural Net

Regularizing Neural Nets (Dropout)

Issue: Some "neurons" might depend only on a handful of "neurons" from the last layer.

- We want diversity!
- Drop some connections during training.
- ***Use all connections at inference!***



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

How to decide which network connections to drop? Select all that apply

Scale of different features

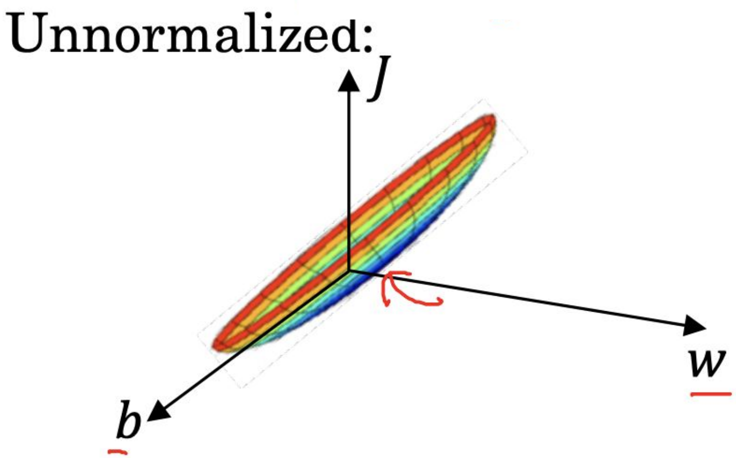
Consider a single layer $y = Wx$

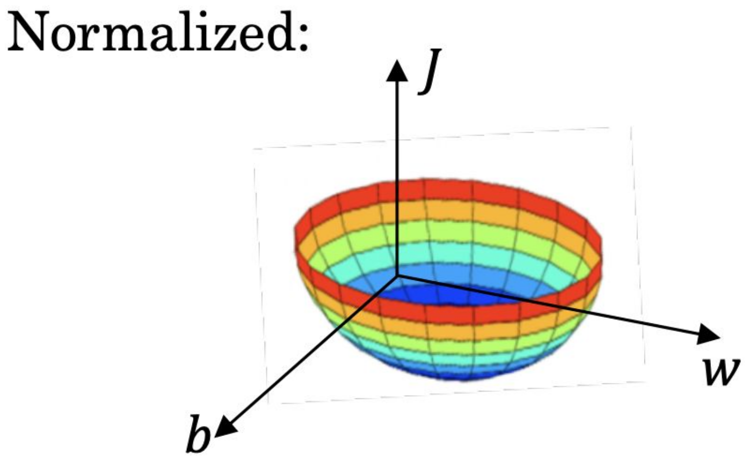
The following could lead to tough optimization:

- Inputs x are not centered around zero
- Inputs x have different scaling per element (entries in W will need to vary a lot)

Idea: force inputs to be “nicely scaled” at each layer!

Solution: Feature Normalization

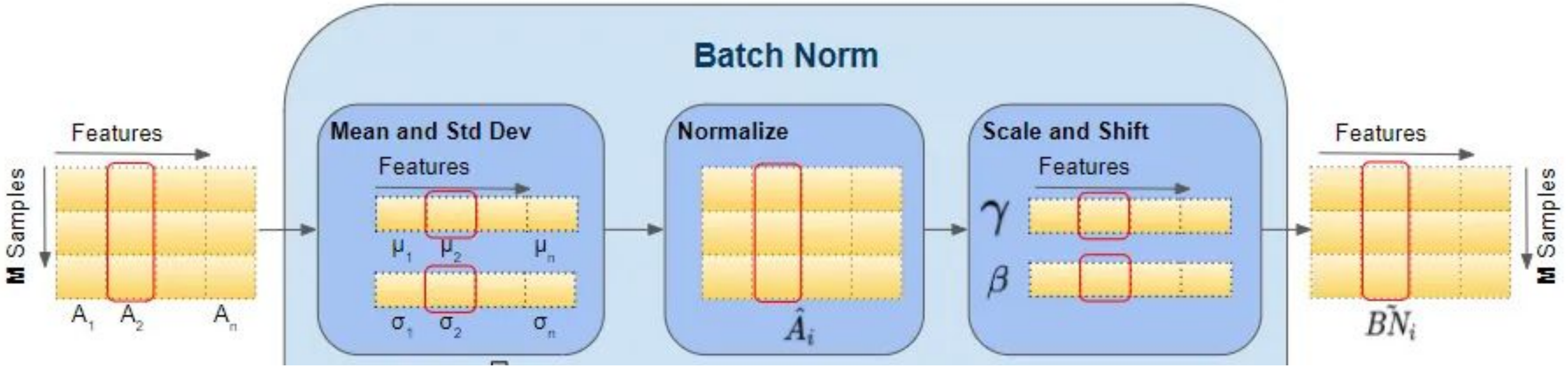




- Center the values around zero.
- Scale the values to fall between a fixed range.

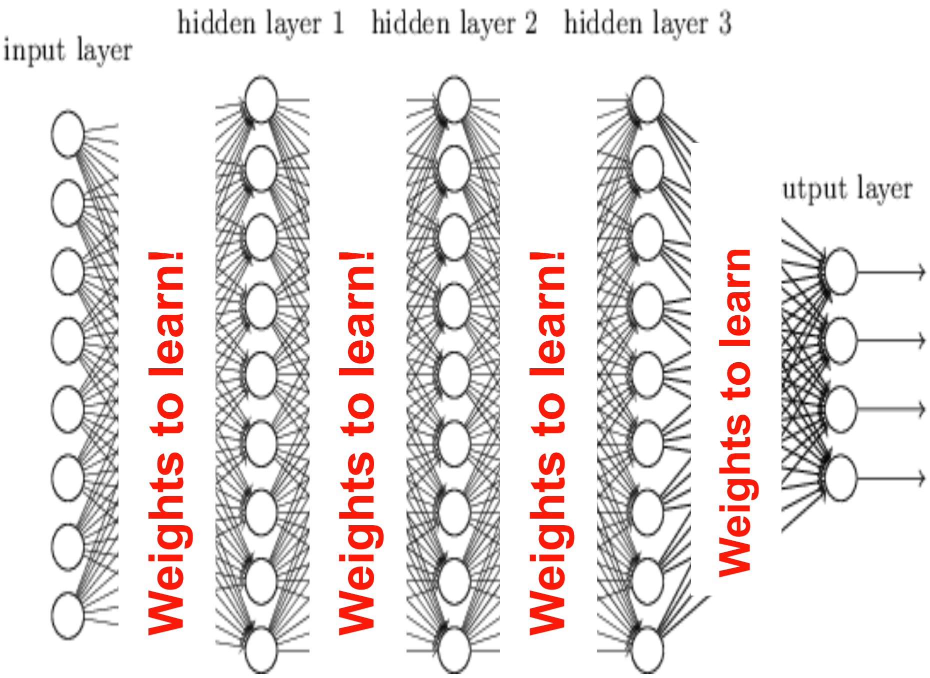
Batch Normalization:

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

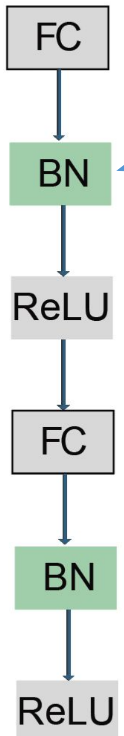


Discussion

1. Where all in the deep net pipeline should we introduce the normalization?
2. Why?



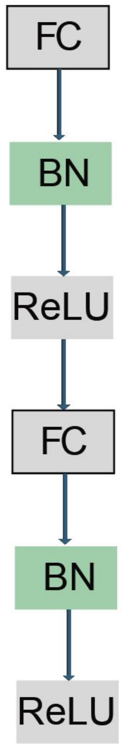
Batch Normalization



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

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

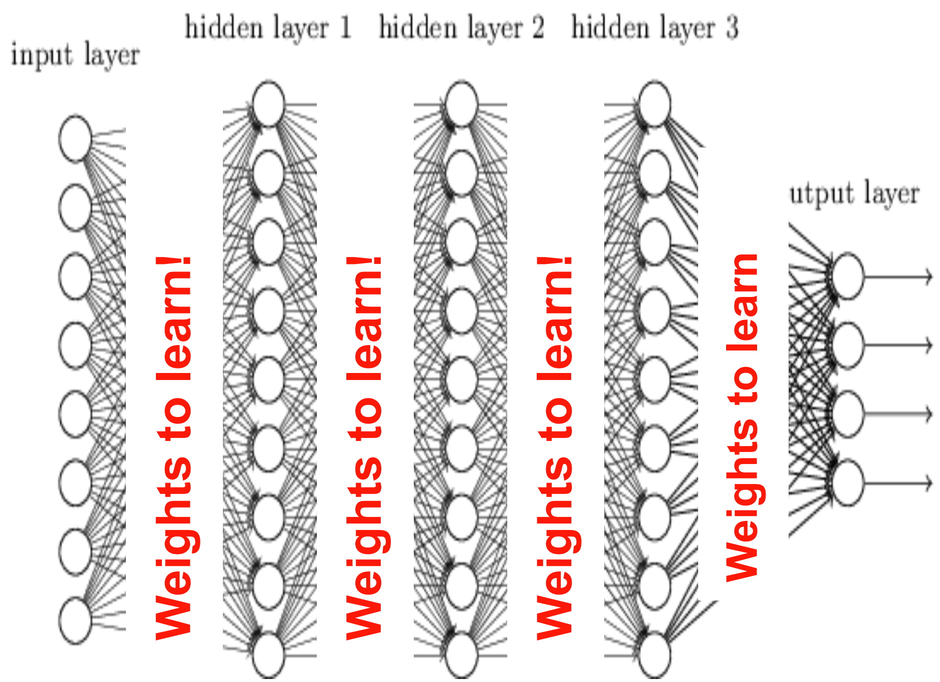
Batch Normalization:



- Allows higher learning rates, faster convergence
- Acts as a kind of regularization during training

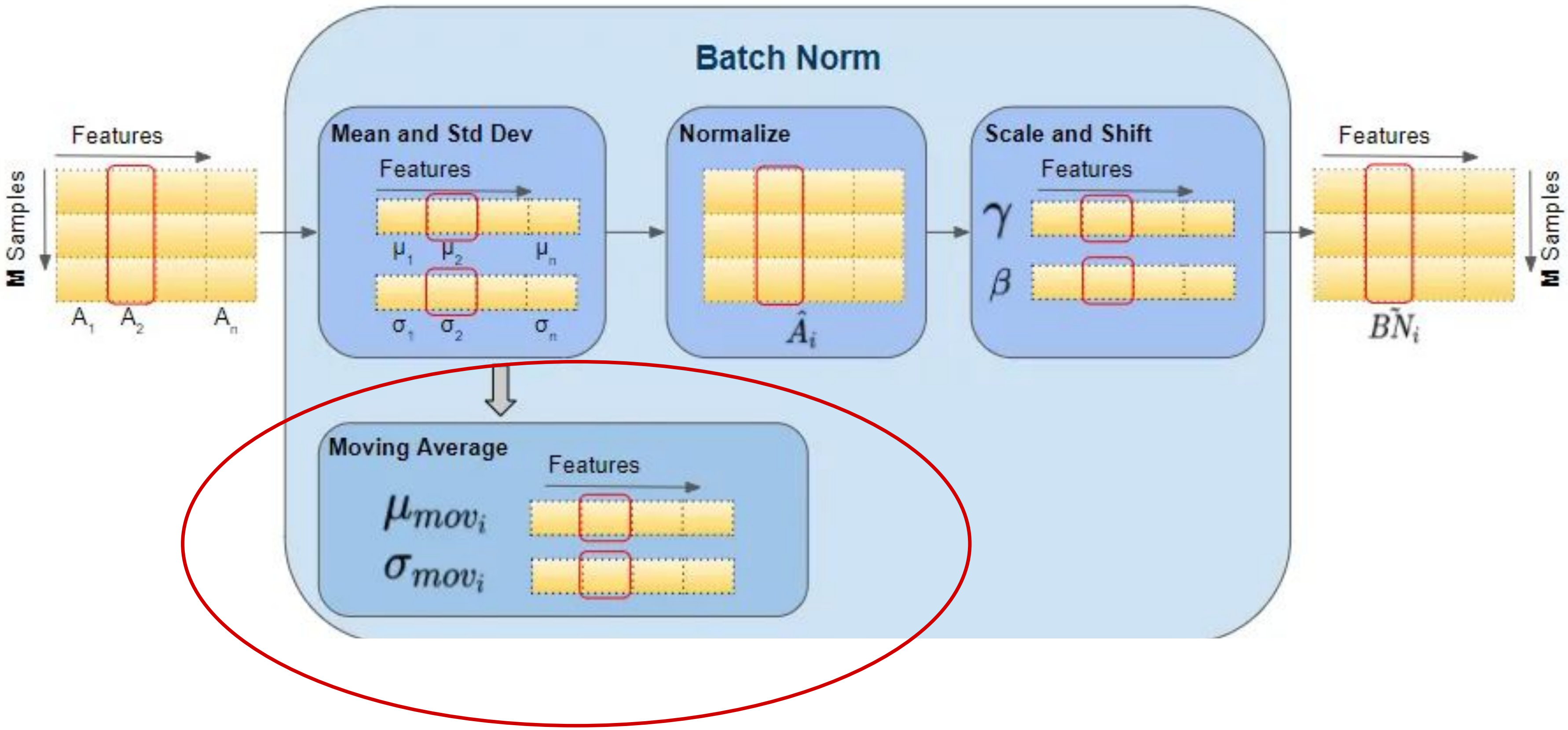
Discussion

- As the training progresses, should we update the mean and variance?
- Why or why not?



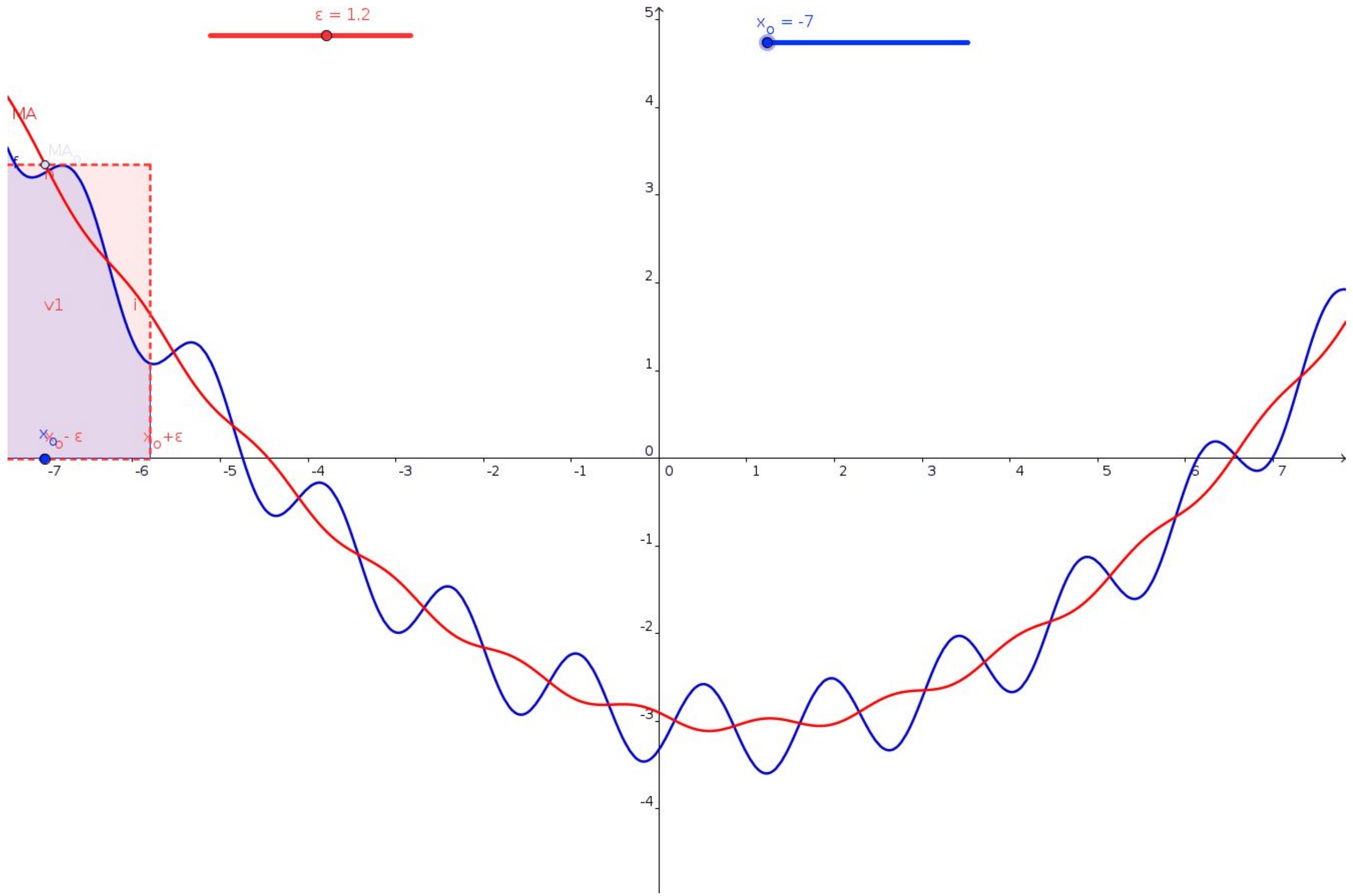
As the training progresses, should we update the mean and variance? Select all that apply

Should we update the mean and variance?



Why or why not? To keep up with the shifting data.

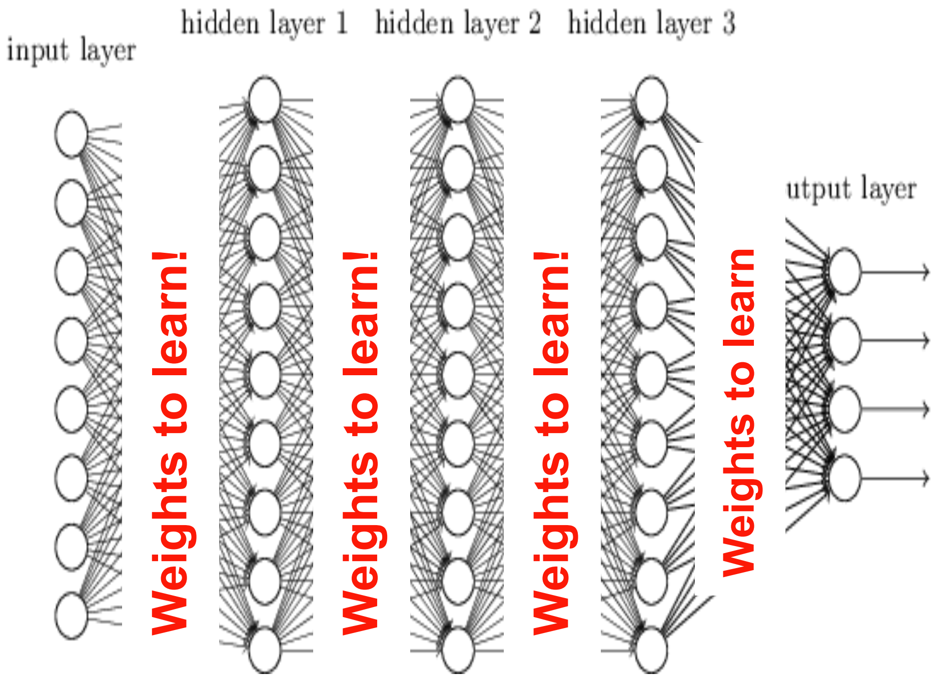
Moving average



- Moving average results in smoother updates to mean, variance.

Discussion

How do we know what mean and variance to use during inference?



How do we know what mean and variance to use during inference?

How do we know what mean and variance to use during inference?

- **Retain** the ones computed during training.

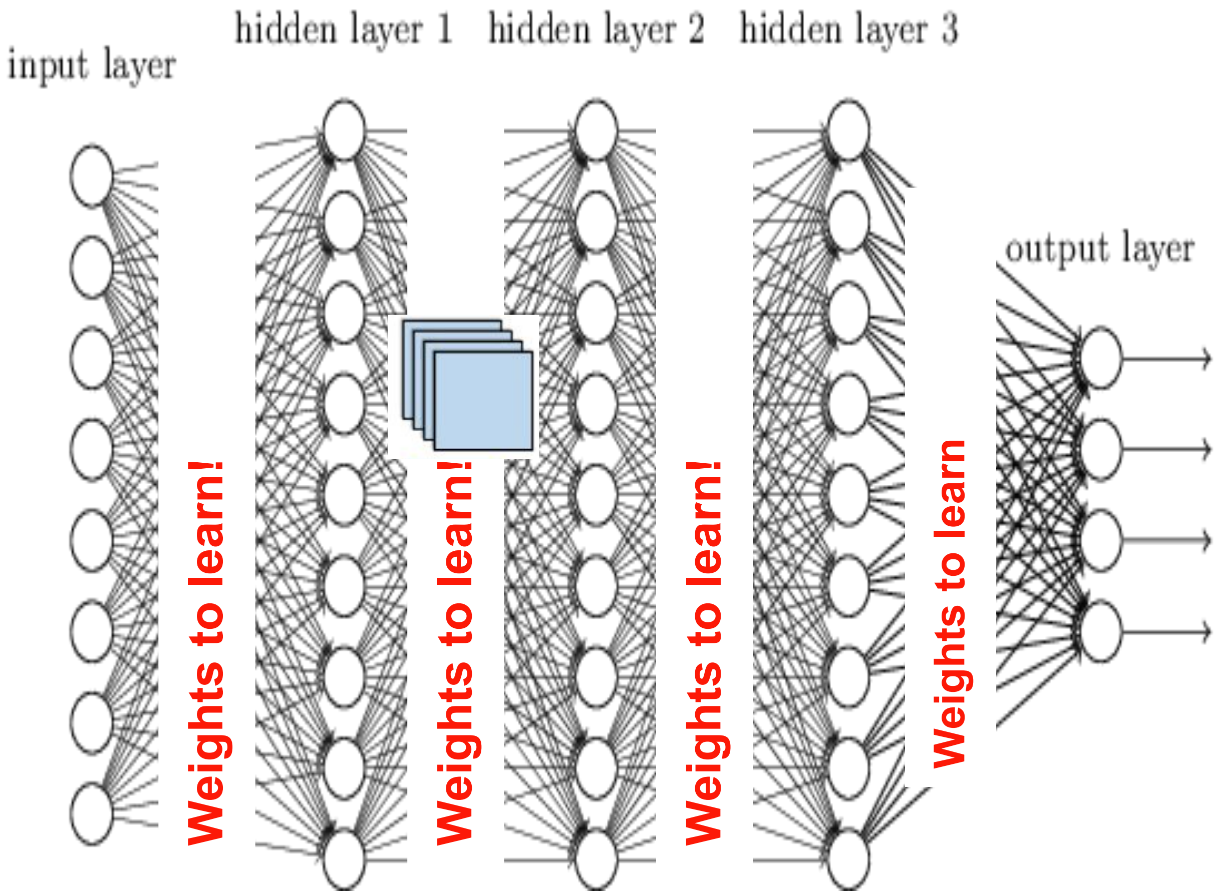
- This is a very common source of bugs!

Summary so far

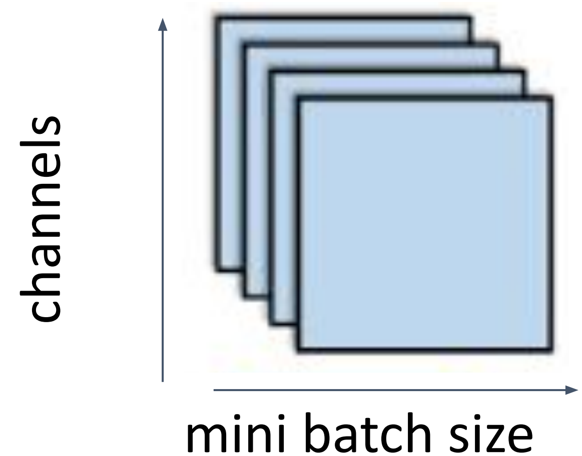
1. Network regularization
    a. Dropout
    b. Batch normalization

2. Data Augmentation
3. Convolutional Neural Networks

How does the output of every layer look like?

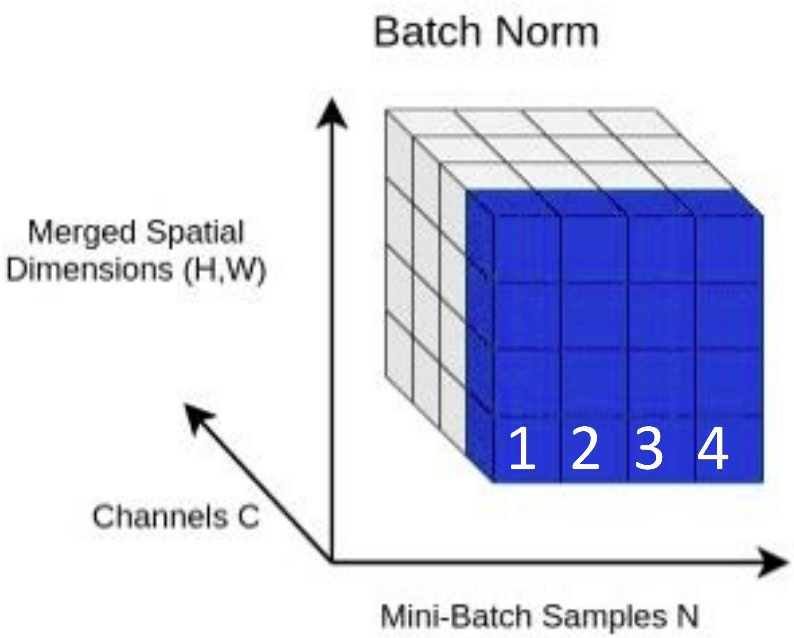


How does the output of every layer look like?



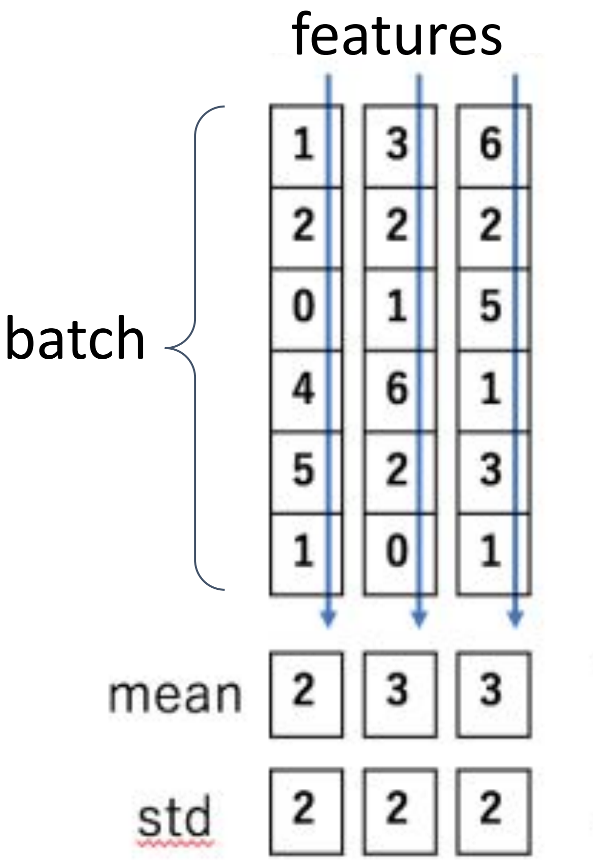
What does this plot convey?

The pixels in blue are normalized by the same mean and variance



- For a given features (HW X C), across different batches, we apply the same mean

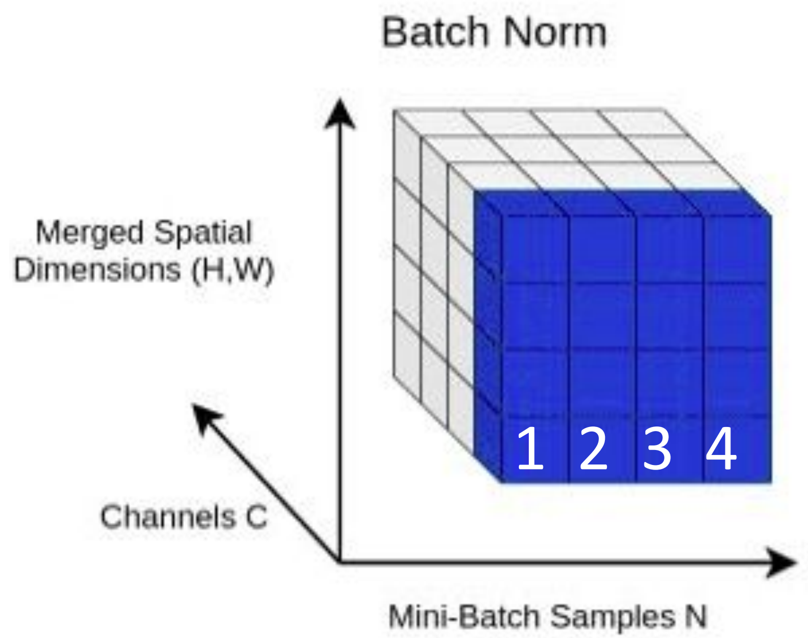
Easier to visualize in 2D



- For a given features, across different batches, we apply the same mean

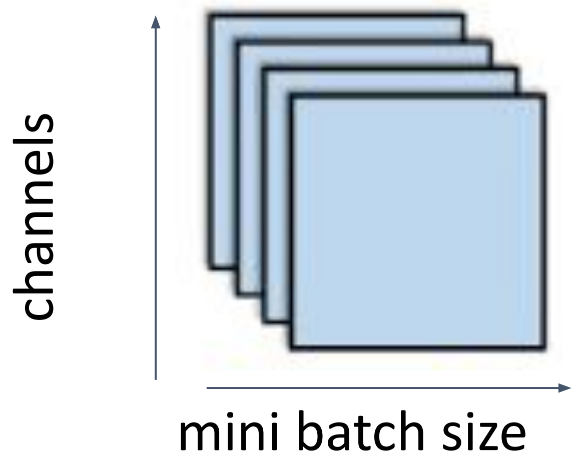
What does this plot convey?

The pixels in blue are normalized by the same mean and variance



- For a given features (HW X C), across different batches, we apply the same mean

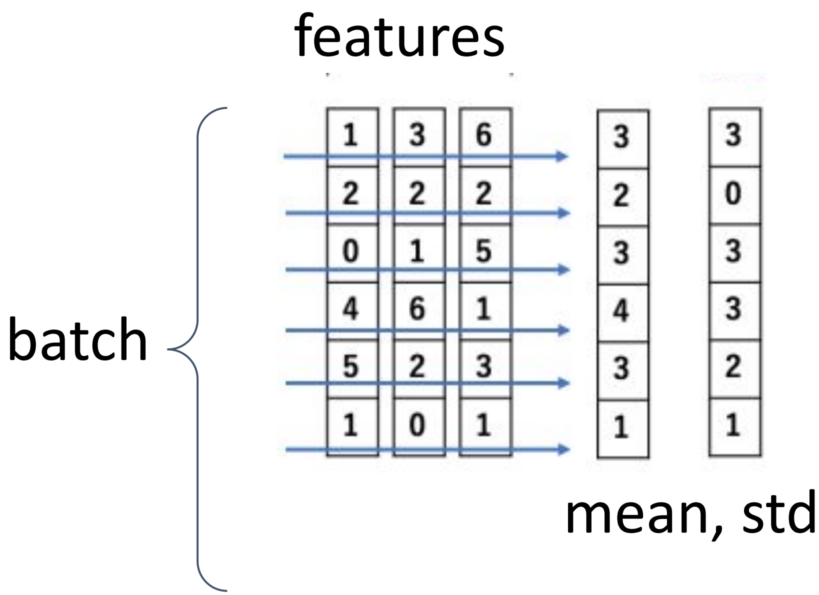
What other forms of normalizations does this offer?



What other forms of normalizations can you think of? Select all that apply

Layer Normalization in 2D

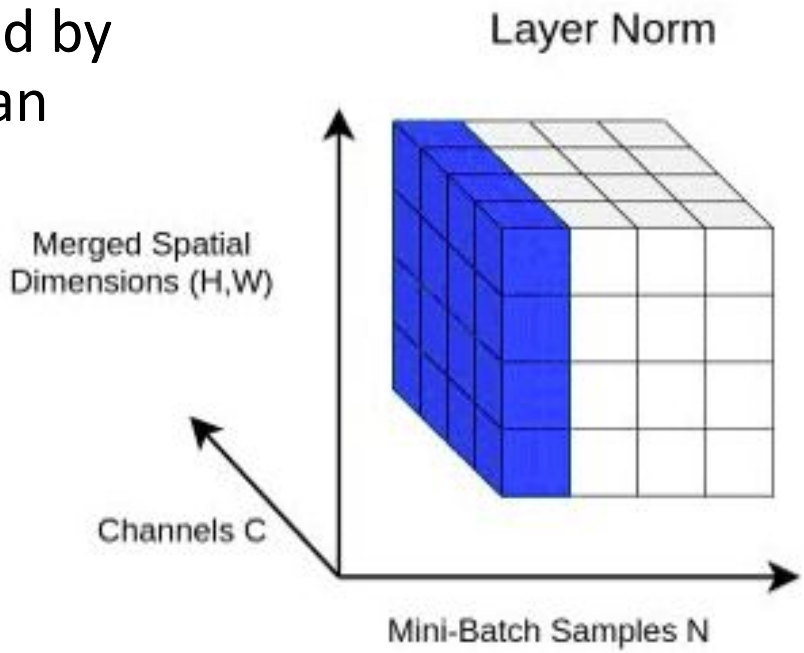
- Used for feature dimension for a single sample



- Same mean and variance for all features

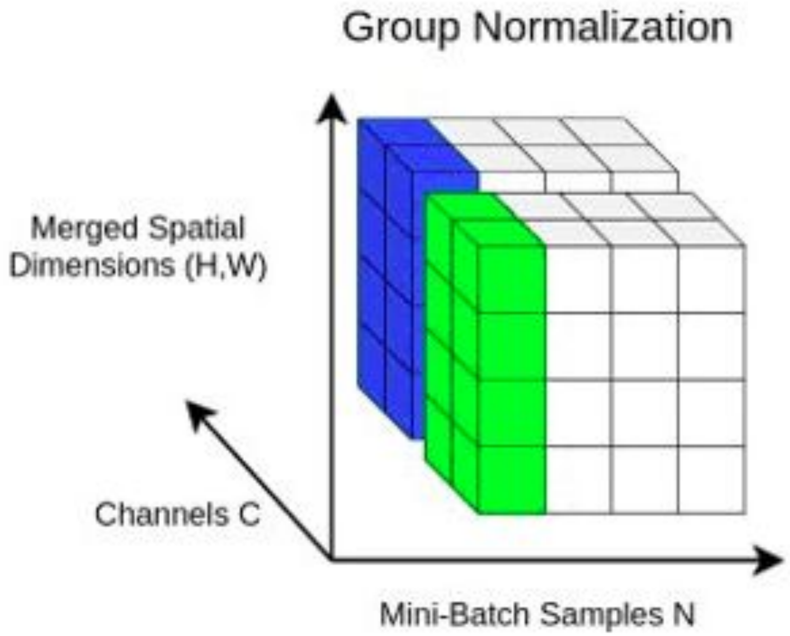
Layer norm in 3D

The pixels in blue are normalized by the same mean and variance



Group norm in 3D

The pixels in blue and green are normalized by the same mean and variance



Summary so far

1. Network regularization
   a. Dropout
   b. Batch normalization
   c. Layer norm
   d. Group norm
2. Data Augmentation
3. Convolutional Neural Networks

Today

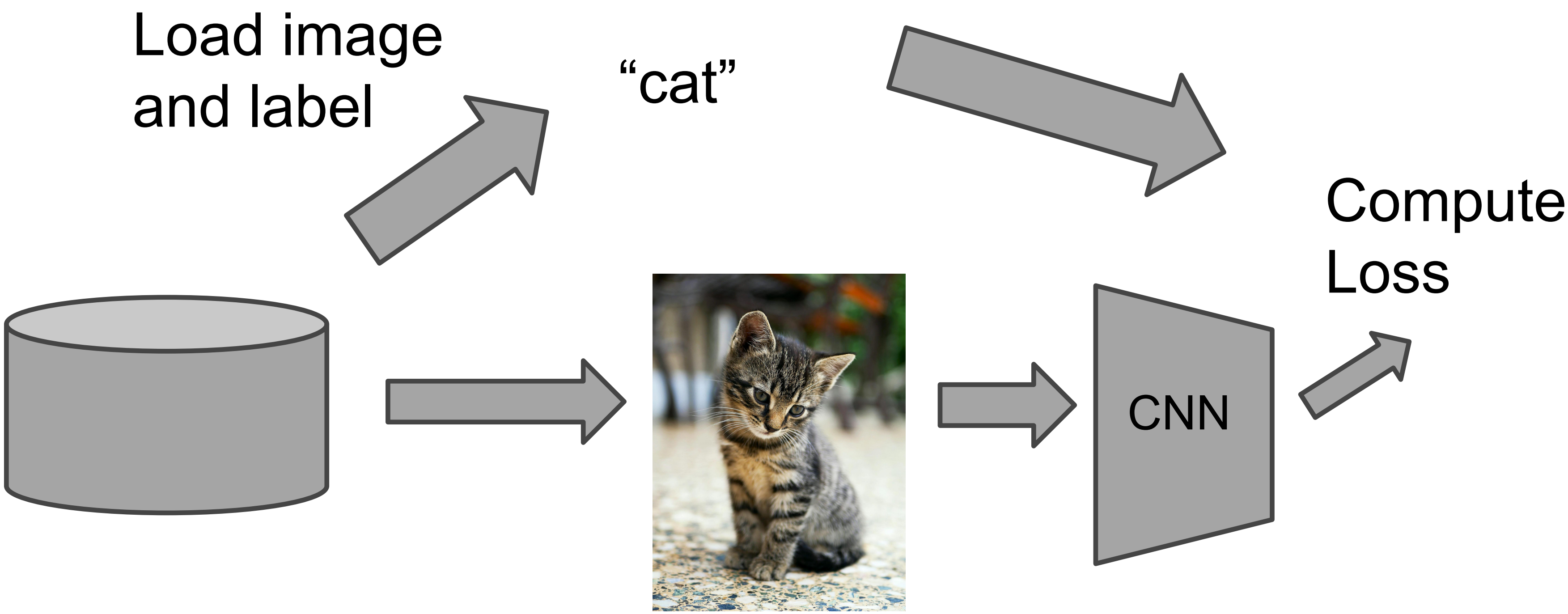
1. Network regularization
   a. Dropout
   b. Batch normalization
   c. Layer norm
   d. Group norm
2. Data Augmentation
3. Convolutional Neural Networks

Data Augmentation:

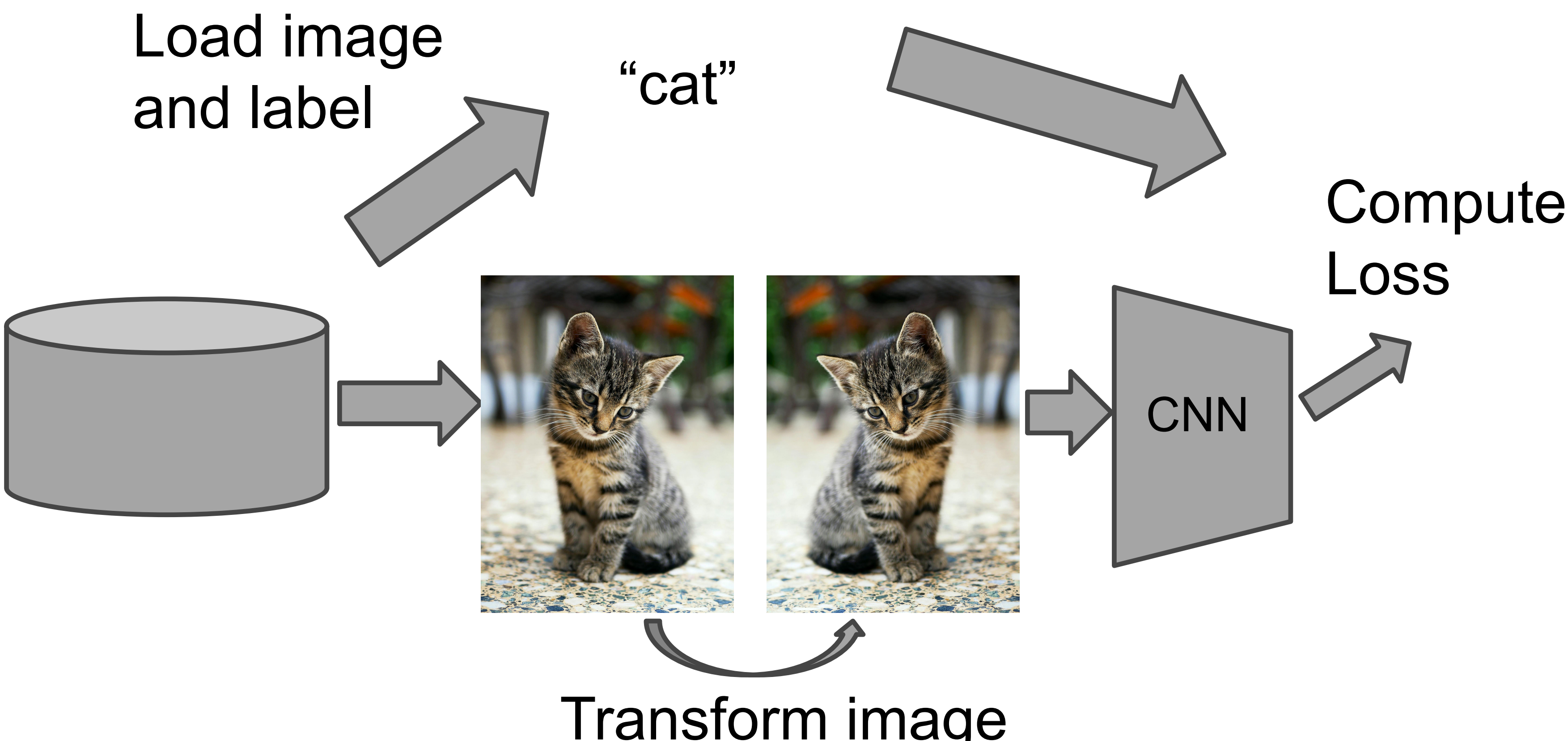
- Creates new samples from existing datasets
- Generally, more data = better performance
- **Goal:** Alter the data without changing the label

Vertical Flip

Data Augmentation:

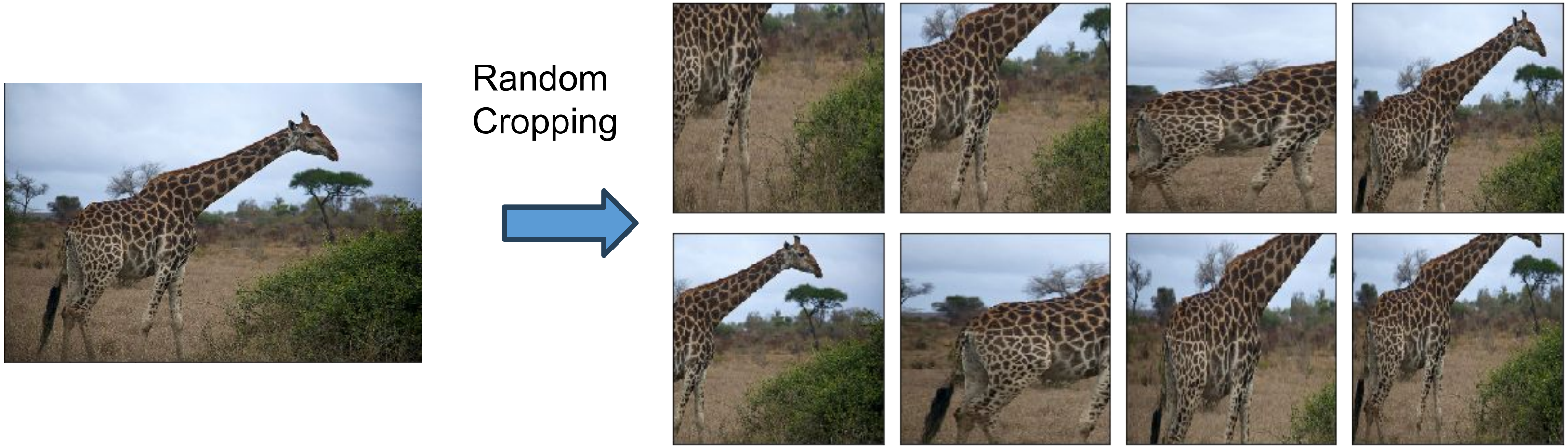


Data Augmentation:



Data Augmentation:

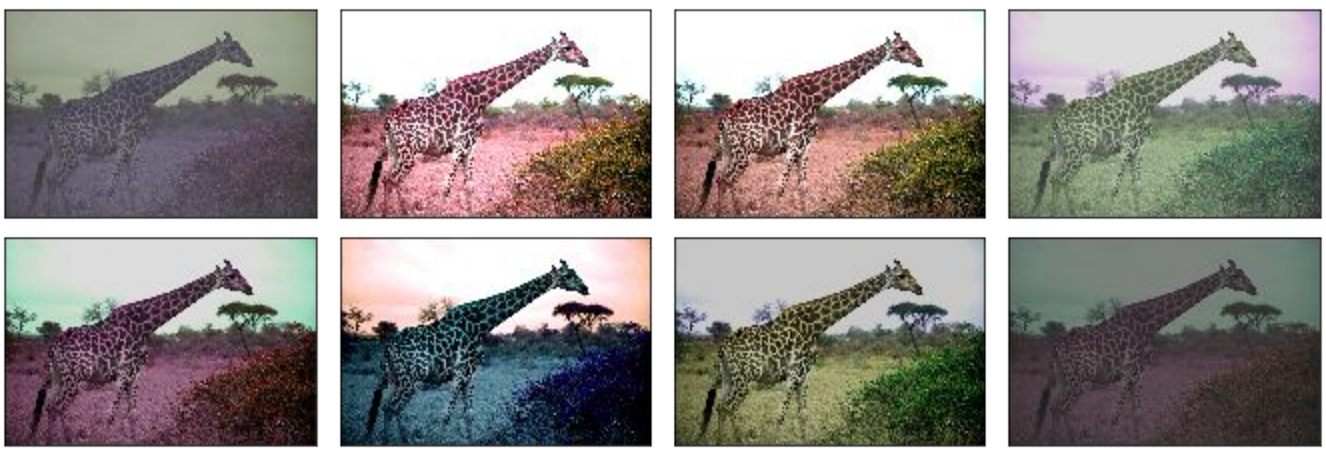
Random Cropping – Sample random crops / scales



Data Augmentation:

Examples of data augmentation:

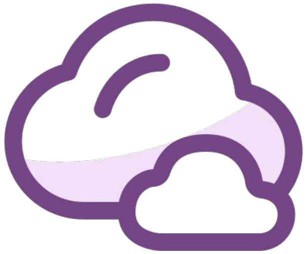
- Translation
- Rotation
- Color Jittering (randomize brightness, contrast, hue etc,)
- Stretching



Is there a benefit of data augmentation using image transformations? Select all that apply

Video data augmentations?

- 1-2 min ⏰
  - Enter in slido in the next slide.
- Do not include the spatial transformations (color jittering etc.) we discussed.



Types of video augmentations?

Types of video augmentations?