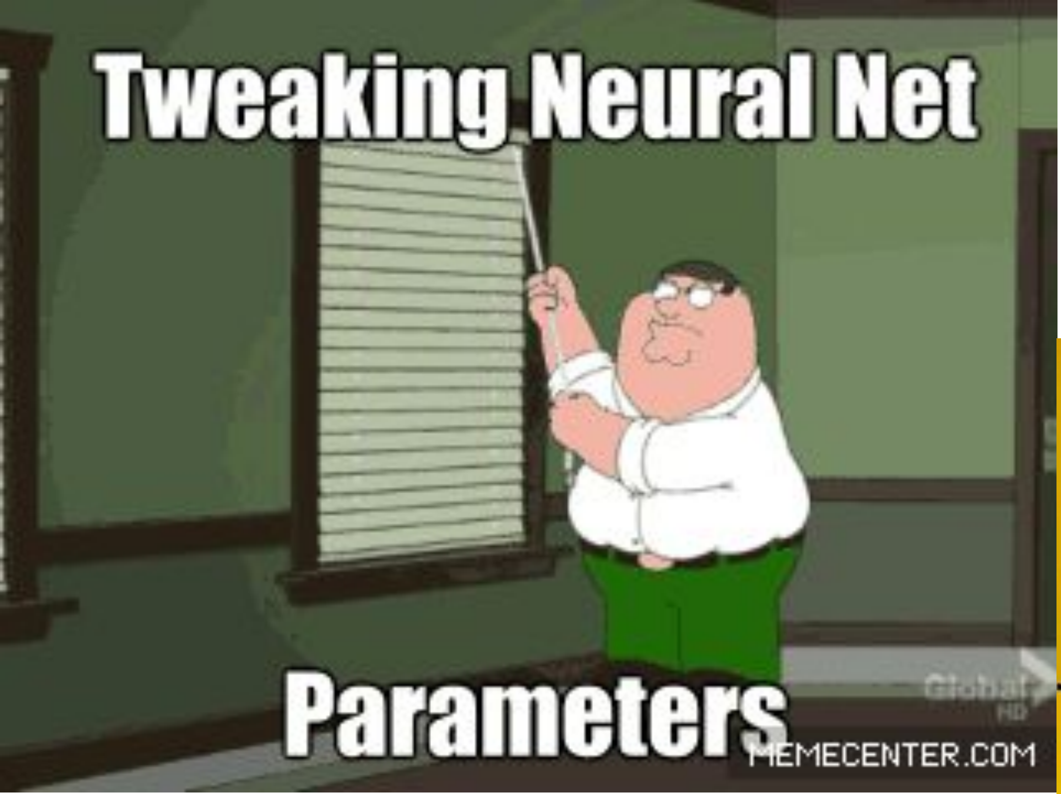
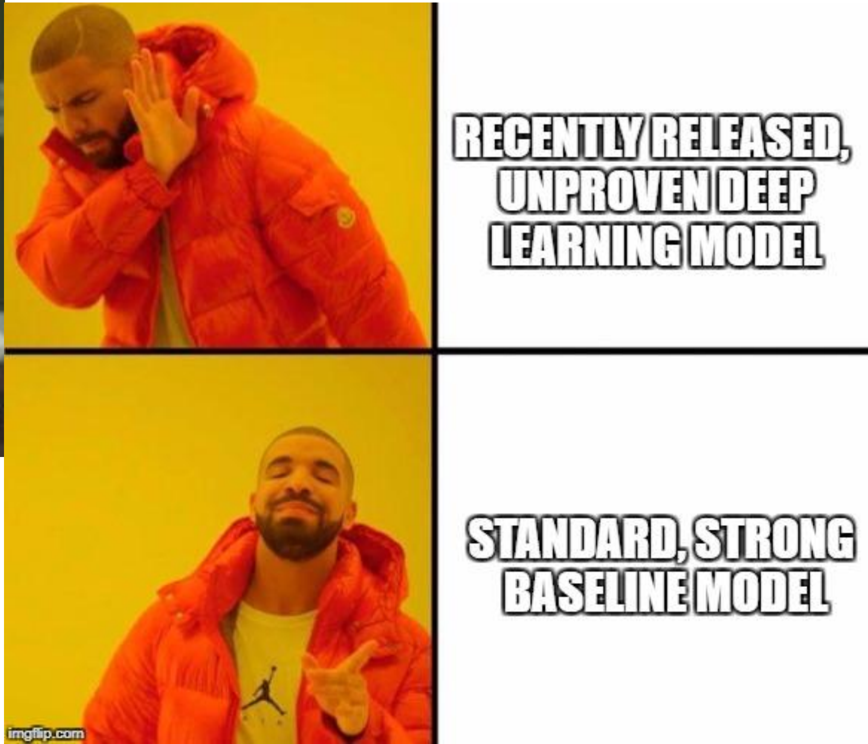

Tweaking Neural Net
Parameters
MEMECENTER.COM


RECENTLY RELEASED, UNPROVEN DEEP LEARNING MODEL
STANDARD, STRONG BASELINE MODEL
imgflip.com

CNNs – General Layout

- One-time setup
  - **Architecture**
  - Activation functions (sigmoid, ReLU, …)
  - Regularization (batch norm, dropout)
- Training
  - Data collection: Preprocessing, Augmentation
  - Loss function: cross entropy loss…
  - Training via SGD (update rules)

The Humble Beginnings

LeNet-5

State-of-the-art performance on MNIST digit recognition (< 1%)

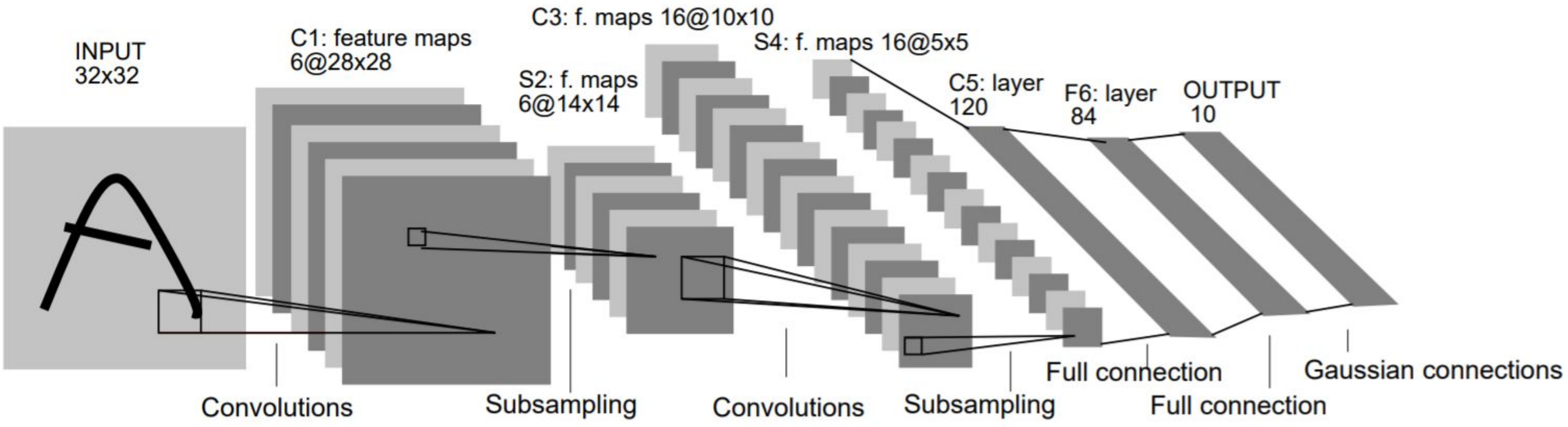
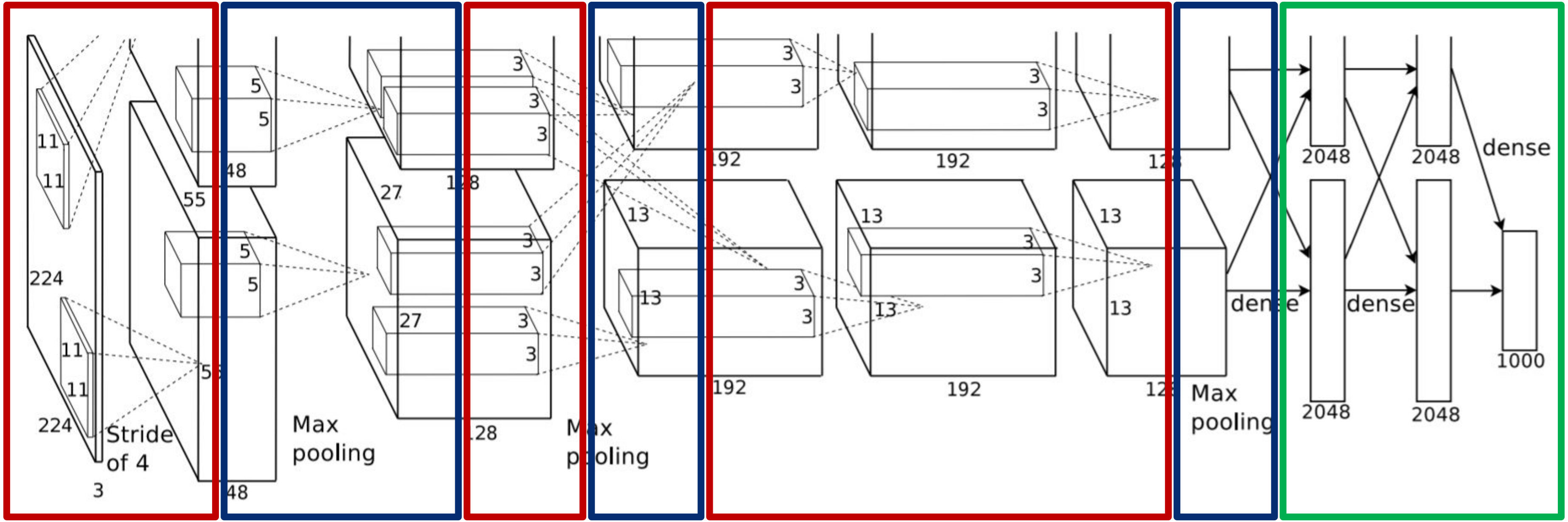


Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.



Conv 1	MaxPool 1	Conv 2	MaxPool 2	Conv 3,4,5			MaxPool 3	FCN 6,7,8		
96 11x11 Stride 4 Pad 0	3x3 Stride 2	256 5x5 Stride 2 Pad 2	3x3 Stride 2	384 3x3 Stride 1 Pad 1	384 3x3 Stride 1 Pad 1	256 3x3 Stride 1 Pad 1	3x3 Stride 2	4096	4096	1000
55x55x96	27x27x96	27x27x256	13x13x256	13x13x384	13x13x384	13x13x256	6x6x256	4096	4096	1000

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. NeurIPS (pp. 1097-1105).

VGG

Q: Why use smaller filter?

Let's do the other way around: Why user larger filters?

→ Receptive field!

Stack of 3 3x3 stride 1 convolutional layers has **same effective receptive field** as 7x7 layer!

But deeper → **More non-linearities**

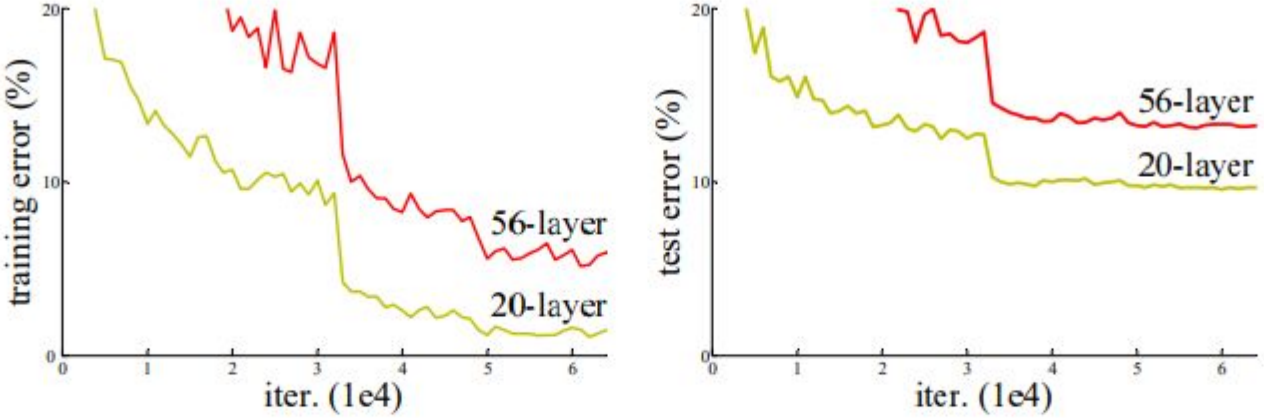
Fewer parameters!

$3*(3^2\ C)$ vs $7^2\ C$ for C channels per layer

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

ResNet

Intuition: Deeper network should perform at least as well as a shallower version. When implemented naively, does this intuition hold?

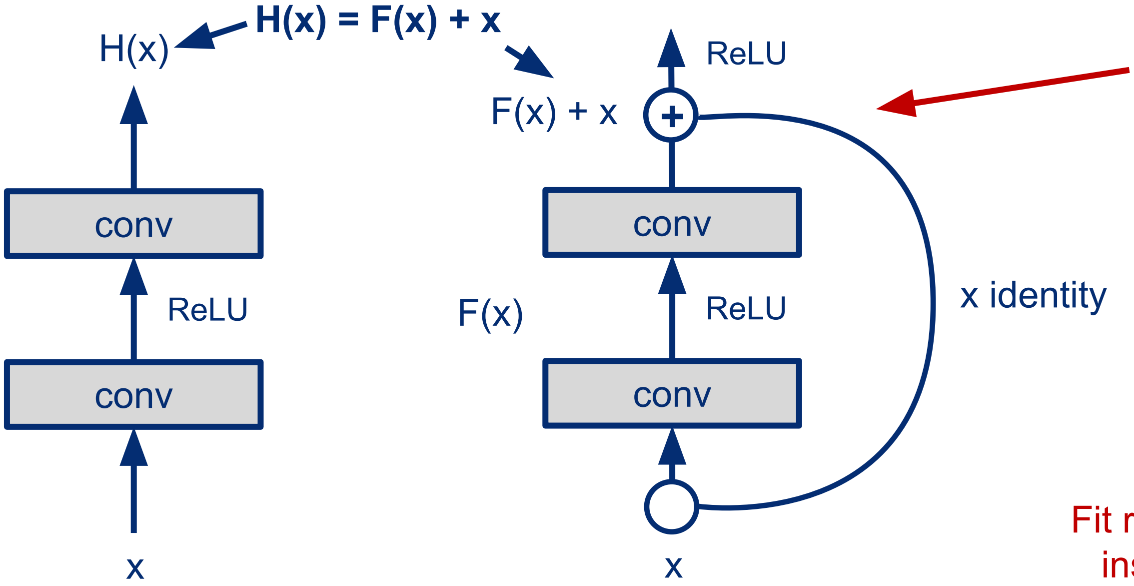


He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognitionIEEE conference on computer vision and pattern recognition (pp. 770-778).

ResNet

Solution

Use network layers to fit residual mapping (rather than desired mapping directly)



“Plain” conv layers

Residual block

Q: What about gradients at addition nodes?

- Add: Distribute gradient!
- Gradient flows unhindered
- “Gradient highway”

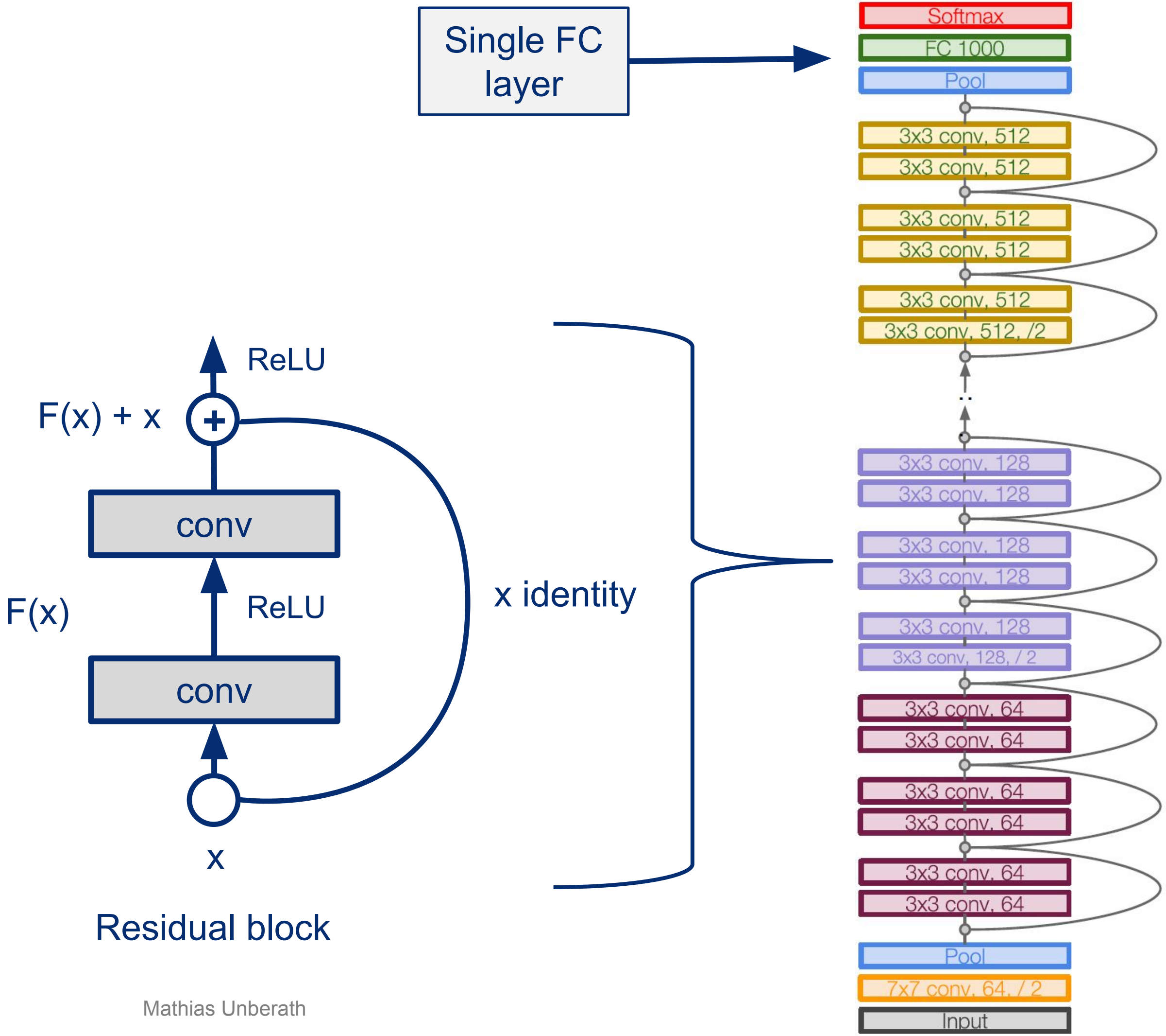
□ Makes training easier!

Fit residual F(x) = H(x) – x instead of H(x) directly

ResNet

Full architecture

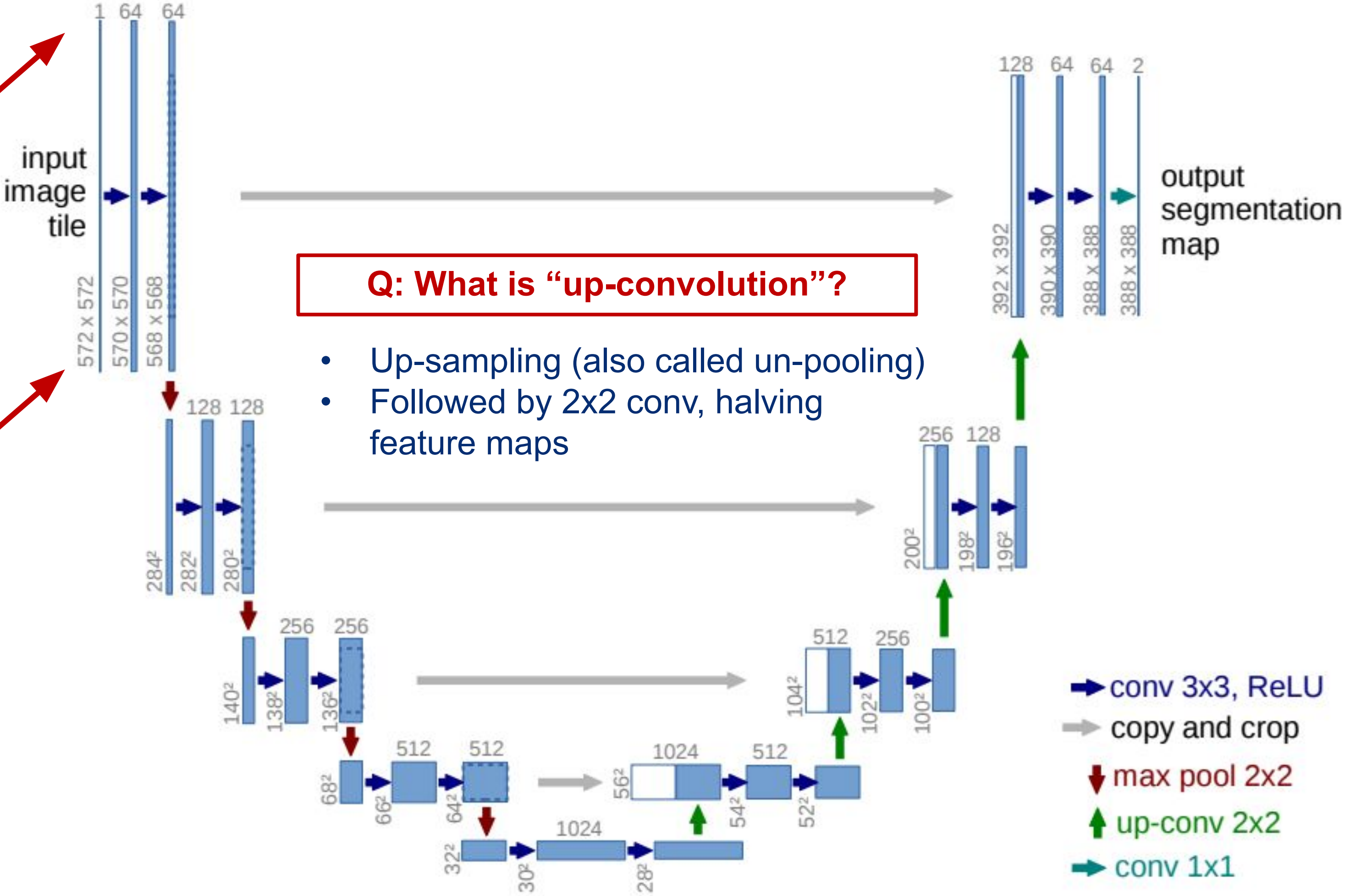
- Stack residual blocks
- Residual block has two 3x3 conv layers
- Periodically, double # filters and down-sample with stride 2
- Additional conv layer at beginning
- Only one FC layer



Residual block

feature maps

Size



Q: What is “up-convolution”?

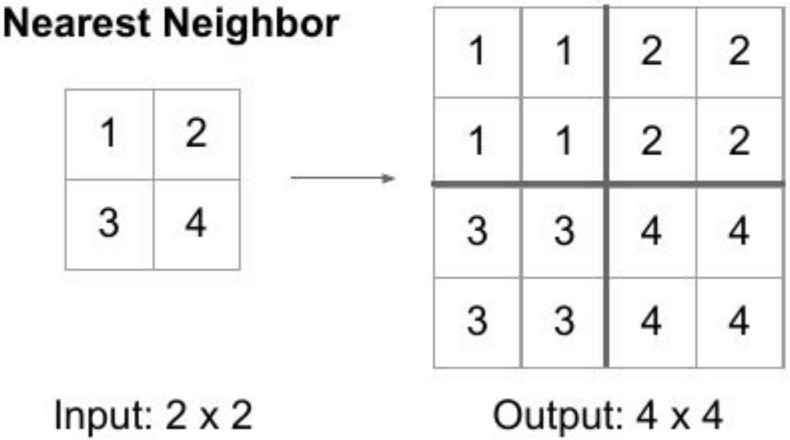
- Up-sampling (also called un-pooling)
- Followed by 2x2 conv, halving feature maps

Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. MICCAI (pp. 234-241). Springer, Cham.

Upsampling

Goal: Increase the spatial dimensions of your matrix

Methods (of interpolation): Nearest neighbor,

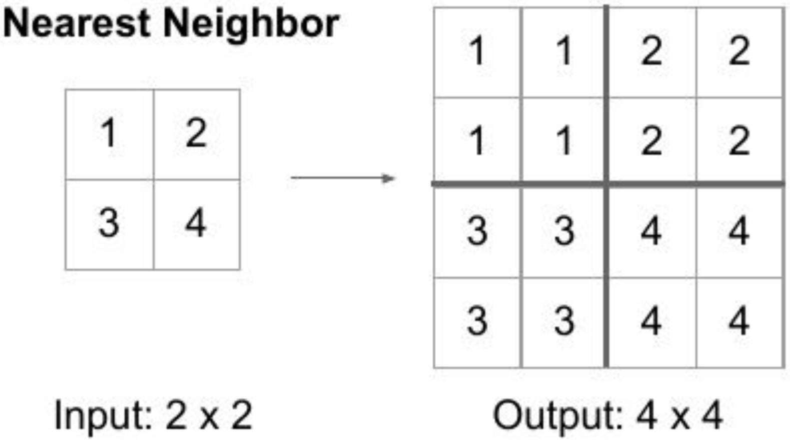


- **mode** (*str, optional*) – the upsampling algorithm: one of `'nearest'`, `'linear'`, `'bilinear'`, `'bicubic'` and `'trilinear'`. Default: `'nearest'`

Upsampling

Goal: Increase the spatial dimensions of your matrix

Methods (of interpolation): Nearest neighbor,



When would you want to use which?

- **mode** (*str*, *optional*) – the upsampling algorithm: one of `'nearest'`, `'linear'`, `'bilinear'`, `'bicubic'` and `'trilinear'`. Default: `'nearest'`

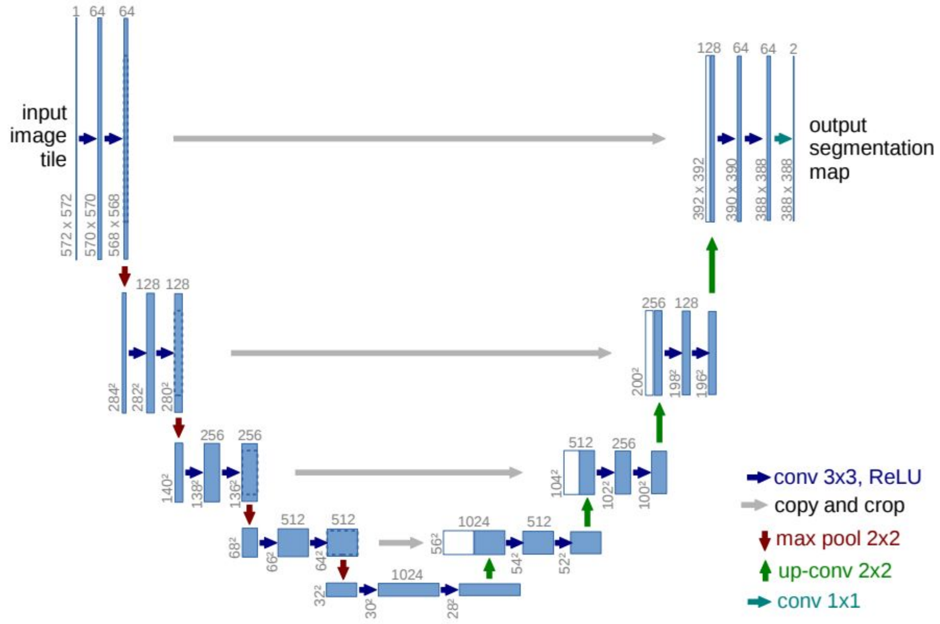
Upsampling

How many parameters does this have?

Upsampling

How many parameters does this have? **0**

There's nothing to learn… do we want to learn something?



Transposed Convolution

Basically the same thing as upsampling but… with learnable parameters

There's nothing to learn… do we want to learn something?