RADONC AI CURRICULUM

Introduction to Deep Learning

LECTURE 5: Convolutional Neural Networks

ANDREW Y. K. FOONG, PH.D., WITH SLIDES BY THOMAS TAVOLARA, PH.D.

May 1st 2026

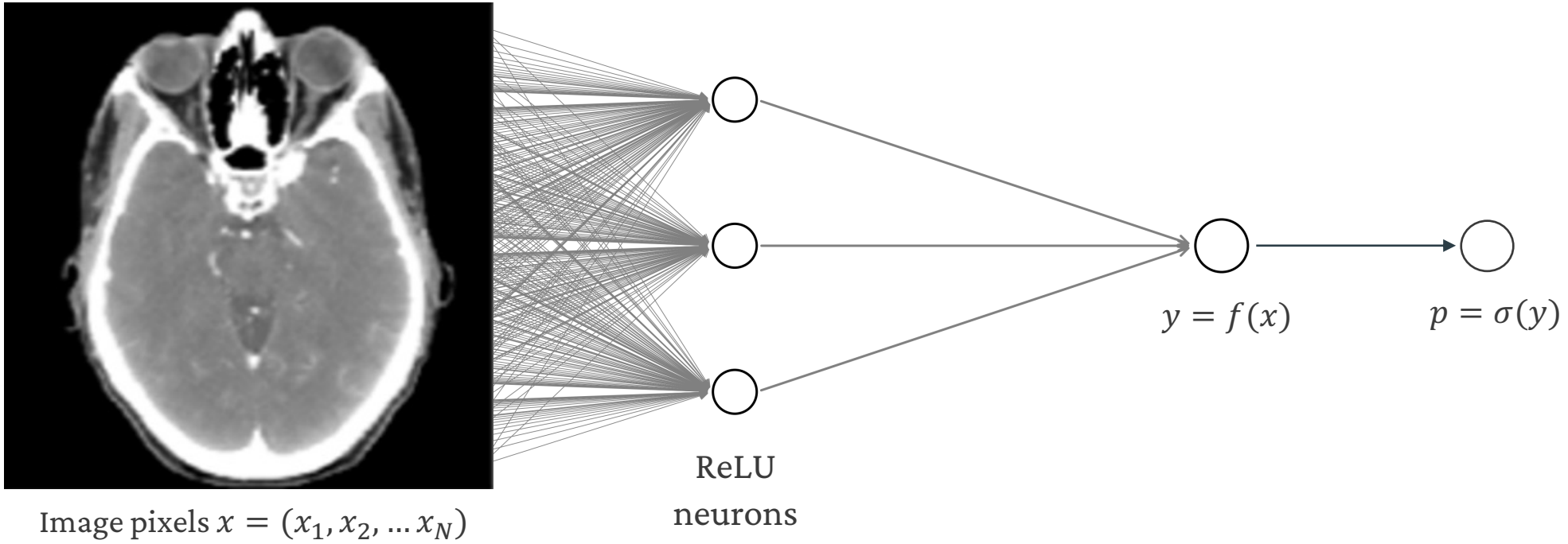
MAYO CLINIC

Radiation Oncology
AI & Data Analytics
AIDA

Today's lecture

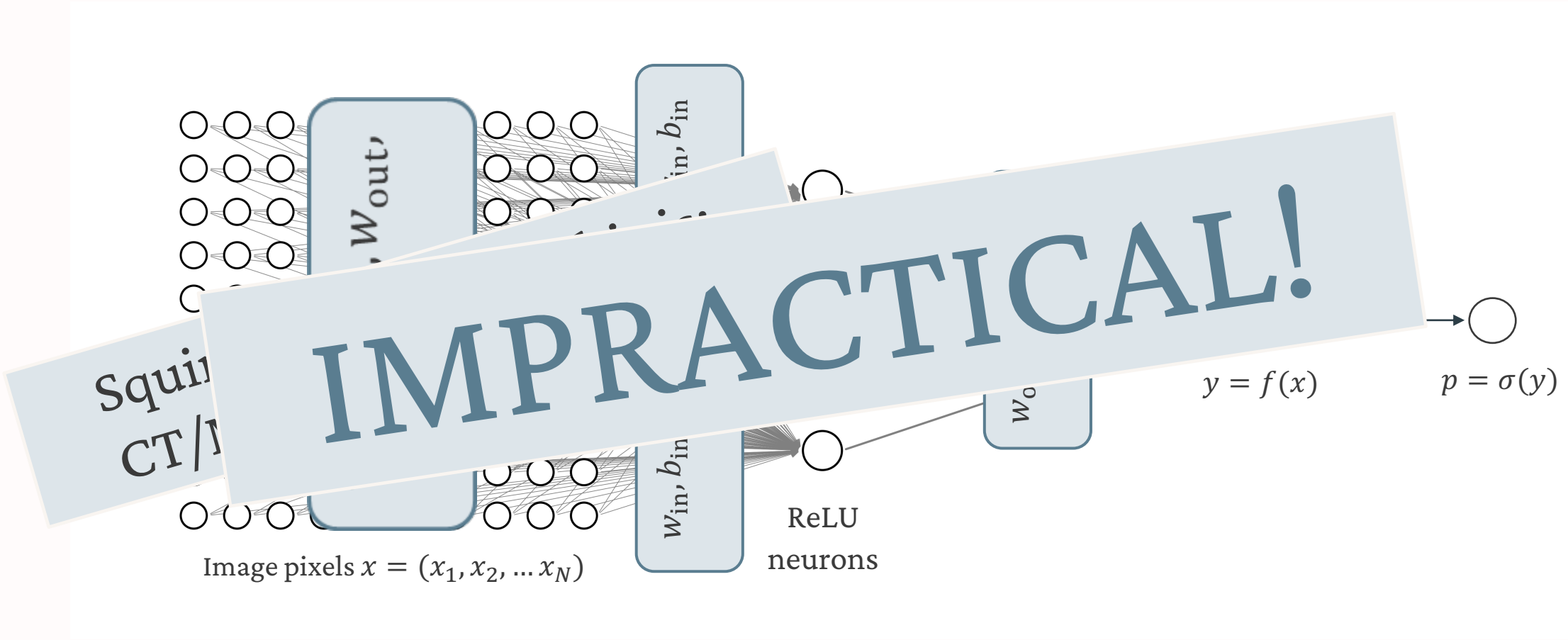
- Why are CNNs needed
  - i.e. why fully connected networks fail
- What convolution means
  - Local receptive fields
  - Feature reuse
  - What is convolution
  - What is a "filter"
  - CNN "layers"
- Why CNNs work better
- Examples

Motivation



Fully-connected neural network

Motivation



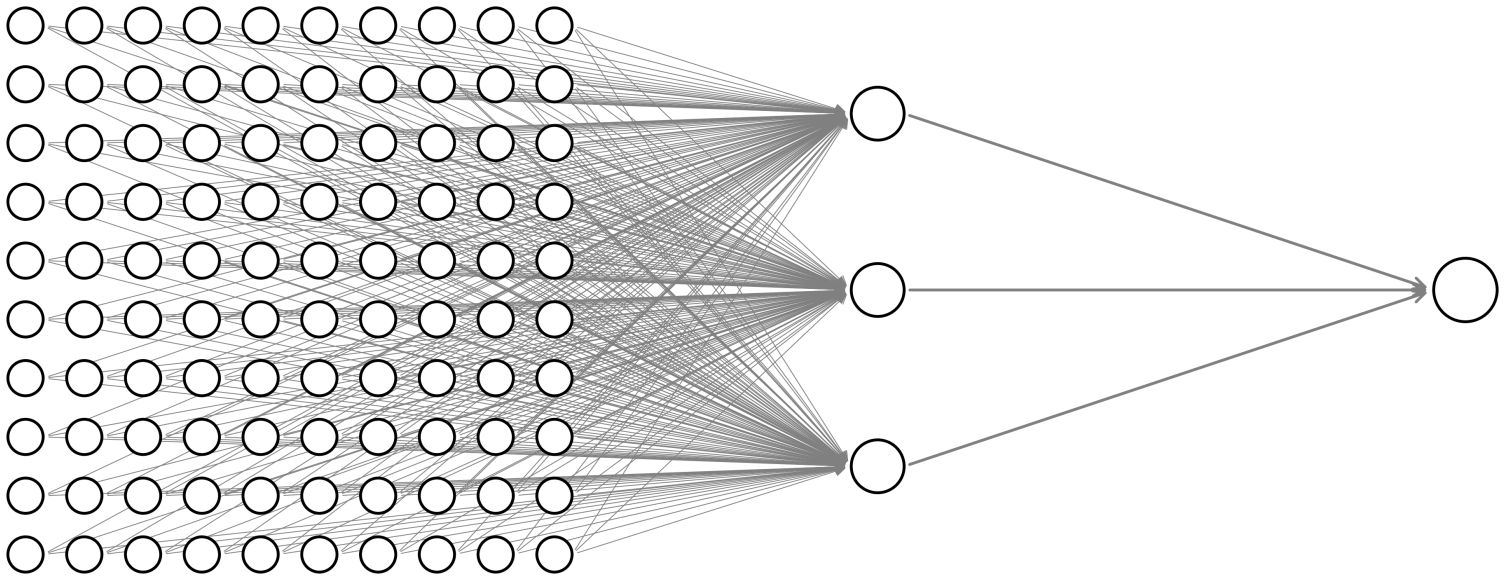
Fully-connected neural network

Fully connected networks fail!

(for "large" images at least...)

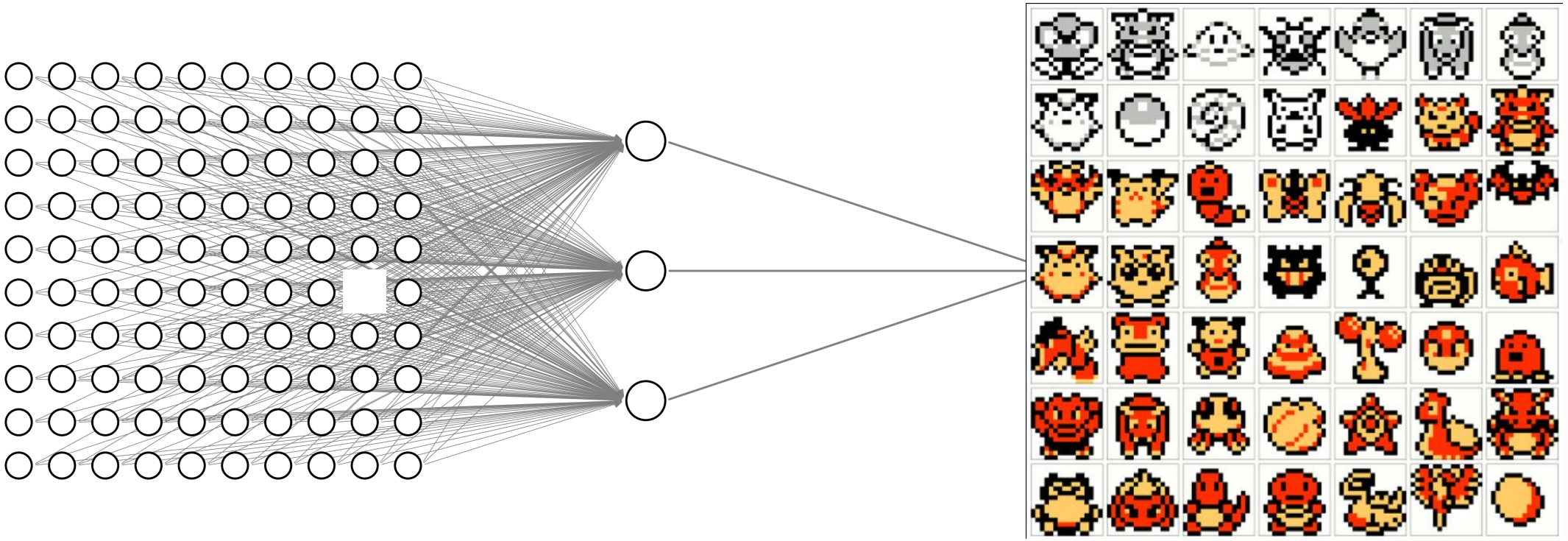
But what if we try anyway?

- Connect every pixel to a neuron
- 10x10 pixels x 3 connections = 300 weights



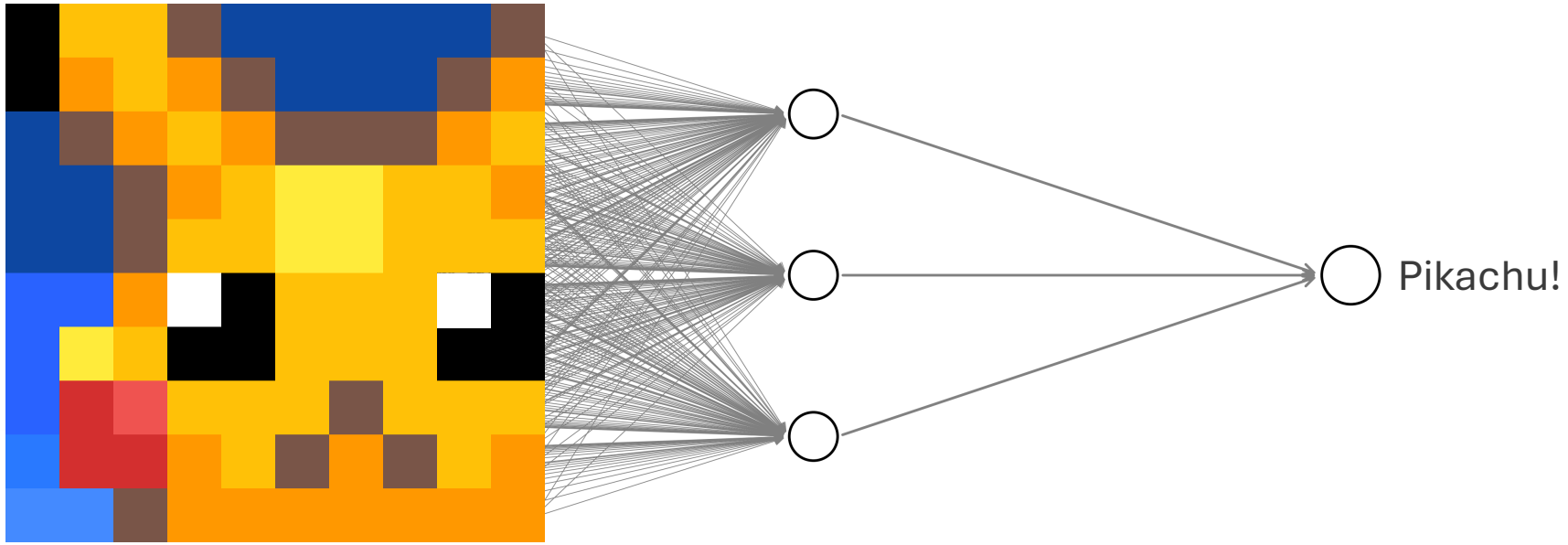
But what if we try anyway?

- Connect every pixel to a neuron
- 10x10 pixels x 3 connections = 300 weights



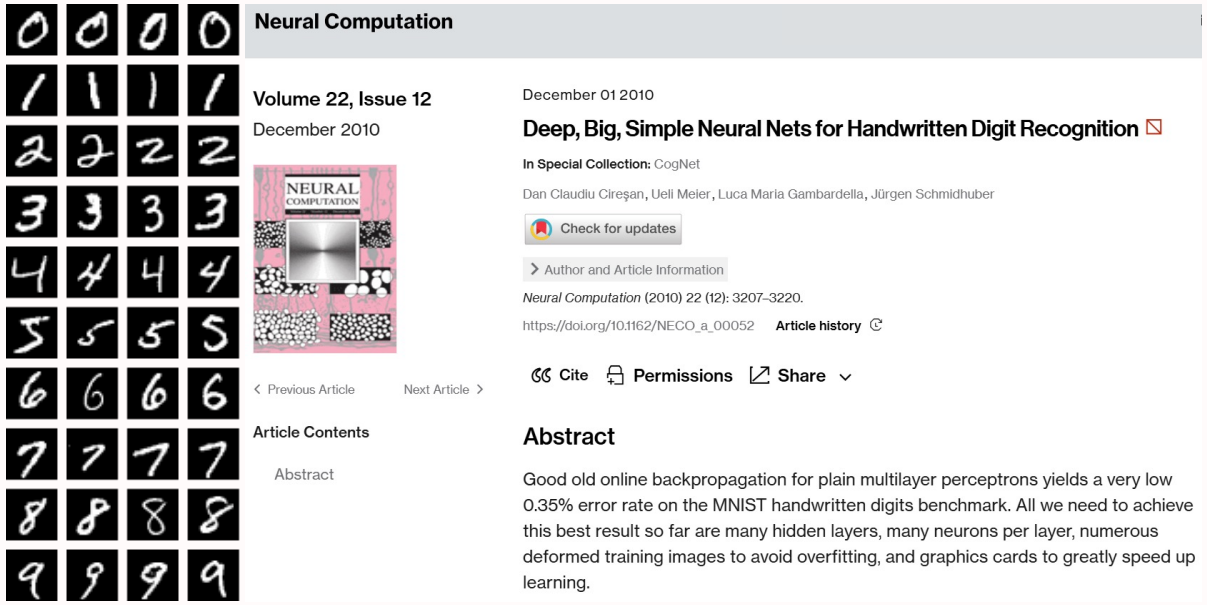
But what if we try anyway?

- Connect every pixel to a neuron
- 10x10 pixels x 3 connections = 300 weights



But what if try anyway?

- MNIST
  - Hand-written digits
  - Initially created by USPS
  - Binarized
    - Pixel values either 0 or 1
  - 60000 training images
  - 10000 testing images



Neural Computation

Volume 22, Issue 12
December 2010

December 01 2010

Deep, Big, Simple Neural Nets for Handwritten Digit Recognition

In Special Collection: CogNet

Dan Claudiu Cireşan, Ueli Meier, Luca Maria Gambardella, Jürgen Schmidhuber

Check for updates

Author and Article Information

Neural Computation (2010) 22 (12): 3207–3220.

https://doi.org/10.1162/NECO_a_00052 **Article history**

Cite Permissions Share

Previous Article Next Article

Article Contents

Abstract

Abstract

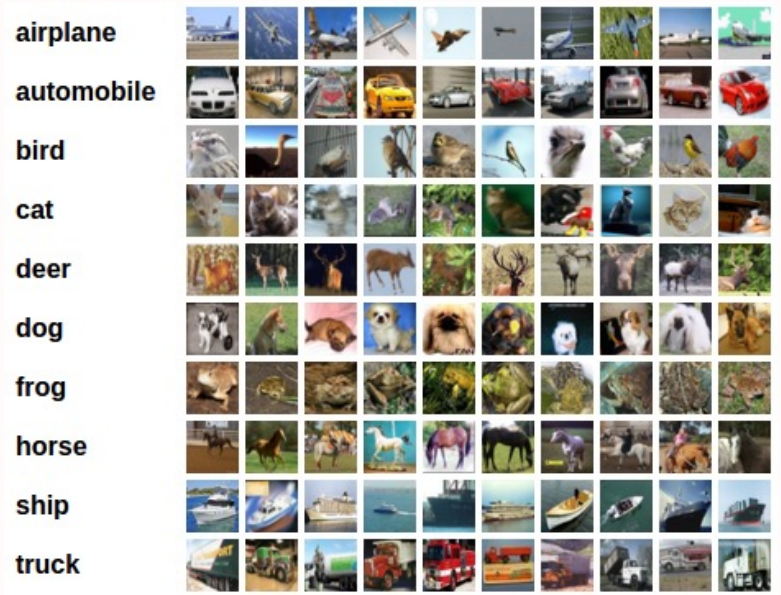
Good old online backpropagation for plain multilayer perceptrons yields a very low 0.35% error rate on the MNIST handwritten digits benchmark. All we need to achieve this best result so far are many hidden layers, many neurons per layer, numerous deformed training images to avoid overfitting, and graphics cards to greatly speed up learning.

- CIFAR-10
  - 32x32 pixels
  - 10 classes (like digits)

Learning Multiple Layers of Features from Tiny Images

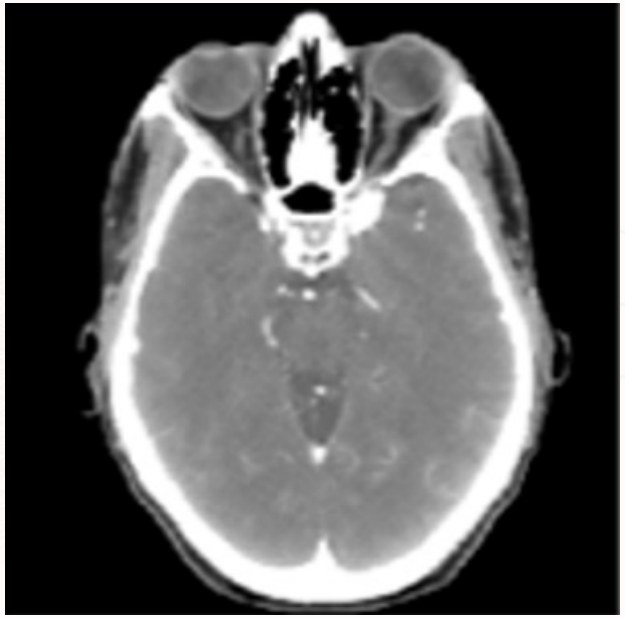
Alex Krizhevsky

April 8, 2009



Reality check

- Real images are much larger
- CT or MRI image size typically 256x256 or 512x512
  - i.e. 65536 pixels or 262144 pixels



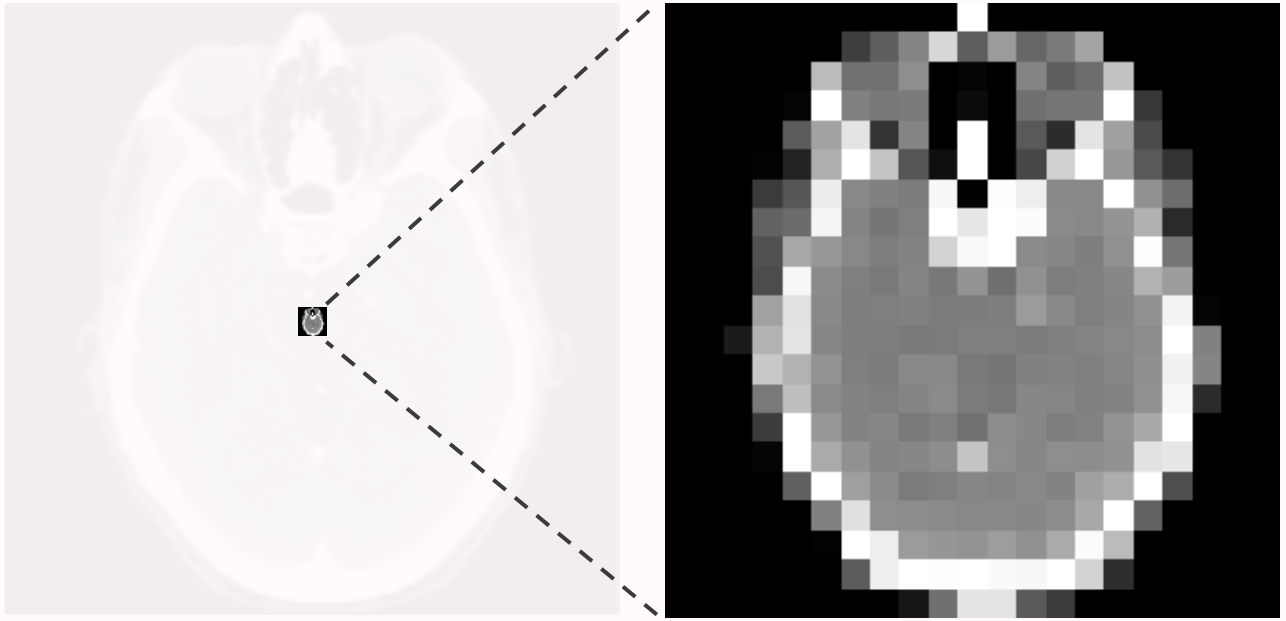
- And that's only one layer

Reality check

- What about just resizing the image?
  - 256x256 -> 16x16
  - Loss of features ∝ resizing
    - i.e. you would keep ~6.25% of features in this example

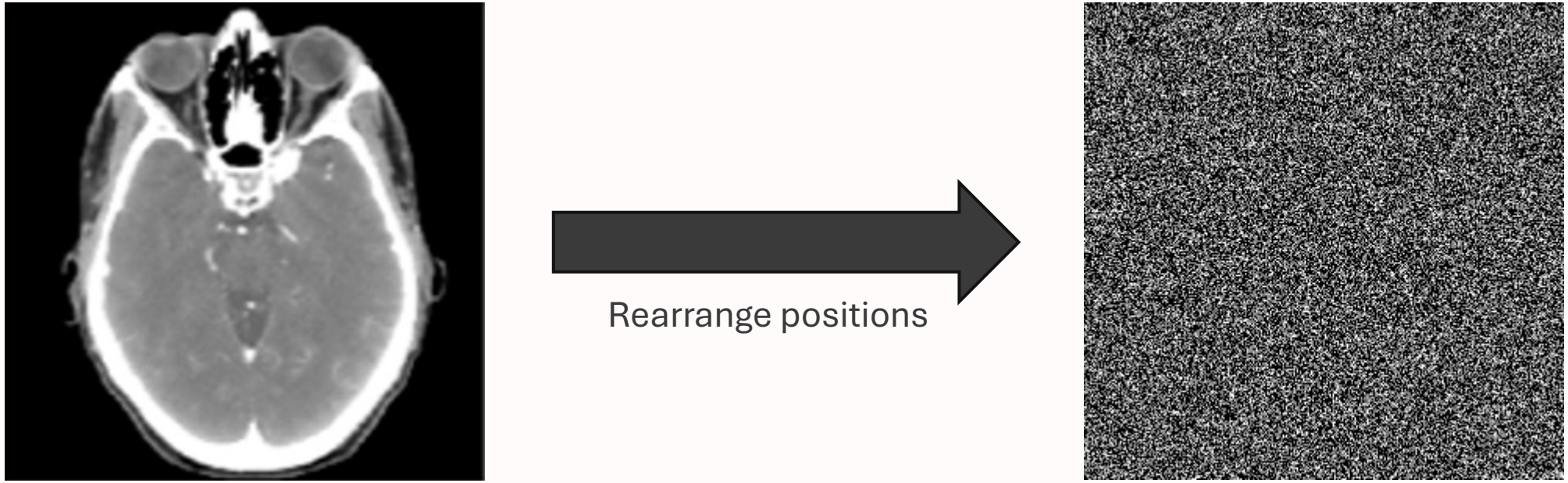
Reality check

- What about just resizing the image?
  - 256x256 -> 16x16
  - Loss of features ∝ resizing
    - i.e. you would keep ~6.25% of features in this example



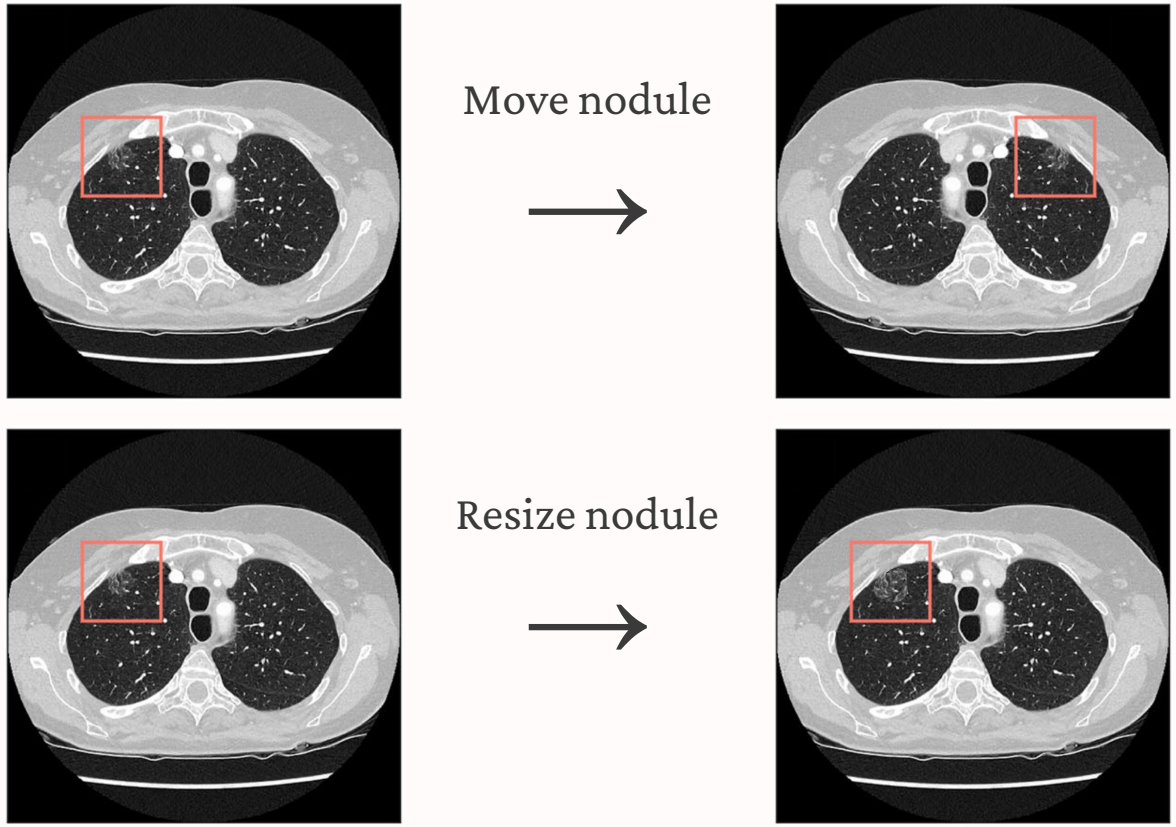
Spatial relationships matter

- Fully connected networks (on images) ignore *spatial* relationships
- One neuron per pixel cannot see structure



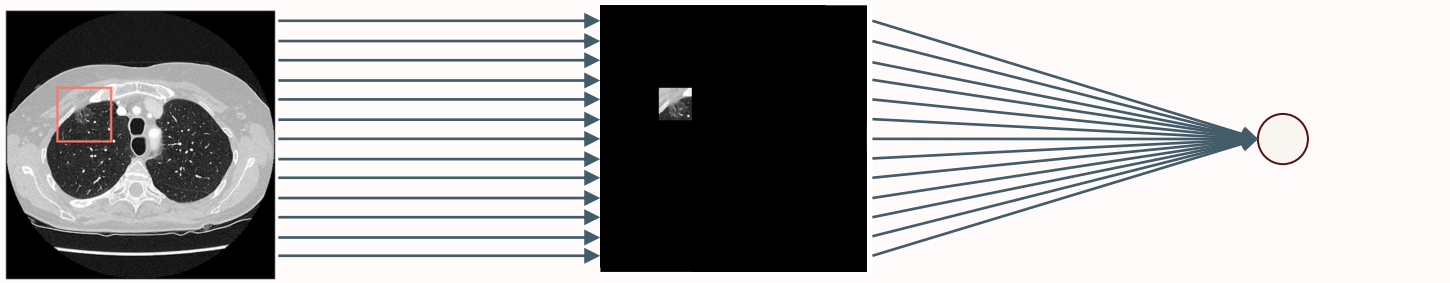
Same feature can be everywhere

- For example, lung nodules
  - Different position
  - Different scale

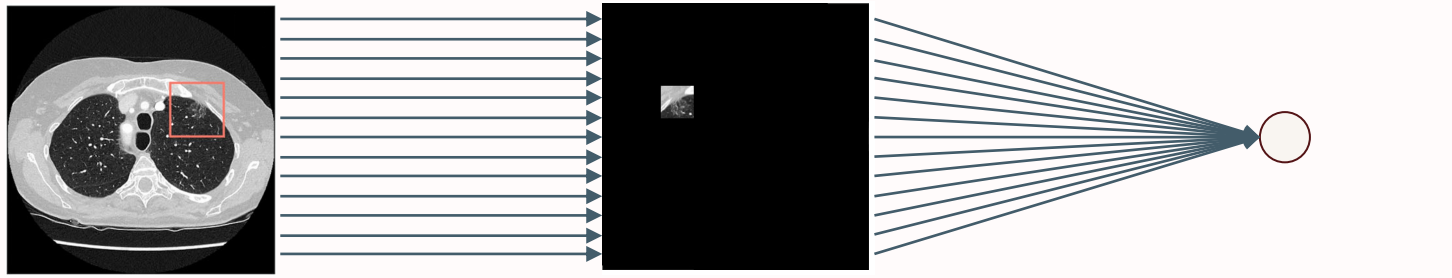


Same feature can be everywhere

- Neuron not activated if nodule in other lung



CT image | Neuron weights | Neuron activated



CT with nodule in other lung | Neuron weights | Neuron inactive

- Fully connected network requires one neuron for each position!

Why fully connected networks fail

- Each neuron must learn what nodules look like independently
  - Examples of left lung nodules tell network nothing about right lung
  - By contrast, humans generalize to new locations easily
- Rule of thumb:
  - More parameters → more data needed to generalize well
  - Collecting and labeling images expensive!
- Same problem in all medical imaging!

Convolutional neural networks (CNNs) to the rescue

Why fully connected networks fail

arXiv > cs > arXiv:1202.2745

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 13 Feb 2012]

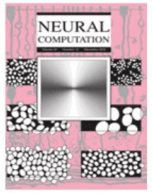
Multi-column Deep Neural Networks for Image Classification

Dan Cireşan, Ueli Meier, Juergen Schmidhuber

Traditional methods of computer vision and machine learning cannot match human performance on tasks such as the recognition of handwritten digits or traffic signs. Our biologically plausible deep artificial neural network architectures can. Small (often minimal) receptive fields of convolutional winner-take-all neurons yield large network depth, resulting in roughly as many sparsely connected neural layers as found in mammals between retina and visual cortex. Only winner neurons are trained. Several deep neural columns become experts on inputs preprocessed in different ways; their predictions are averaged. Graphics cards allow for fast training. On the very competitive MNIST handwriting benchmark, our method is the first to achieve near-human performance. On a traffic sign recognition benchmark it outperforms humans by a factor of two. We also improve the state-of-the-art on a plethora of common image classification benchmarks.

Neural Computation Issues Online Early About Submit Subscribe

Volume 22, Issue 12
December 2010



December 01 2010

Deep, Big, Simple Neural Nets for Handwritten Digit Recognition

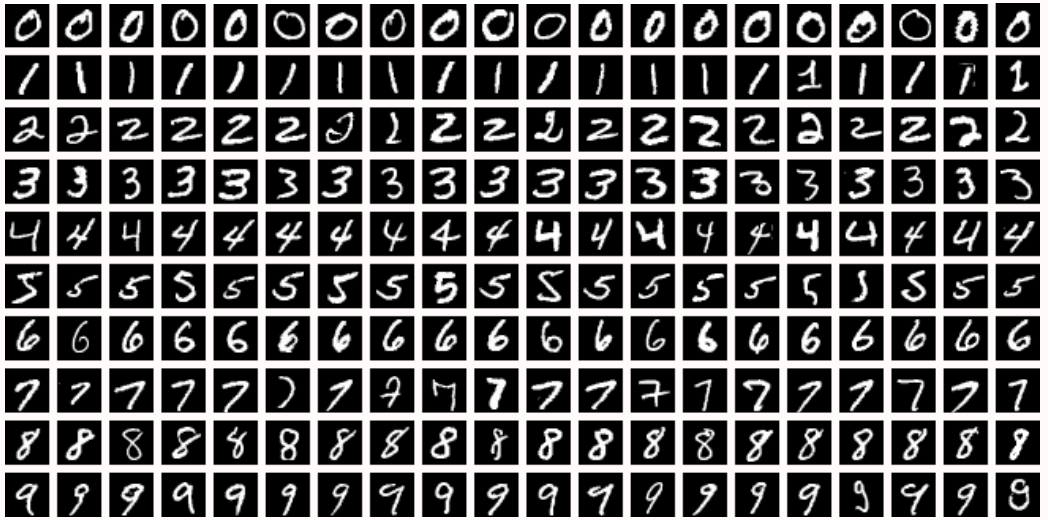
In Special Collection: CogNet

Dan Claudiu Cireşan, Ueli Meier, Luca Maria Gambardella, Jürgen Schmidhuber

Neural Computation (2010) 22 (12): 3207–3220.
https://doi.org/10.1162/NECO_a_00052 **Article history**

Abstract

Good old online backpropagation for plain multilayer perceptrons yields a very low 0.35% error rate on the MNIST handwritten digits benchmark. All we need to achieve this best result so far are many hidden layers, many neurons per layer, numerous deformed training images to avoid overfitting, and graphics cards to greatly speed up learning.



Local patterns

How to reuse features

Local receptive fields

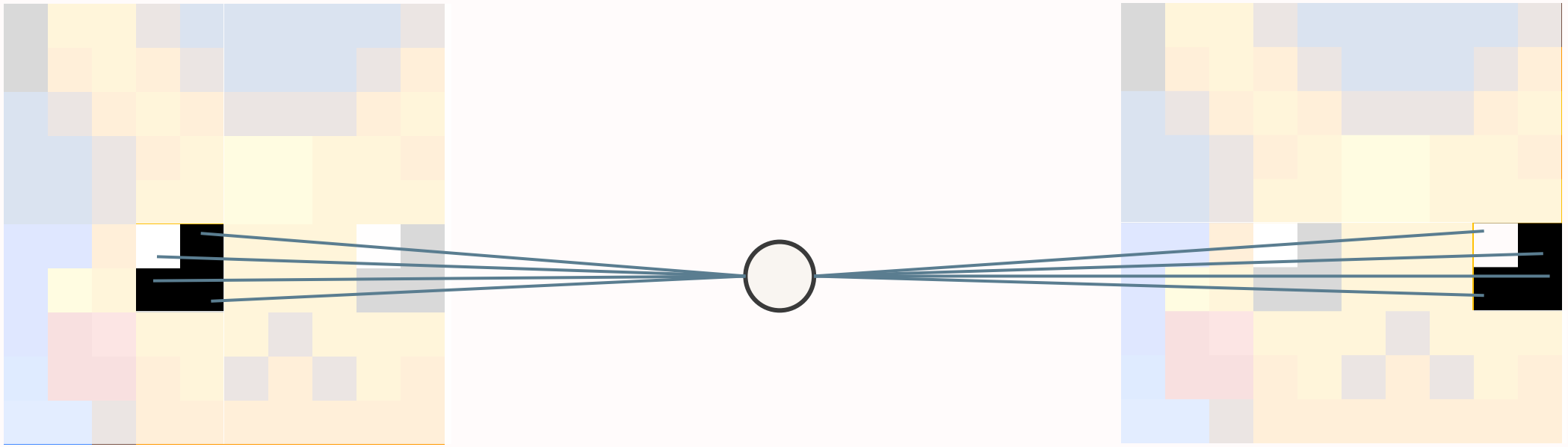
- Receptive field = small region of pixels
  - Typically 3x3 to 7x7 pixels (although can be larger)
- Each weight “looks” at a small region of the image

Local receptive fields

- But wait… now we have 3x3 weights per pixel
- *The trick*
  - We use the same set of 3x3 weights across the whole image!

Feature reuse

- What's useful in one part of the image is useful elsewhere
  - i.e. edges, textures
  - Or simple pixel-y eyes

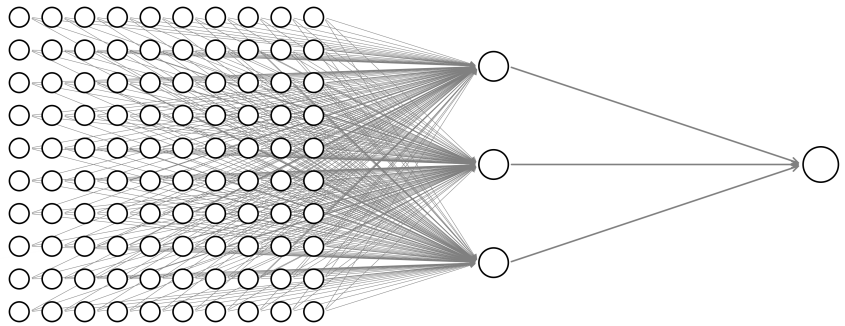


Practical imaging AI

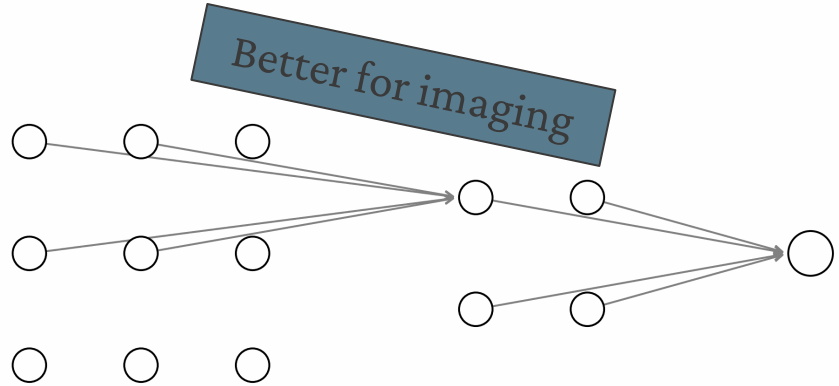
- Practical imaging AI never uses fully-connected networks
  - Needs too many training images

Imaging AI → think convolutional neural networks (CNNs), not fully connected networks

- CNNs type of network architecture:
  - Particular way to connect neurons
  - Especially tailored to imaging tasks



Fully-connected architecture

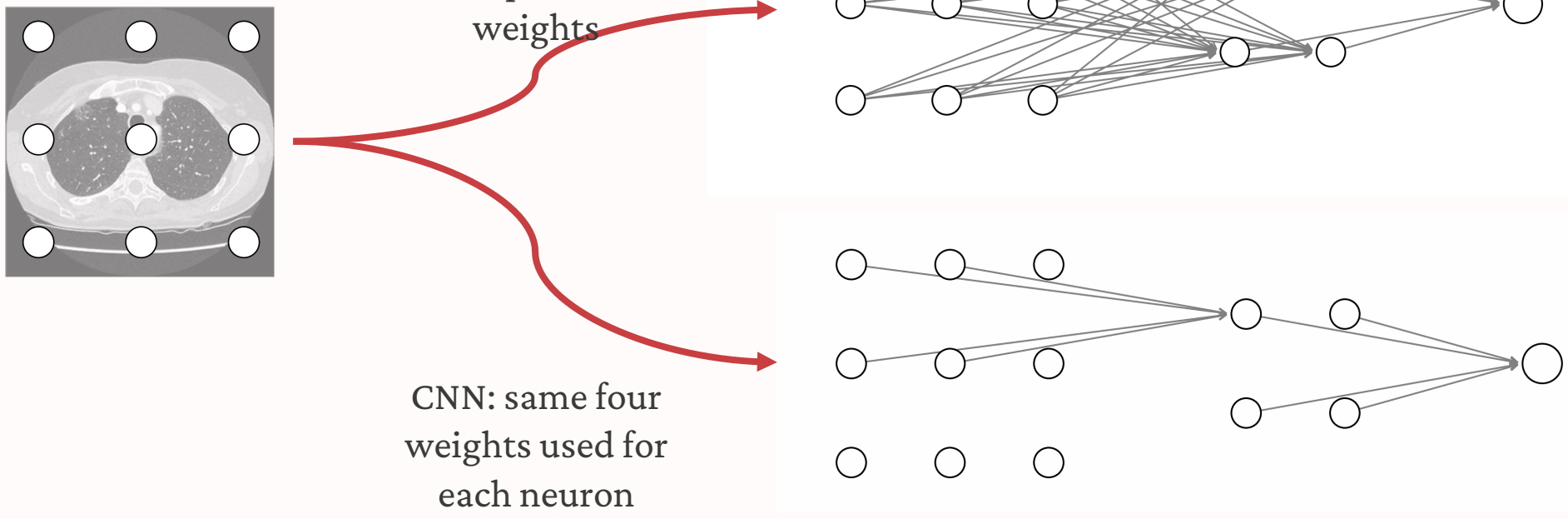


Convolutional neural network (CNN) architecture

Courtesy of Dr. Foong

Convolutional neural networks

- Key idea of CNNs: share weights between different parts of image
- Assumption: features useful in one part also useful in others
- Dragging weights across image: “convolution”



Convolution

CNN operator

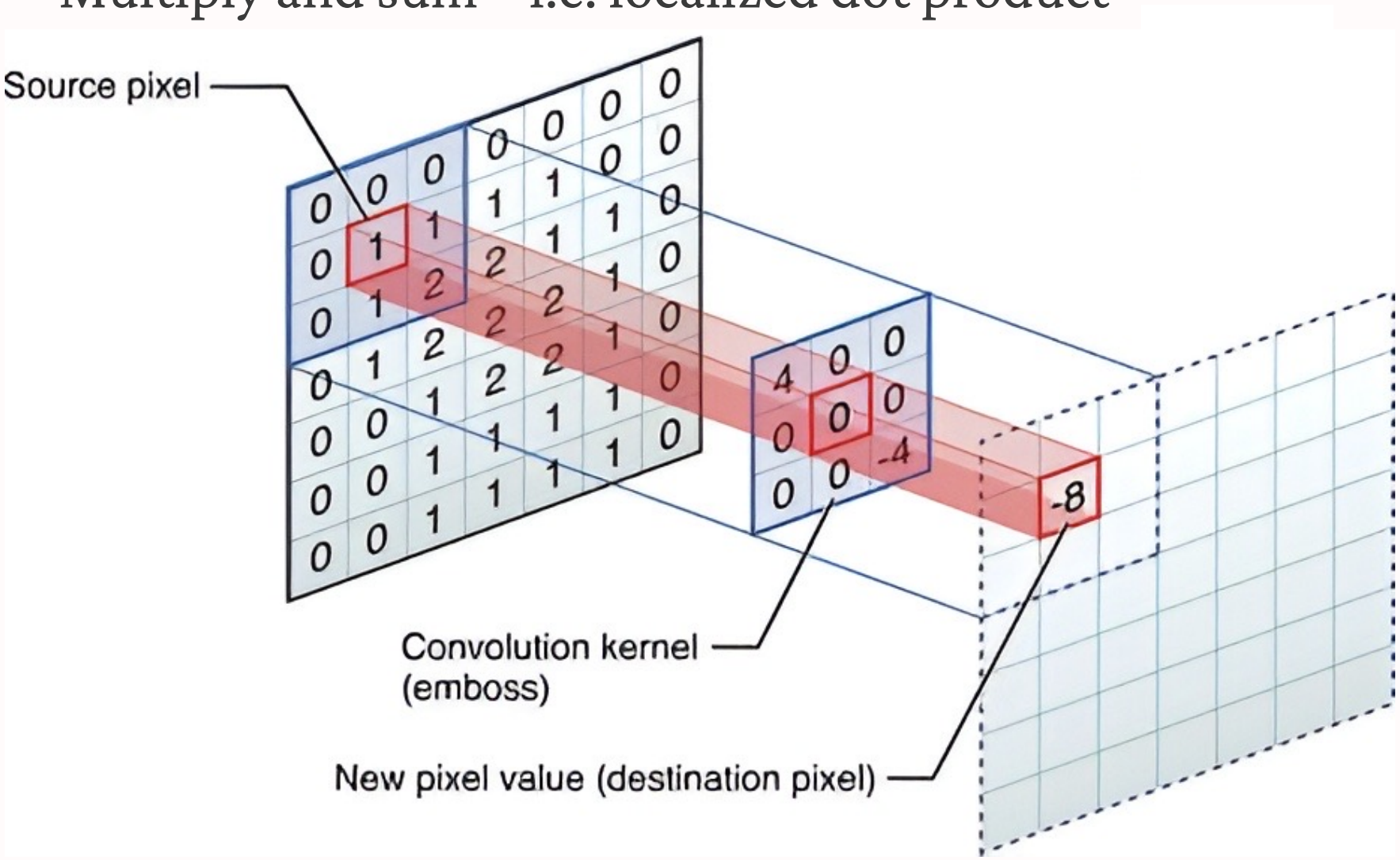
Mathematical intuition

- just kidding

$$(f * g)(t) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} f(\tau)\, g(t-\tau)\, d\tau$$

Mathematical intuition

- Multiply and sum – i.e. localized dot product



(4 x 0)
(0 x 0)
(0 x 0)
(0 x 0)
(0 x 1)
(0 x 1)
(0 x 0)
(0 x 1)
(-4 x 2)

-8

“Practical” example

- Detecting a vertical edge in an image

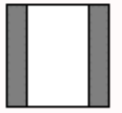
10	10	10	0	0	0
10	10	10	0	0	0
10	10	10	0	0	0
10	10	10	0	0	0
10	10	10	0	0	0
10	10	10	0	0	0

*

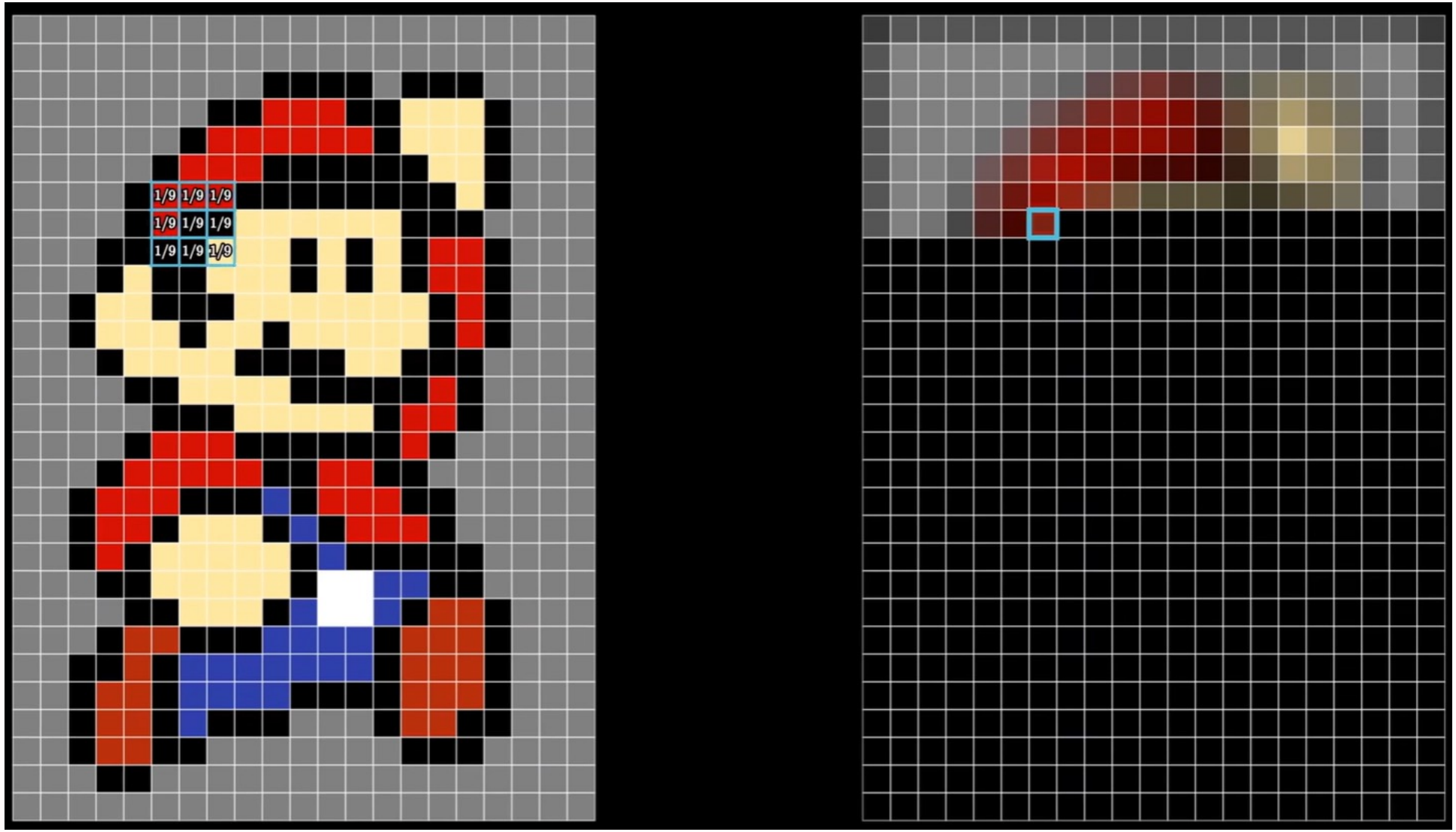
1	0	-1
1	0	-1
1	0	-1

=

0	30	30	0
0	30	30	0
0	30	30	0
0	30	30	0

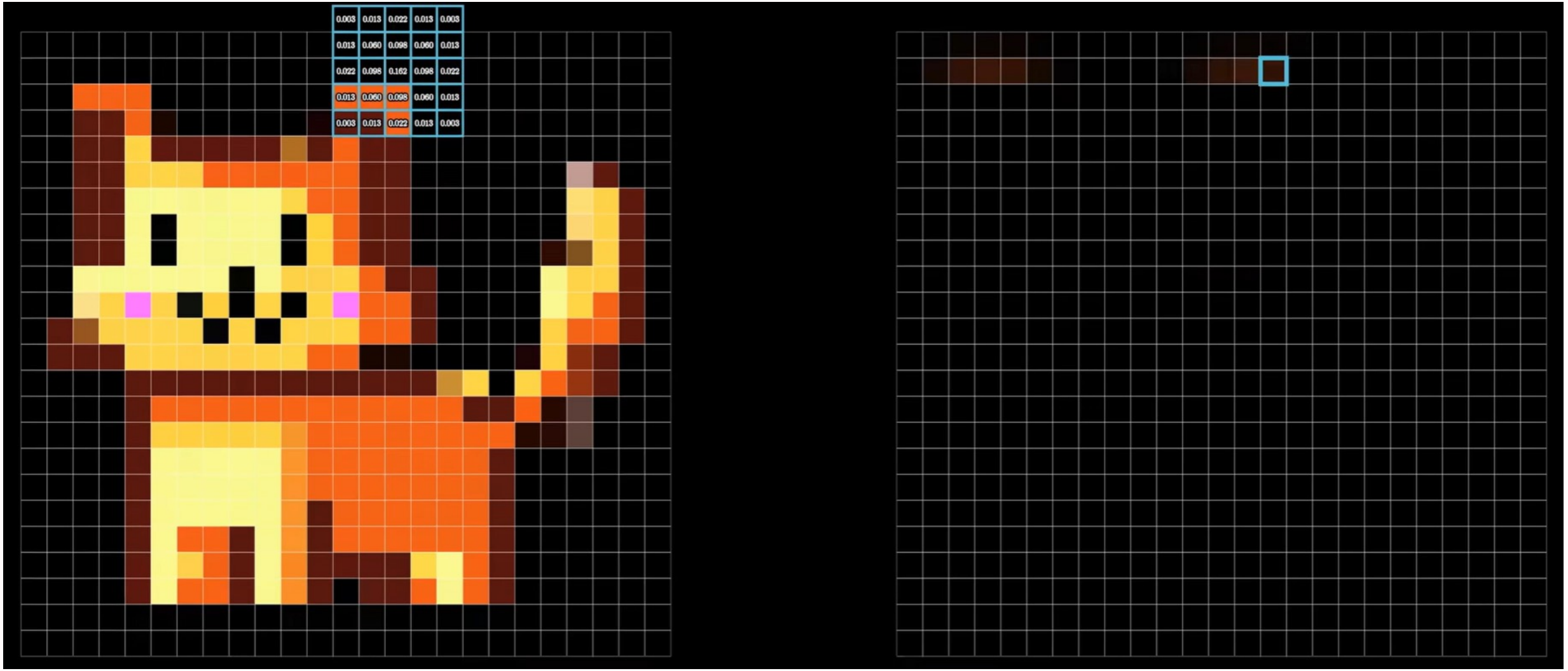


Animated examples



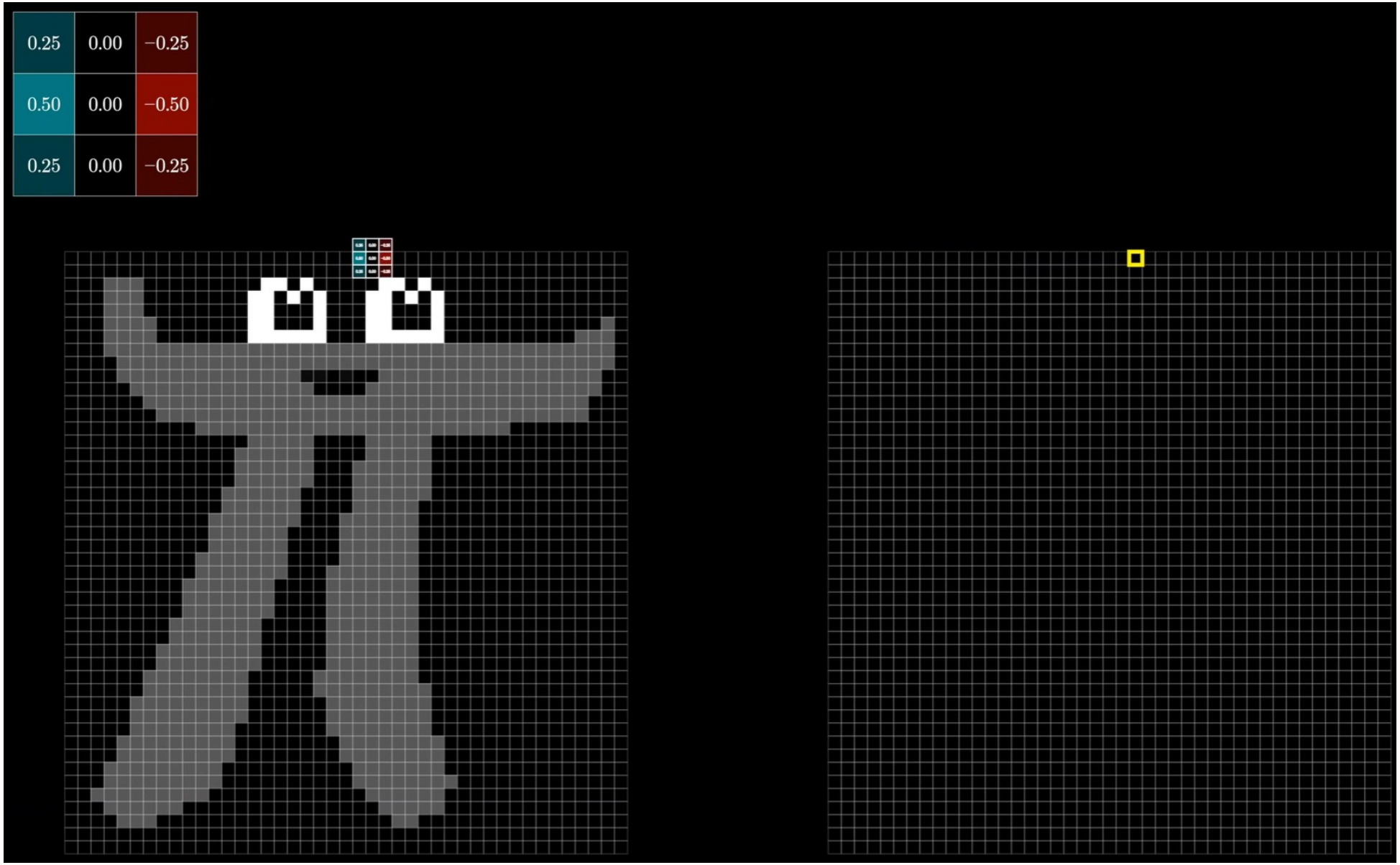
https://www.youtube.com/@3blue1brown

Animated examples



https://www.youtube.com/@3blue1brown

Animated examples

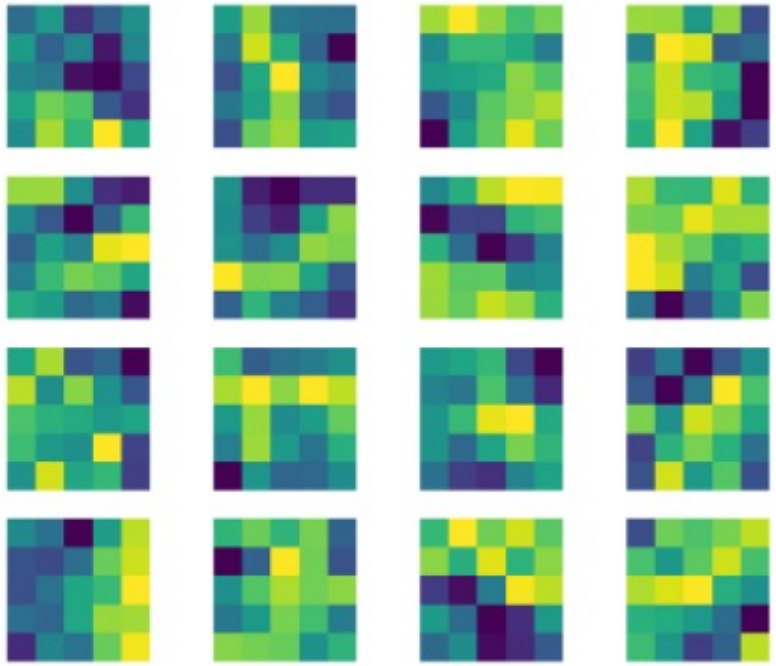


https://www.youtube.com/@3blue1brown

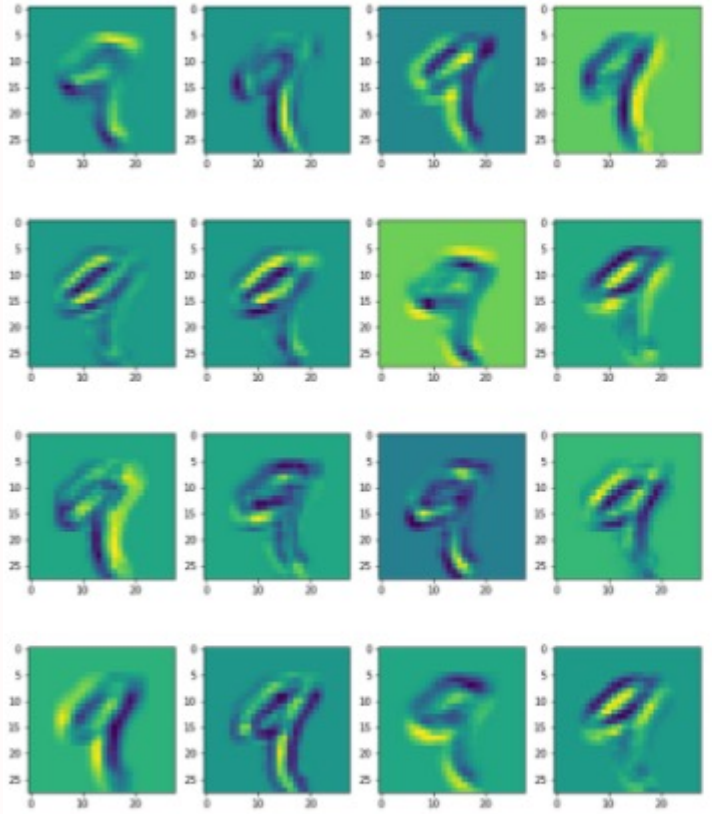
Feature map

- Each convolution/filter yields a “map” of features
  - Like a heatmap of where that feature appears
  - Or a correlation map of the filter and a pixel in the image

Input



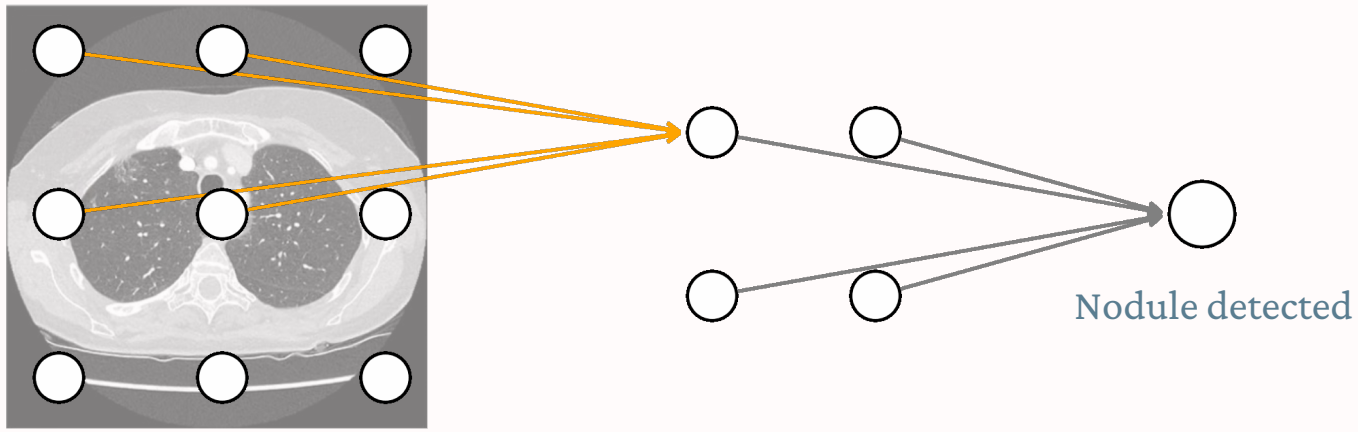
Filters



Output

Feature map

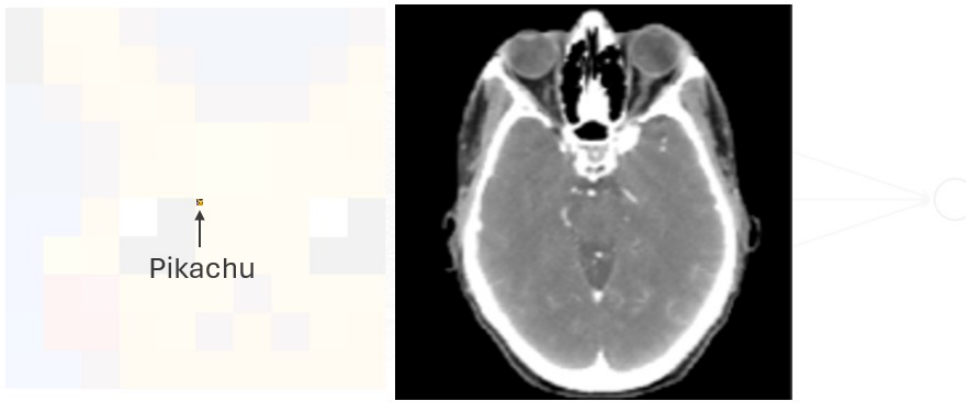
- Squint and imagine:
  - Orange weights detect nodules (2 × 2 pixels).
  - Nodules can be in one of four corners.
- Can detect nodules everywhere once learned to detect anywhere!
  - Imagine tumors, organs, lesions etc.
- More effective with fewer parameters.



But wait… computation?

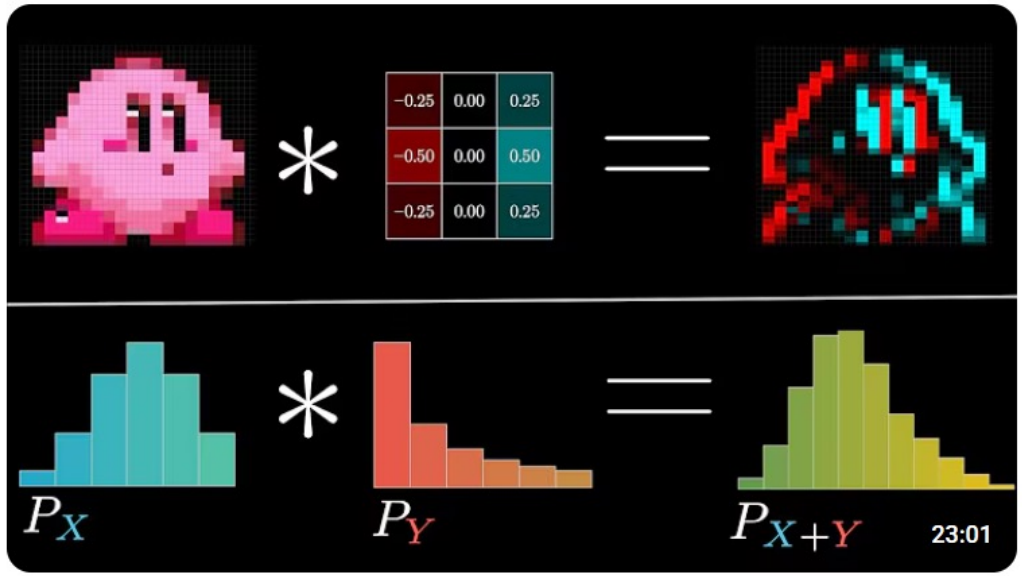
- One of our original motivations
- One feature map for 256x256 image
  - Input: 256x256=65536 pixels
  - Output: 256x256=65536 feature map
  - 3x3 convolution
  - 589824 multiplications (not including sums)
- By comparison, fully connected layer
  - Input: 256x256=65536 pixels
  - Output: Let's say 65536 "features"
  - 4294967296 multiplications (again, not including sums)
- 0.01% multiplications

- Real images are much larger
- CT or MRI matrix size typically 256x256 or 512x512
  - i.e. 65536 pixels or 262144 pixels



If you’re interested…

- Highly recommended to watch



But what is a convolution?

3.2M views • 2 years ago

3Blue1Brown

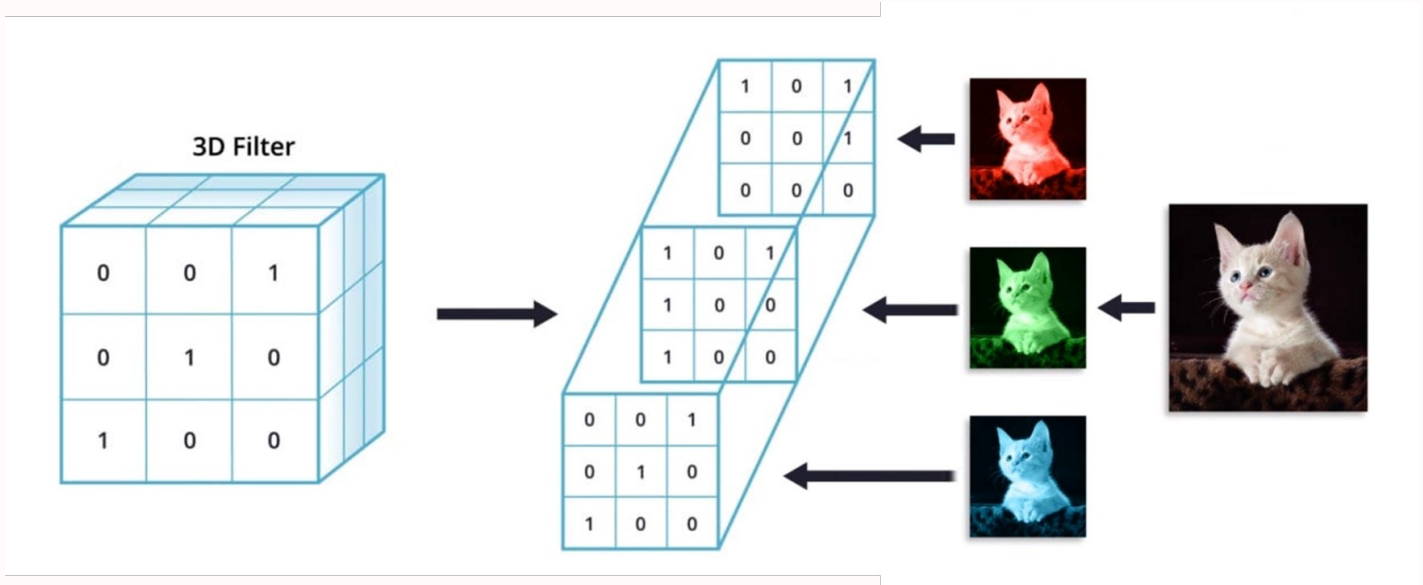
Discrete **convolutions**, from probability to image processing and FFTs. Video on the continuous case: ...

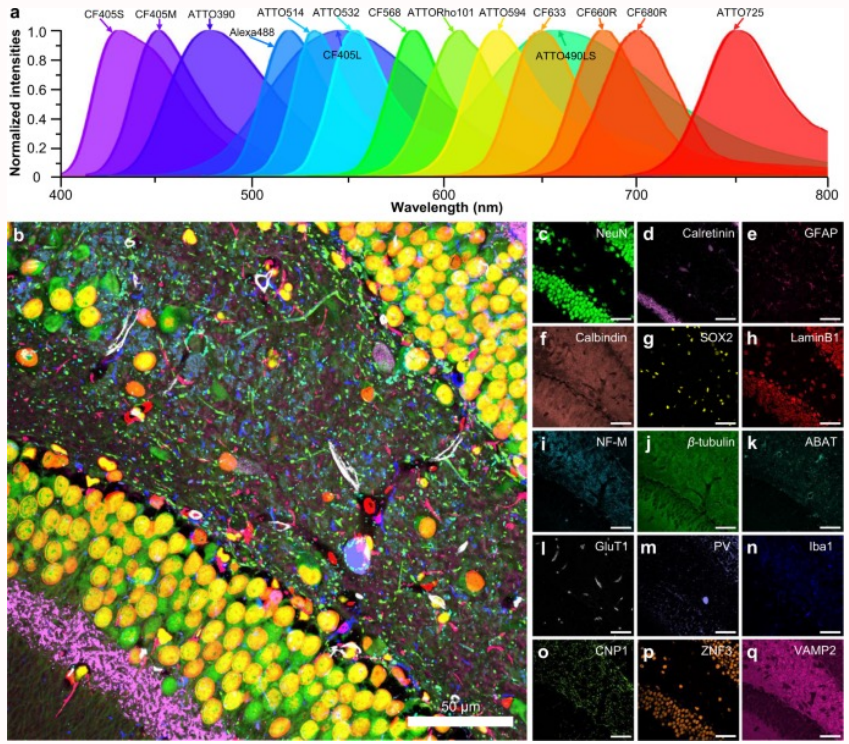
4K CC

9 chapters Where do convolutions show up? | Add two random variables | A simple example |...

One last tidbit

- Filters can be 3D!
- Convolutions applied to volume

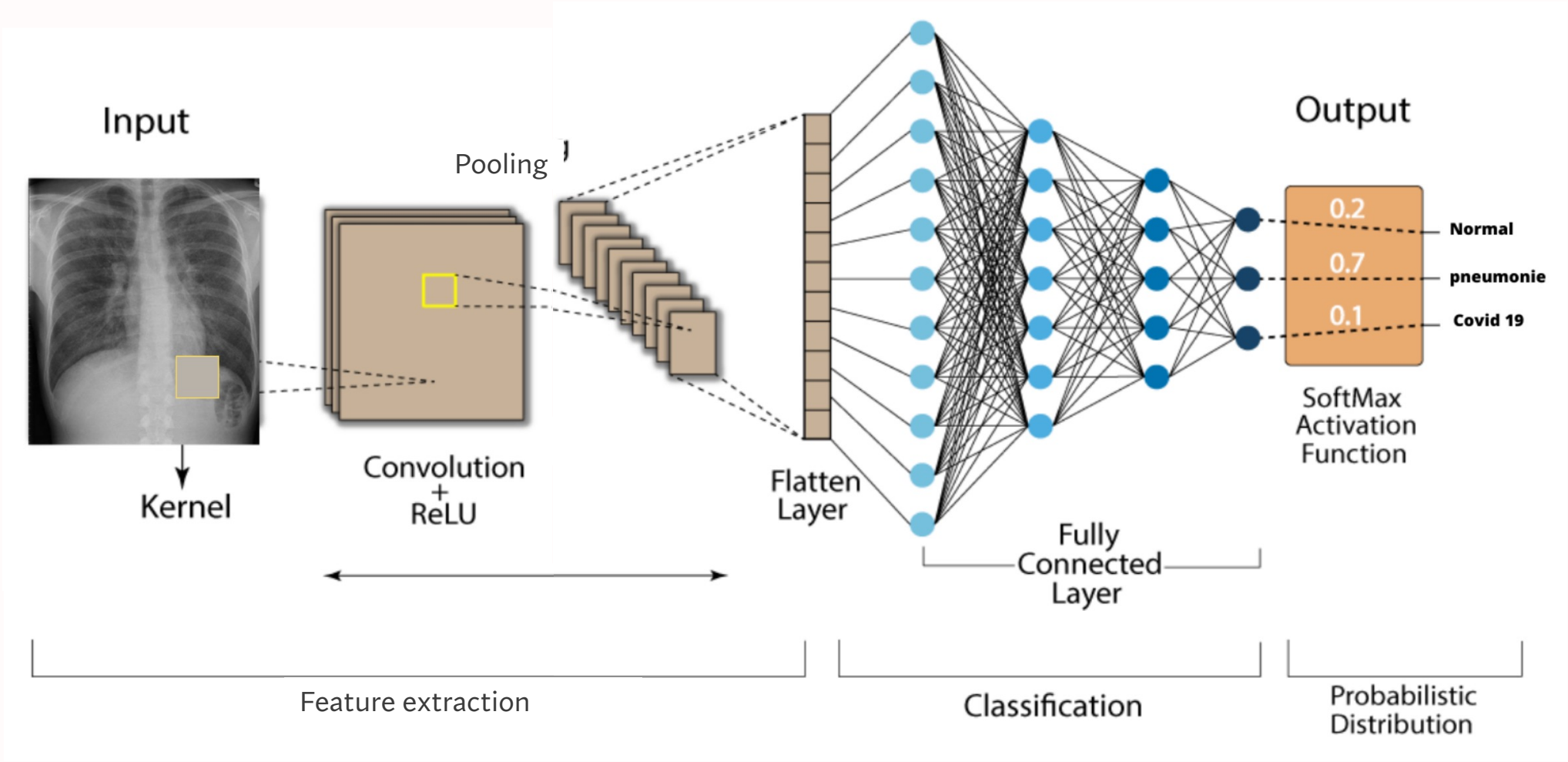




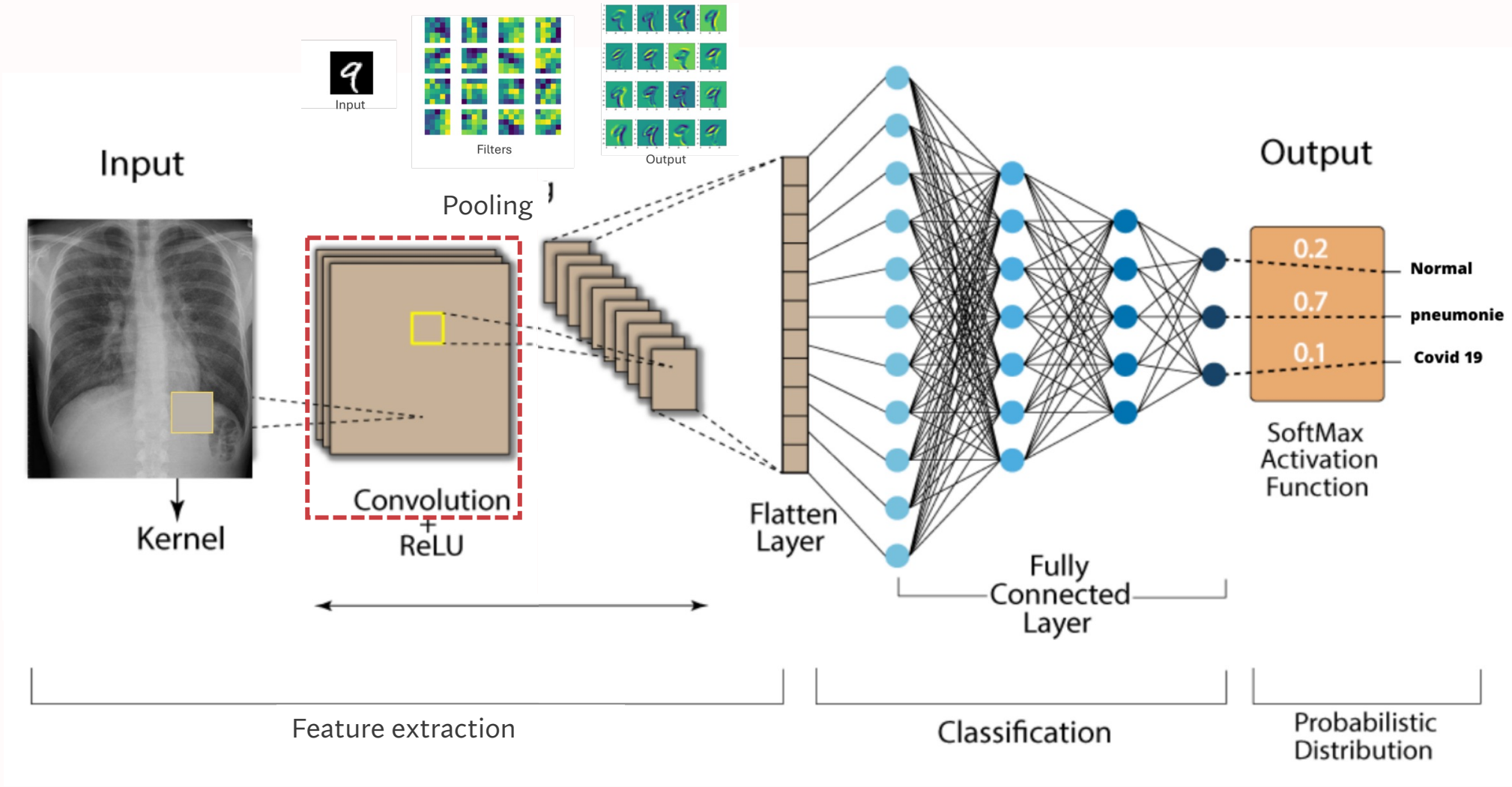
Building a convolution neural network (CNN)

Extracting features to classify

Typical CNN stack

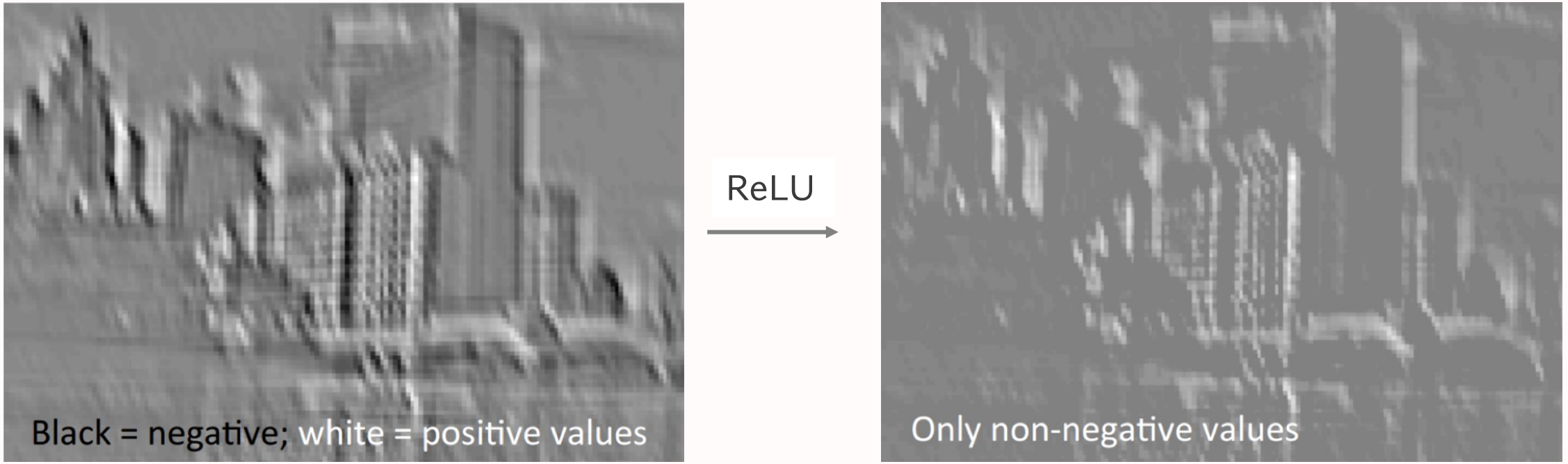


Typical CNN stack



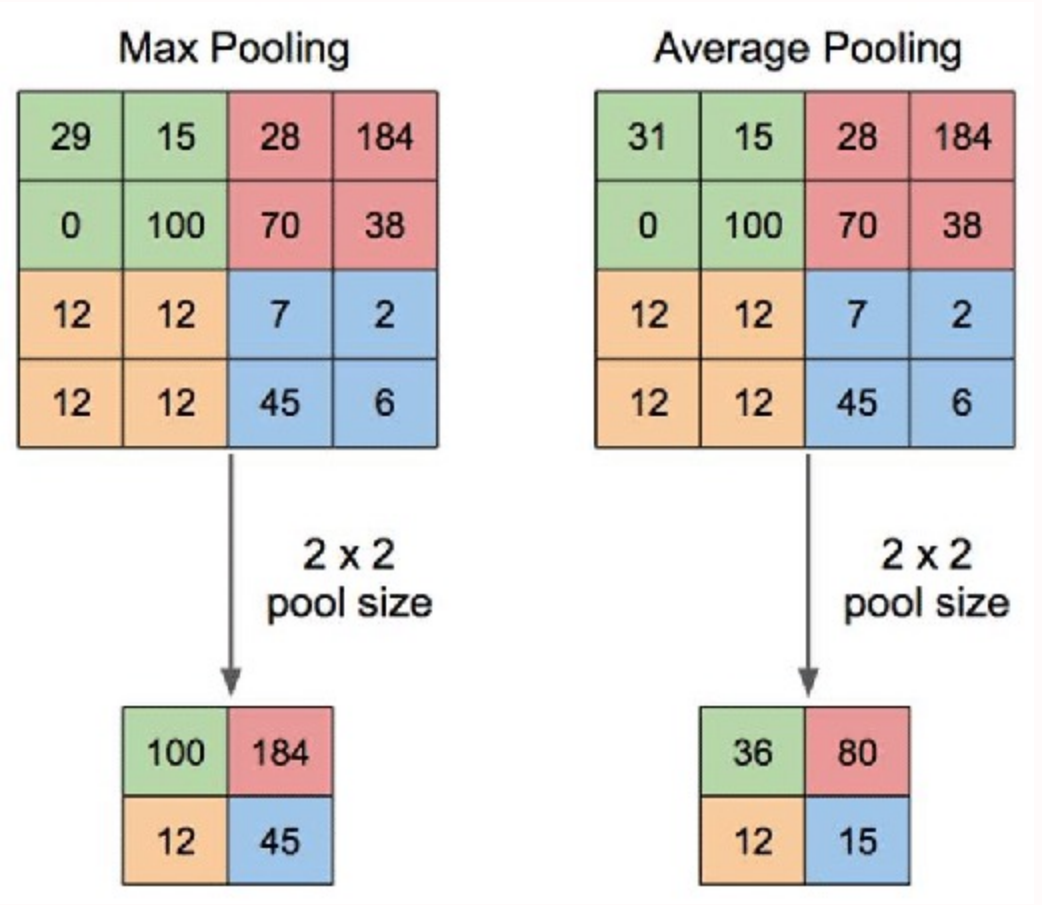
ReLU

- Map negative values to zero
- Allows us to introduce non-linearity
  - Convolution is strictly linear



Pooling

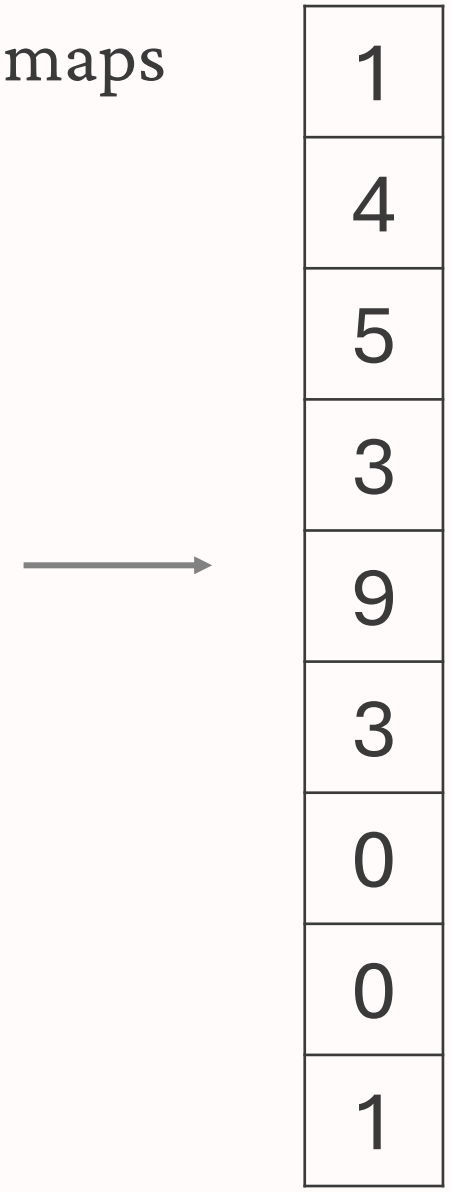
- Allow the network to capture the ‘gist’ of the feature map
  - Lower resolution
- Fewer downstream parameters
- Allows for multi-scale feature extraction
  - More on this later

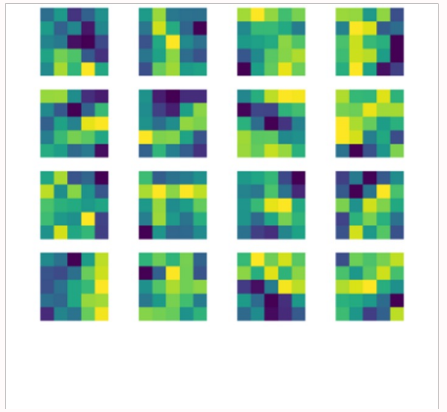


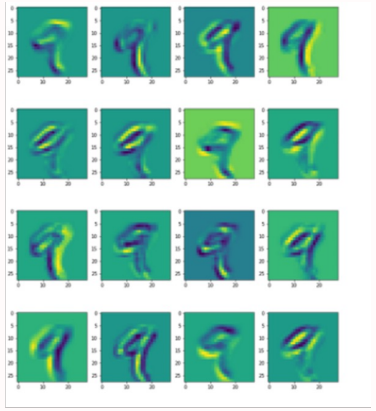
Flattening

- Reshape your feature maps

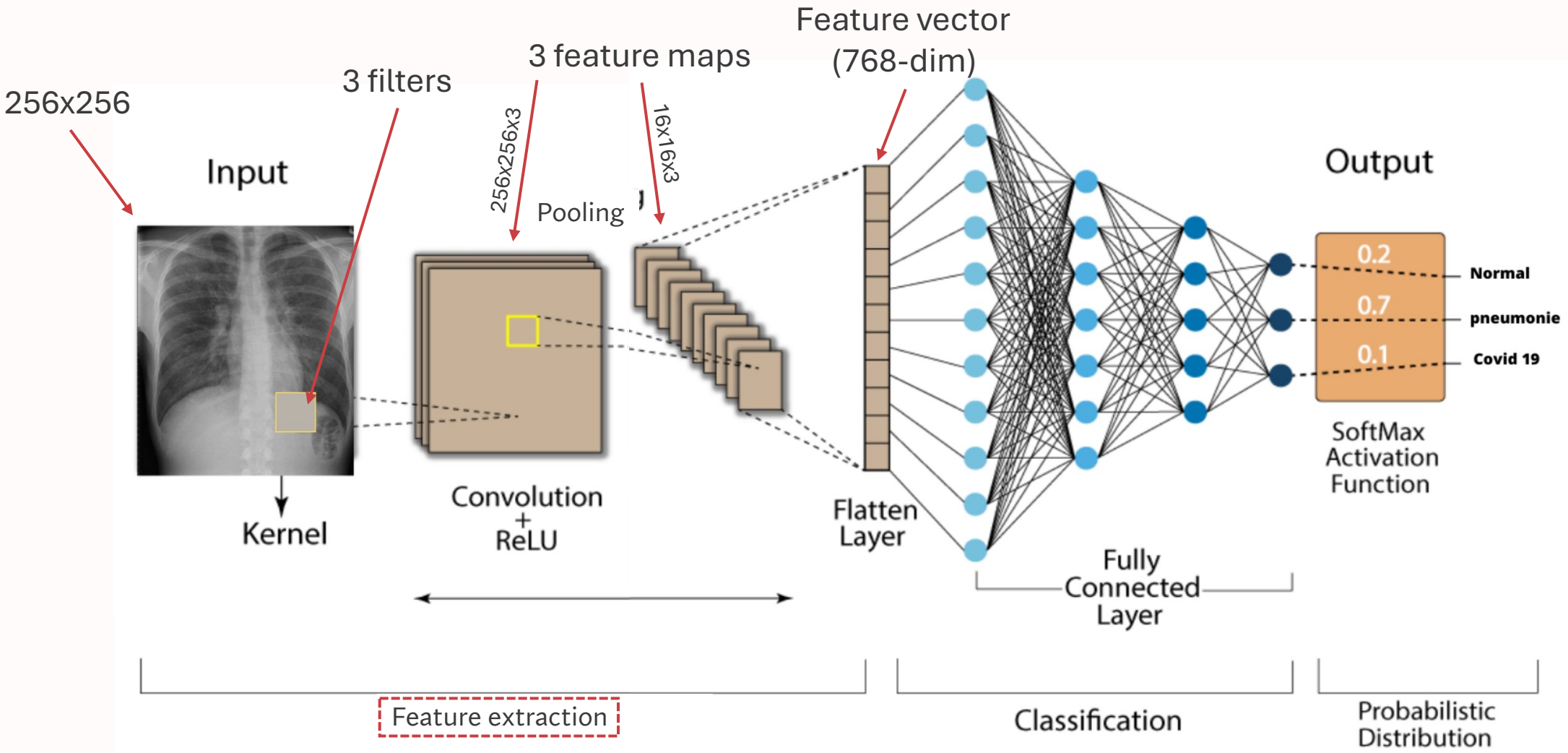
1	4	5
3	9	3
0	0	1







Typical CNN stack

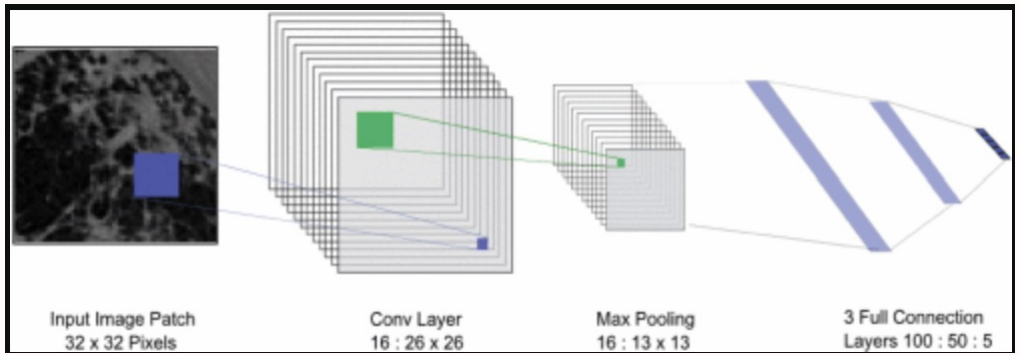


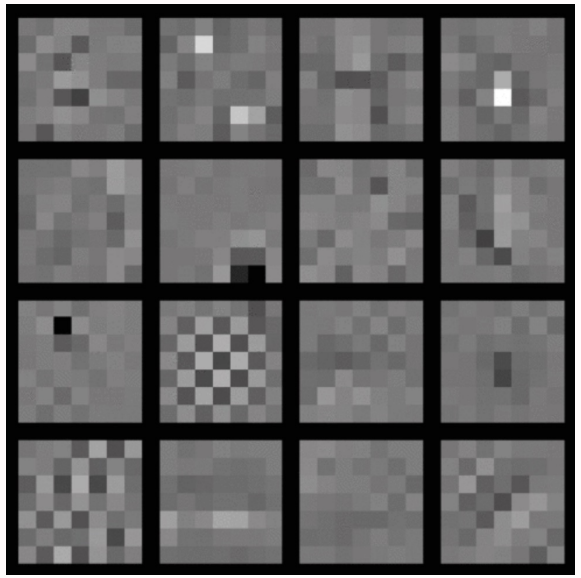
Clinical applications

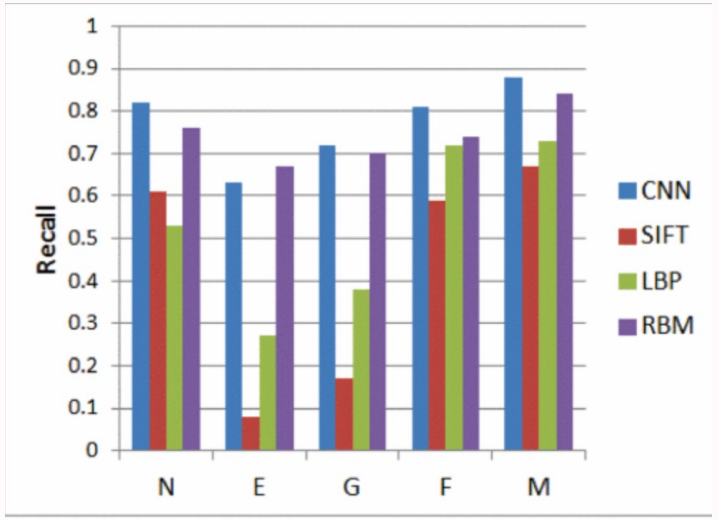
and curiosities

Medical image classification

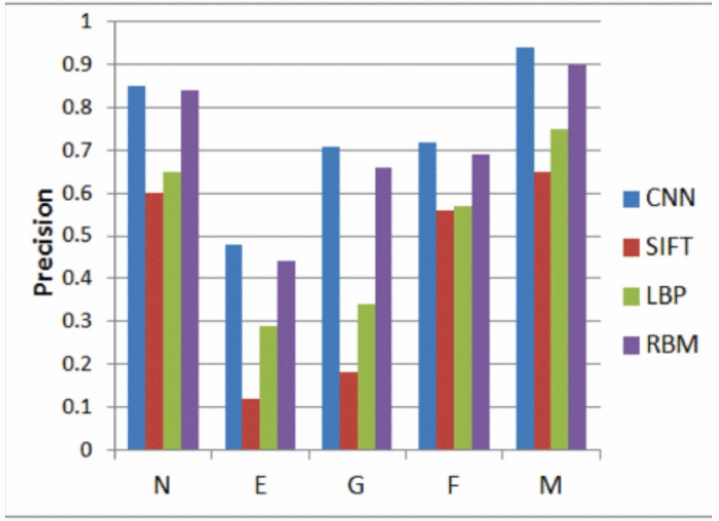
- Classify lung image patches into interstitial lung disease
- HRCT images
- 32x32 pixels







(a)

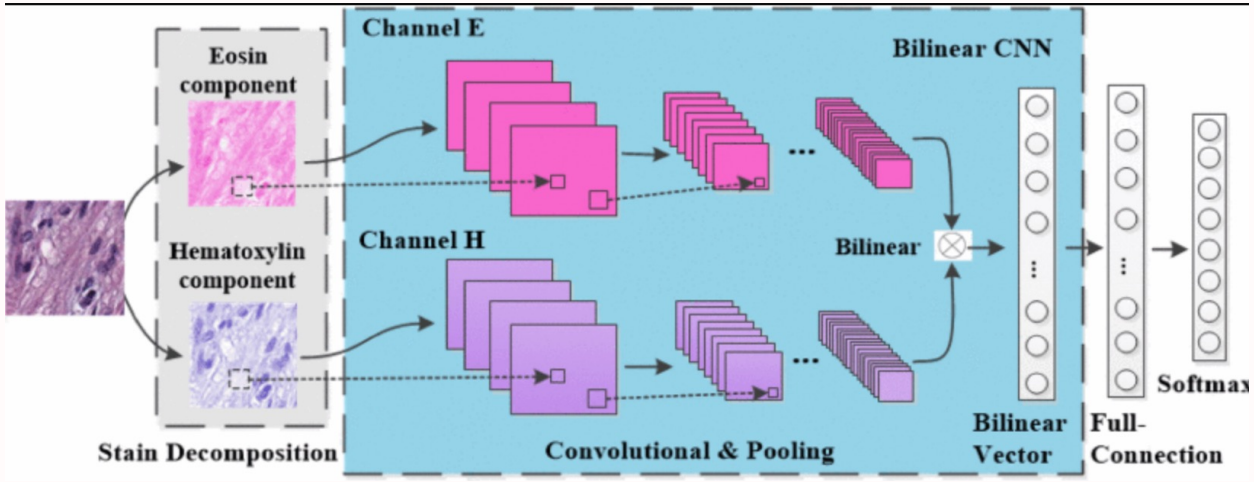


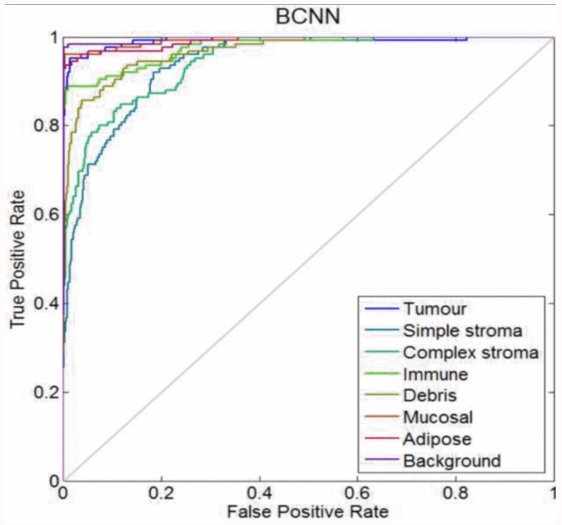
(b)

Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng and M. Chen, "Medical image classification with convolutional neural network," 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV), Singapore, 2014, pp. 844-848

Medical image classification

- Colorectal cancer tissue classification
  - Tumor
  - Stroma
  - Immune cells
  - Debris
  - Mucosa
  - Adipose
  - Background

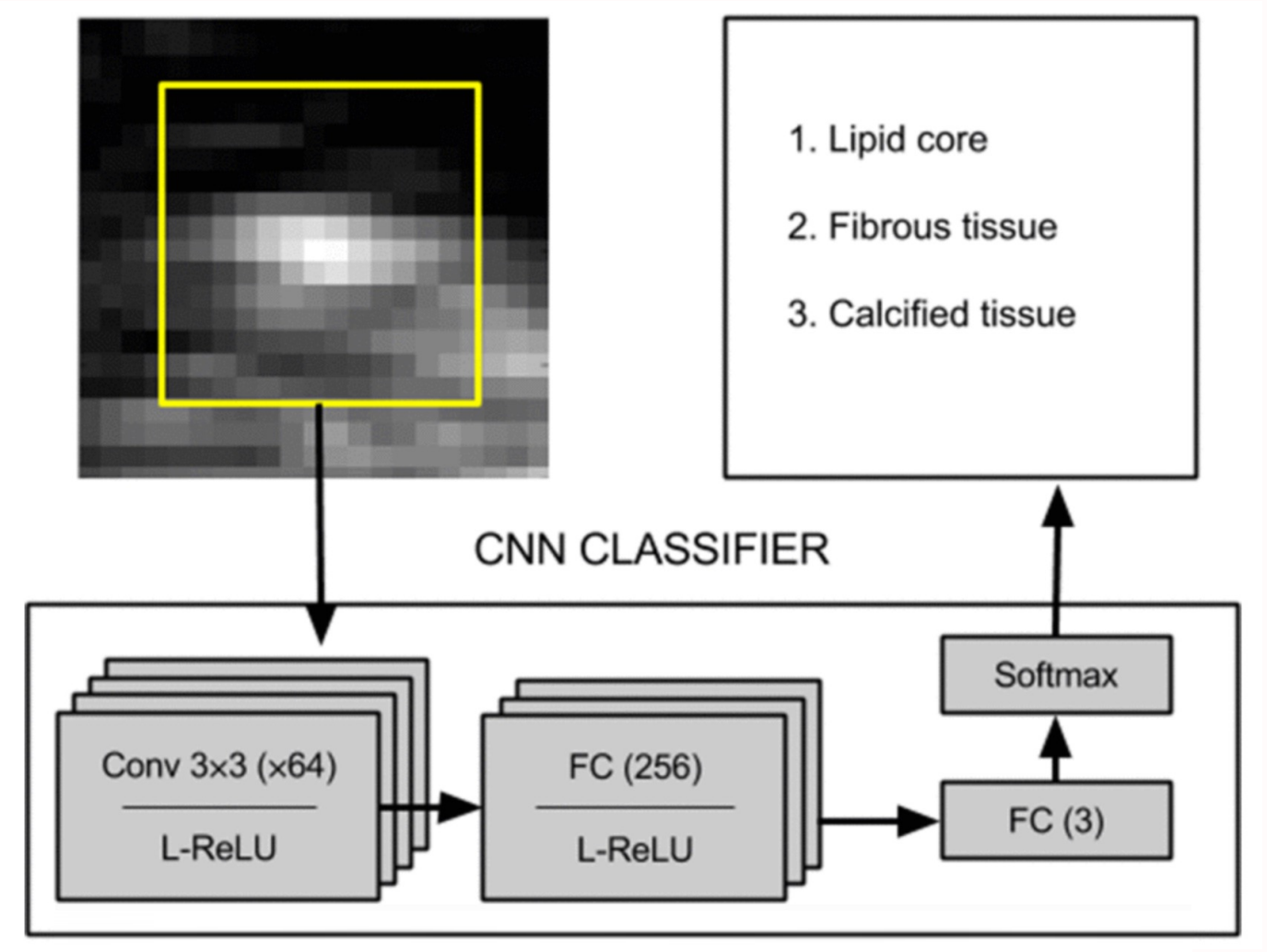




C. Wang, J. Shi, Q. Zhang and S. Ying, "Histopathological image classification with bilinear convolutional neural networks," 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju, Korea (South), 2017, pp. 4050-4053

Medical image classification

- Carotid plaque classification in ultrasound images



Curious case study

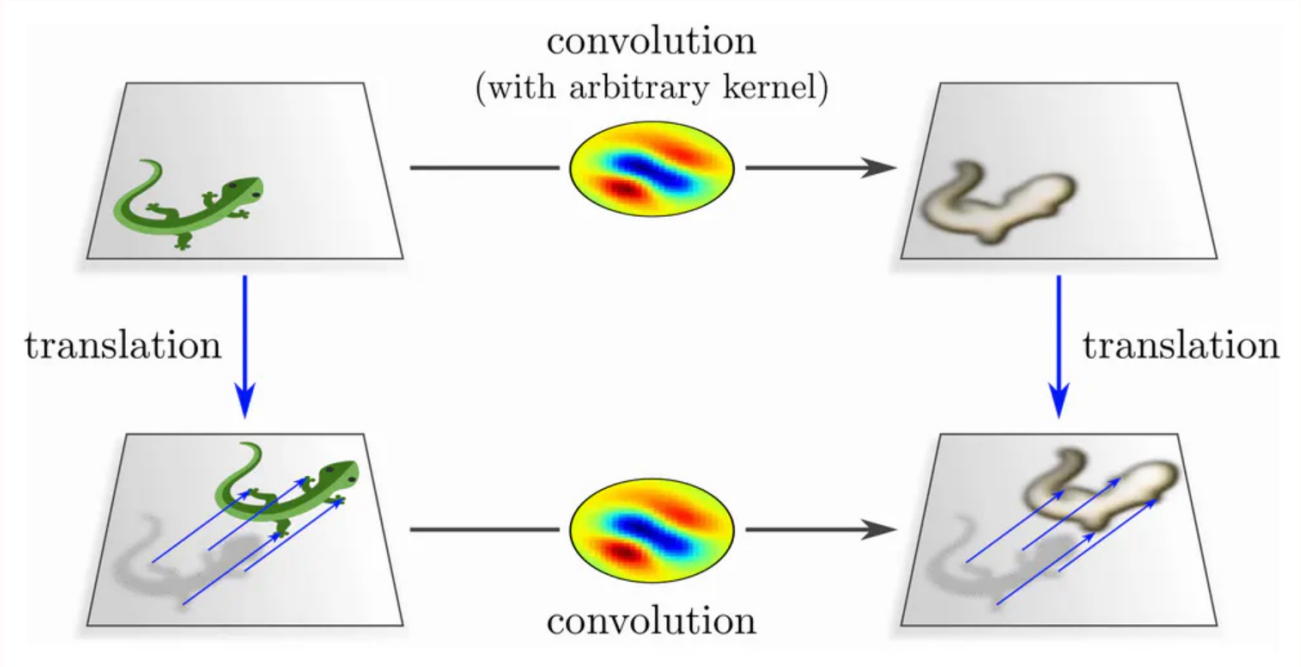
- Deep learning on retinal fundus photos well-established.
  - Monitor diabetic retinopathy.
- Out of curiosity, predict sex using CNN.
  - 80-90% accurate, externally validated.
  - No one knows why!
- Clinician-driven:
  - "… we present the development of a deep learning model by clinicians …"
  - "Our deep learning model was trained using code-free deep learning (CFDL) with the Google Cloud AutoML platform."

Translation equivariance

Detecting a feature anywhere

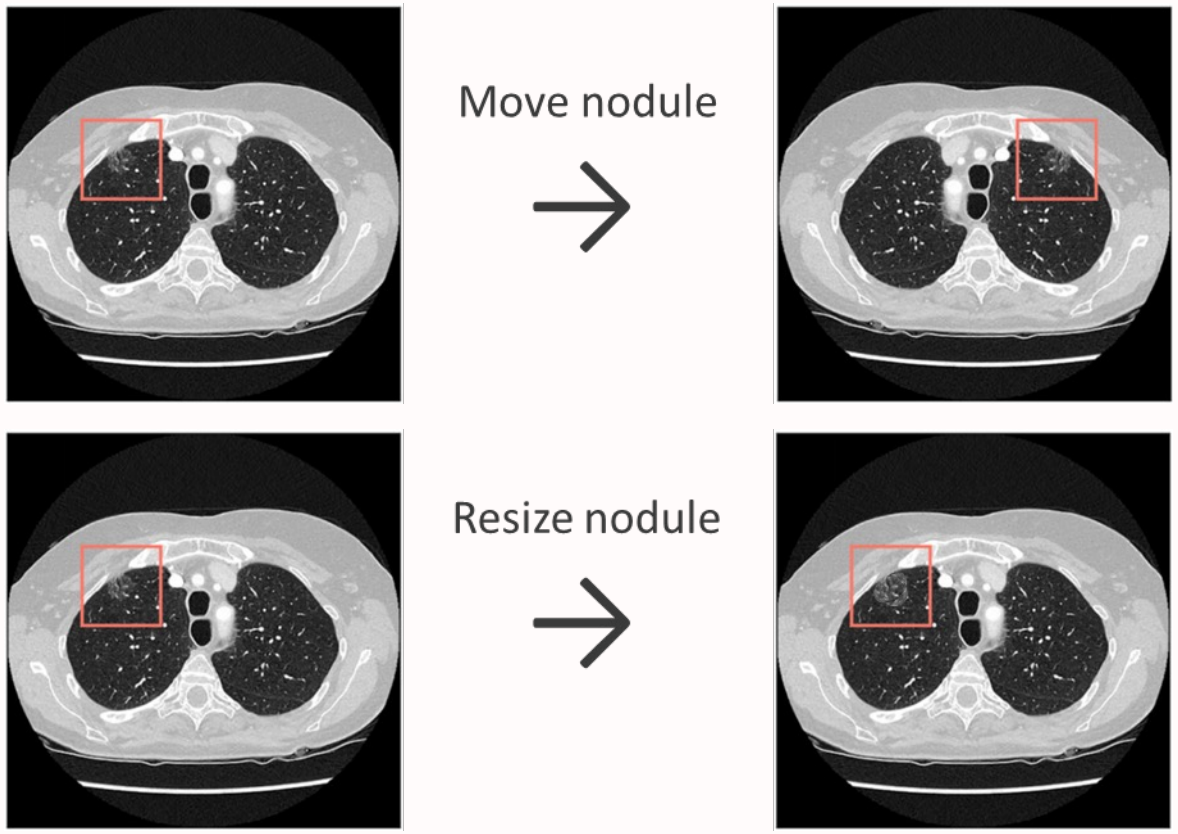
What is it?

- Property of convolution (and other mathematical operations)
- When the input is shifted, the corresponding feature map also shifts
- NOT a property of fully connected networks
  - Remember, a neuron is required for every position of a feature
- Allows CNNs to detect feature anywhere in an image



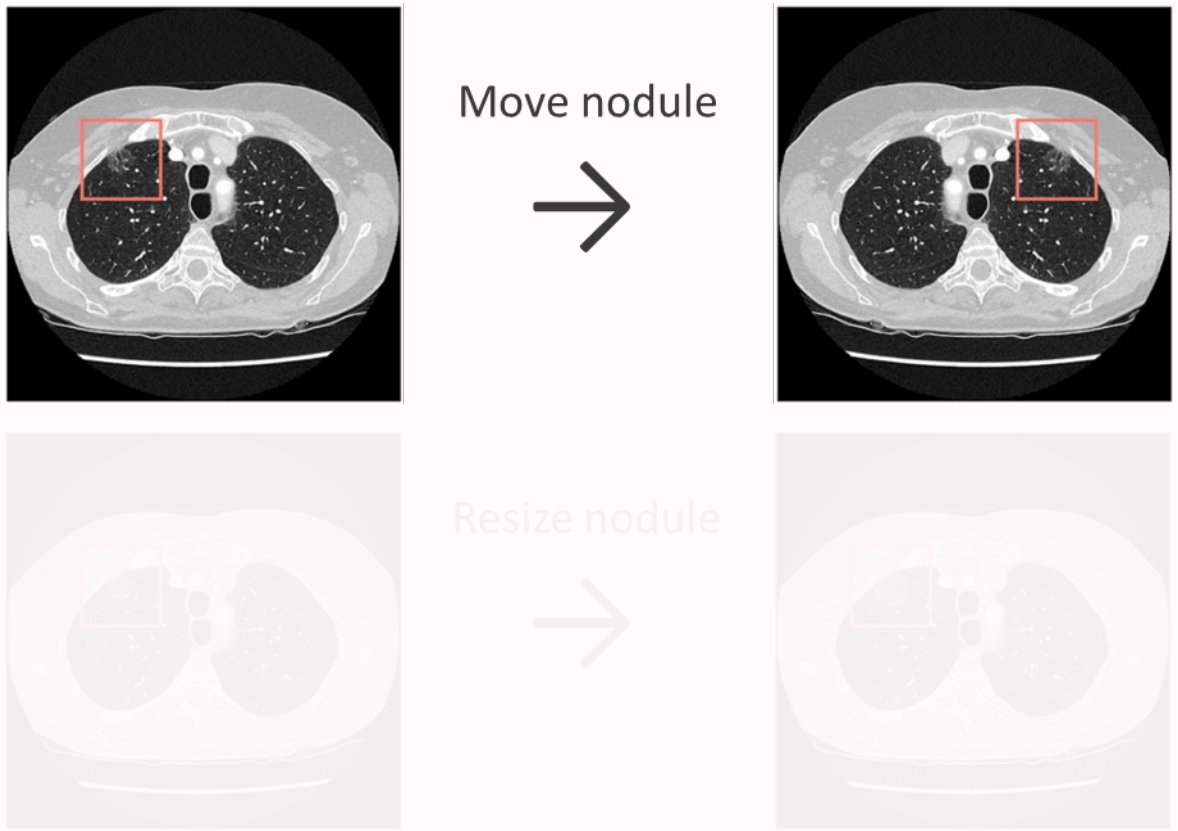
Same feature can be everywhere

- For example, lung nodules
  - **Different position**
  - Different scale



Same feature can be everywhere

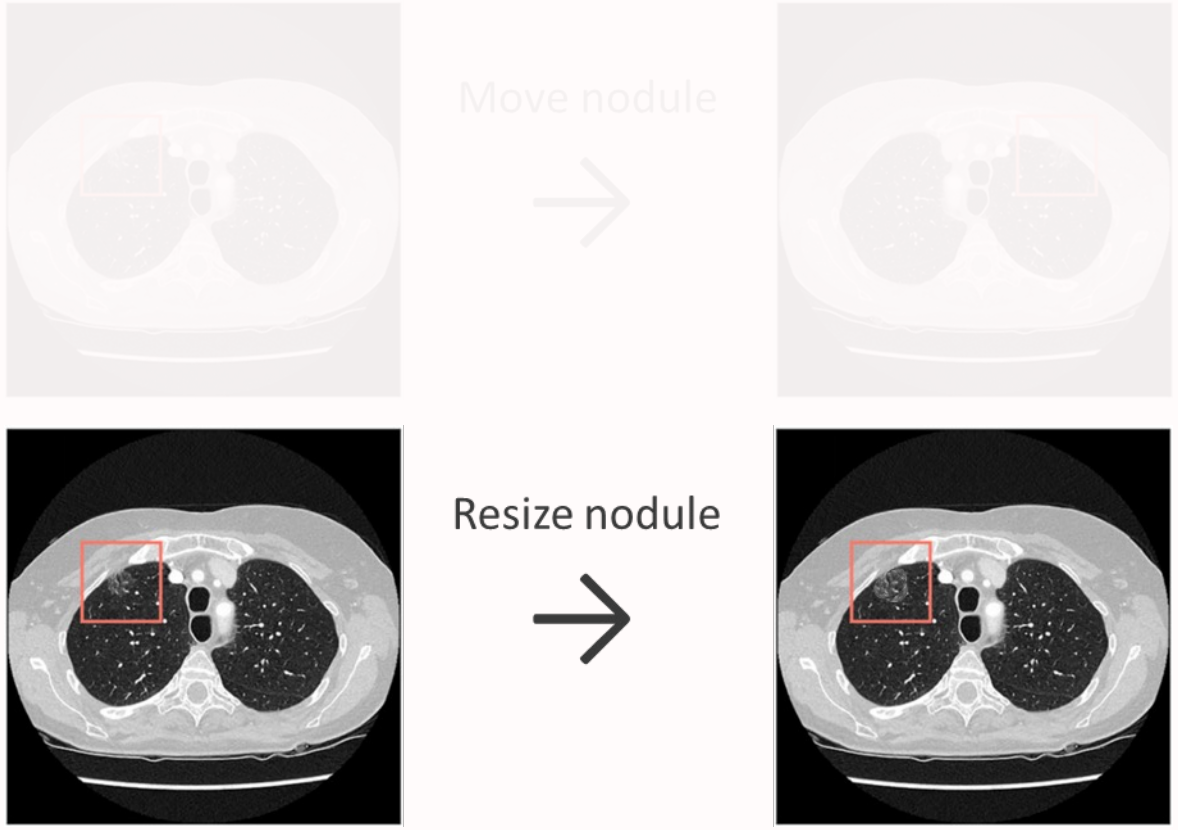
- For example, lung nodules
  - **Different position (i.e. translation)**
  - Different scale



Courtesy of Dr. Foong

Same feature can be everywhere

- For example, lung nodules
  - Different position (i.e. transation)
  - **Different scale**



Courtesy of Dr. Foong

Why not?

- (Some) filters are not scale equivariant

0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	1	0	0	0
0	0	1	1	1	0	0
0	0	0	1	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

*

0	1	0
1	1	1
0	1	0

=

0	0	1	2	0
0	2	2	2	0
1	2	**4**	2	1
0	2	2	2	0
0	0	1	0	0

✚ ✚ 1 detection!

Why not?

- (Some) filters are not scale equivariant

0	0	0	0	0	0	0
0	0	1	1	0	0	0
0	1	1	1	1	0	0
0	1	1	1	1	0	0
0	0	1	1	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

*

0	1	0
1	1	1
0	1	0

=

2	3	3	2	0
3	**4**	**4**	3	1
3	**4**	**4**	3	1
2	3	3	2	0
0	1	1	0	0

4 detections!

Why not?

- (Some) filters are not scale equivariant

0	0	0	0	0	0	0
0	0	1	1	0	0	0
0	1	1	1	1	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

2	3	3	2	0
0	1	1	0	0

4 detections!

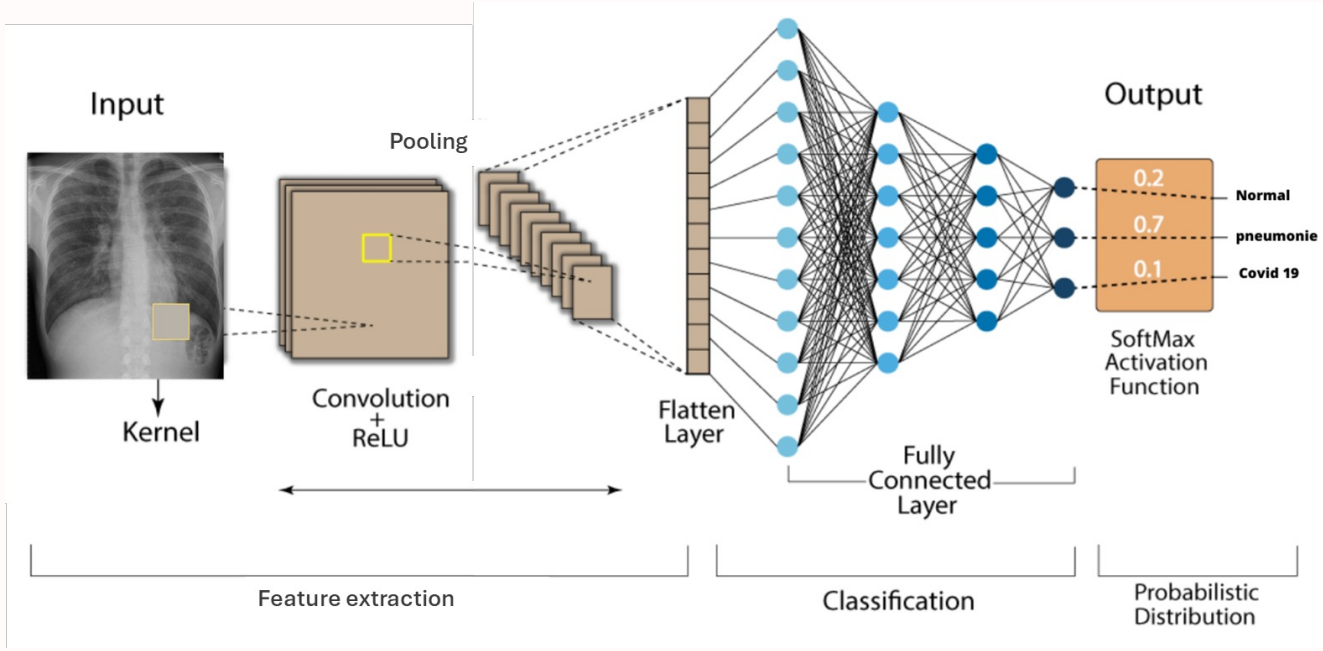
Deep CNNs solve this!

Building Deep CNNs

And how they solve more problems!

Data-driven filters and features

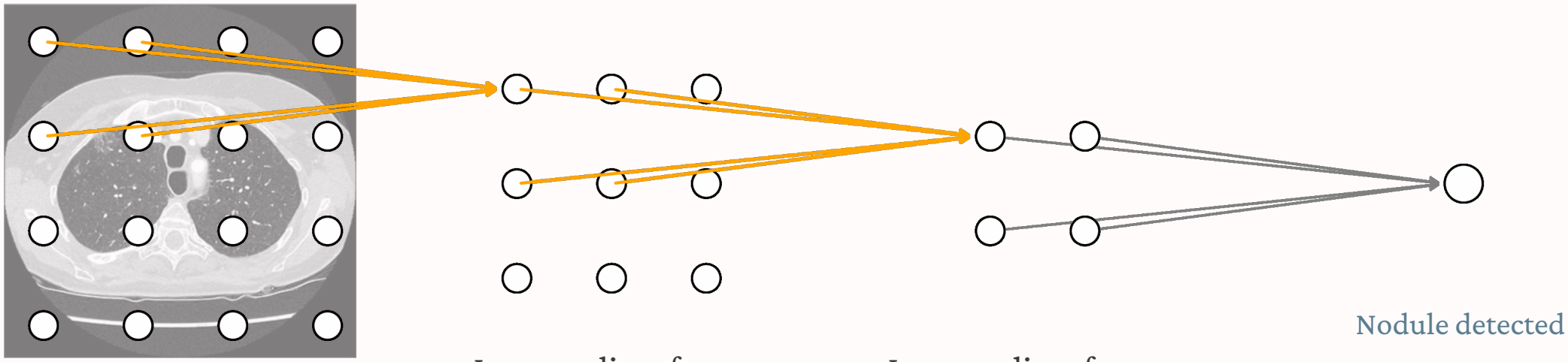
- Traditionally, filters were hand-crafted
  - Yes, there were filter engineers
- Feature extraction was performed separately
  - Then, traditional ML applied on top of features
- Breakthrough in 2012 – AlexNet
  - Feature extraction + ML classifier learned jointly
  - Requirement: lots of data
    - But achieve for free (somewhat)



Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012).

Deep CNNs

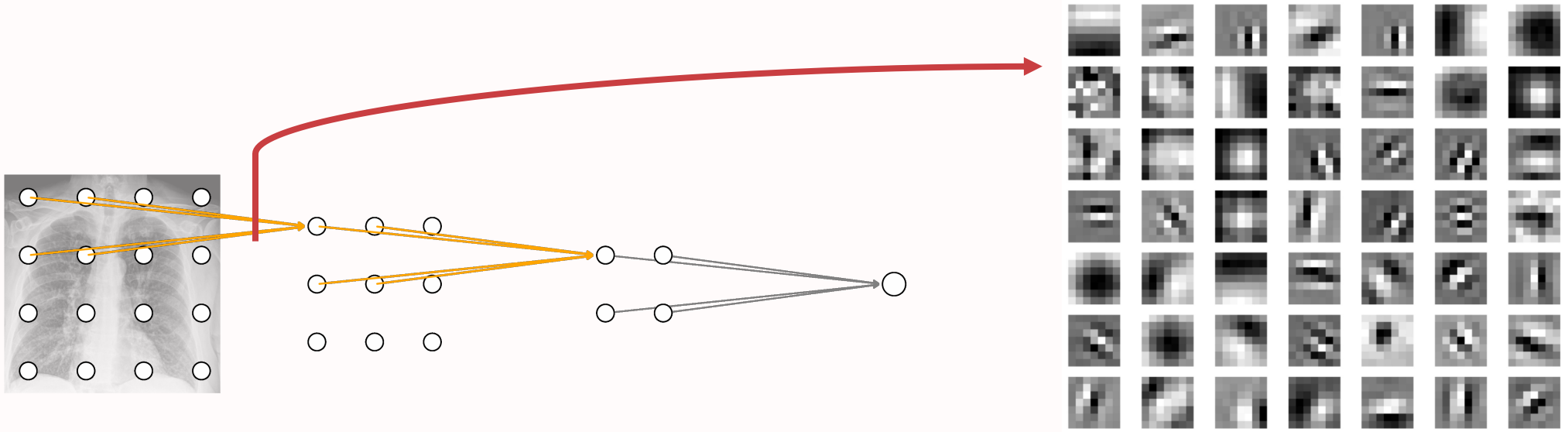
- Idea: stack layers of CNNs on top of each other



- Modern CNNs can have more than 100 layers.
- That's why it's called "deep" learning!

Deep CNNs

- Early weights from actual deep CNN trained on chest X-rays:



What do these remind you of?

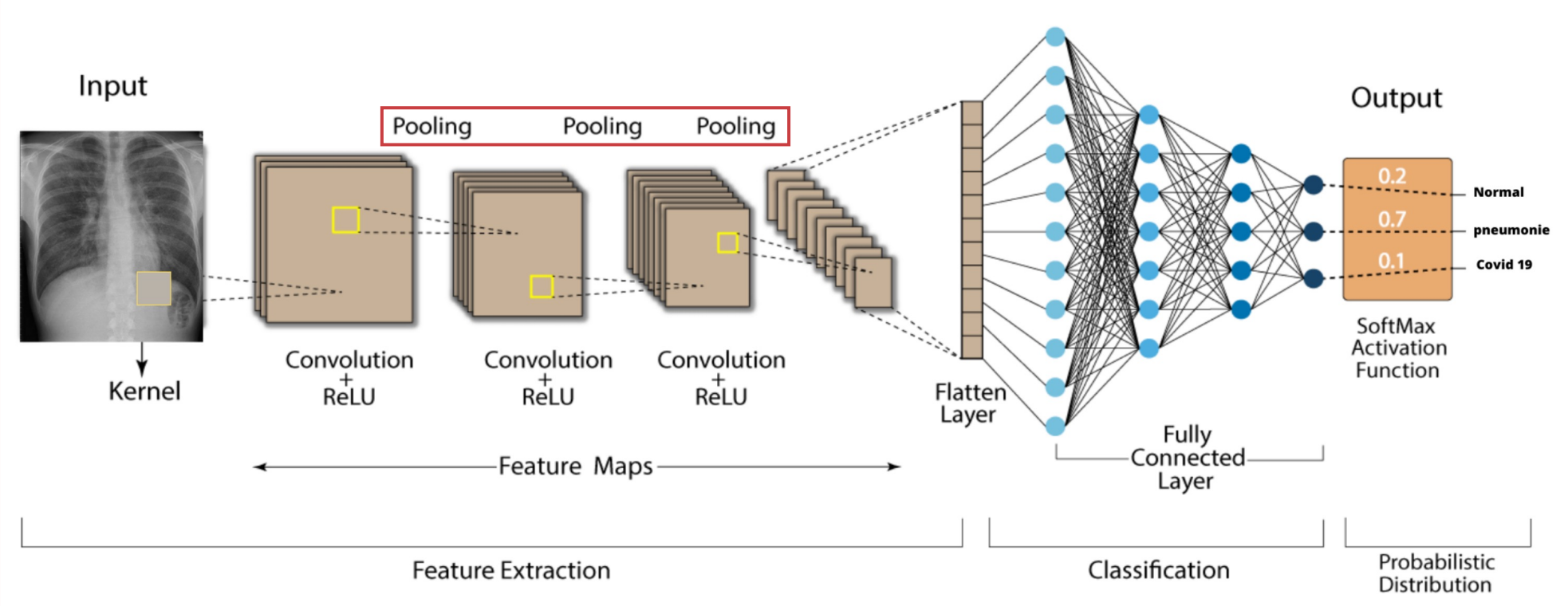
- Real CNNs detect dozens of features each layer.
- Edge and blob detectors in first layer.

Dalton Griner 2023, The Development and Optimization of a Deep-Learning Strategy for COVID-19 Classification in Chest X-ray Radiography, Ph.D. thesis, University of Wisconsin-Madison

Courtesy of Dr. Foong

How does this help scale equivariance?

- It does via pooling!



Pooling enables scale equivariance

0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	1	1	0	0	0
0	0	1	1	1	1	0	0
0	0	1	1	1	1	0	0
0	0	0	1	1	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

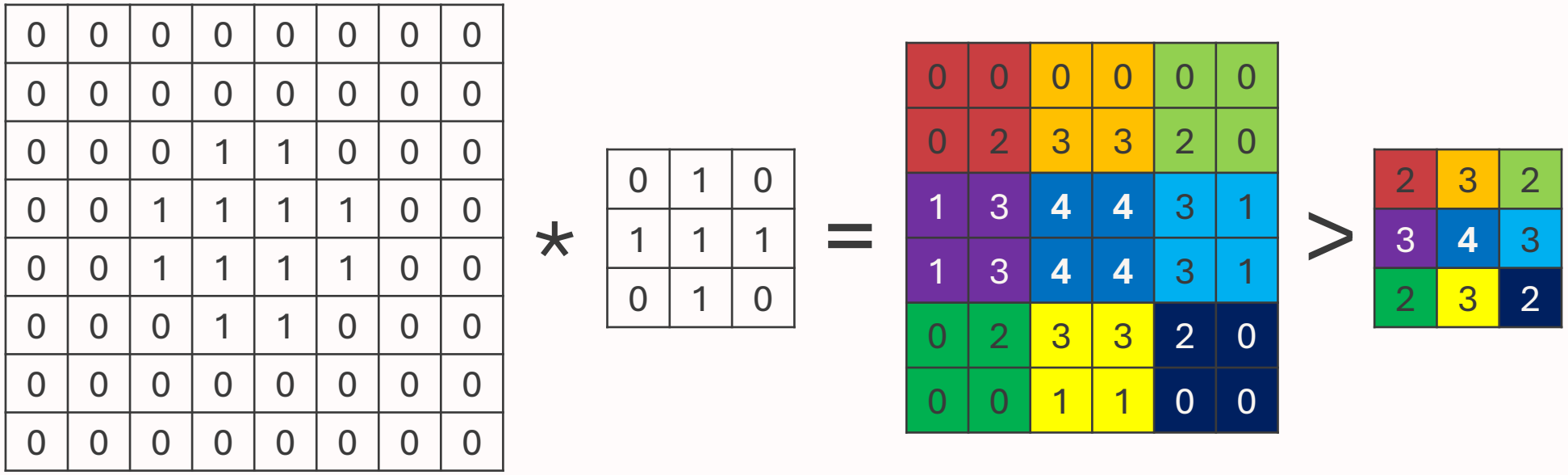
$*$

0	1	0
1	1	1
0	1	0

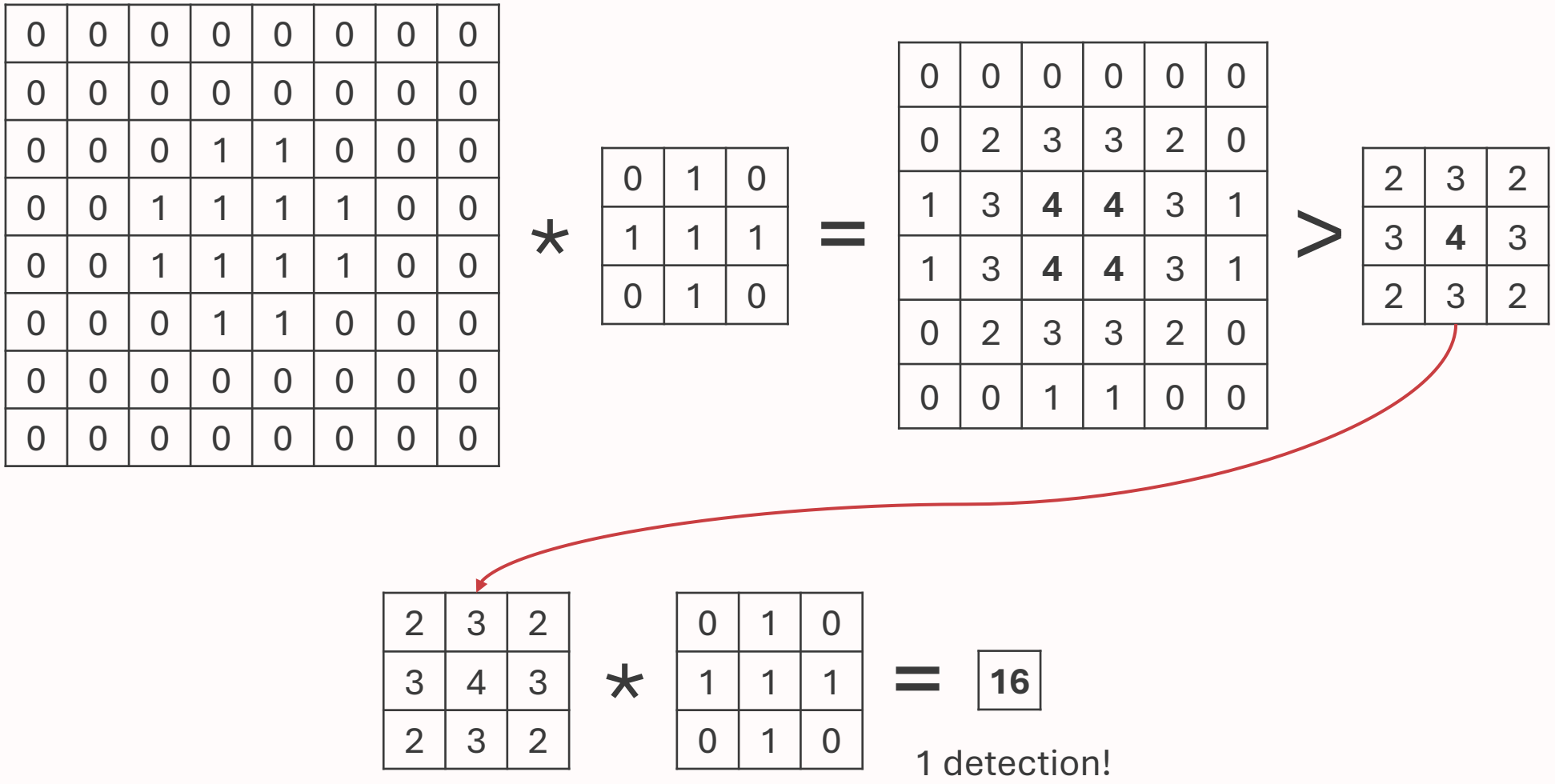
$=$

0	0	0	0	0	0
0	2	3	3	2	0
1	3	**4**	**4**	3	1
1	3	**4**	**4**	3	1
0	2	3	3	2	0
0	0	1	1	0	0

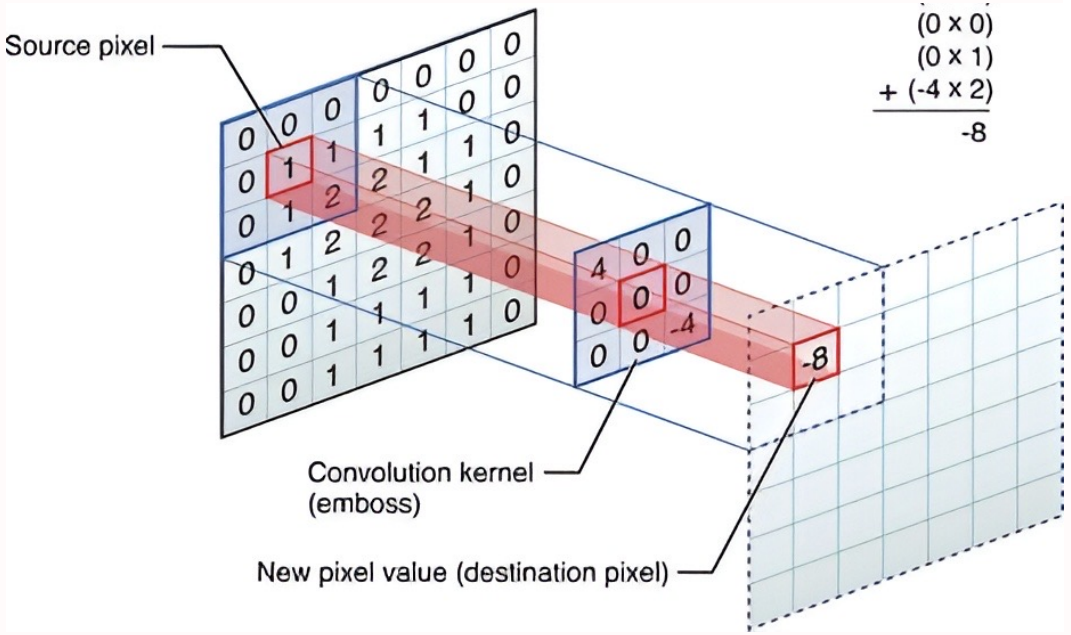
Pooling enables scale equivariance



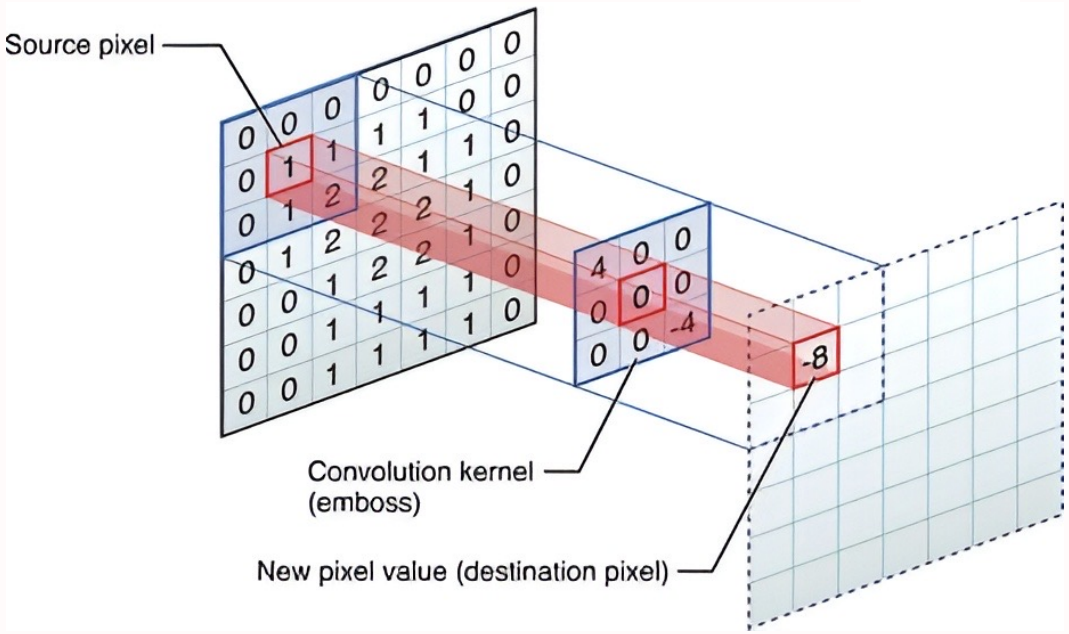
Pooling enables scale equivariance



Pooling increases receptive field

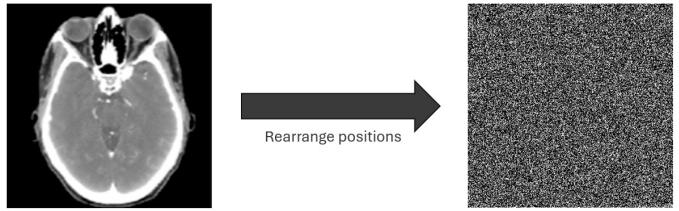


Deep networks increase receptive field

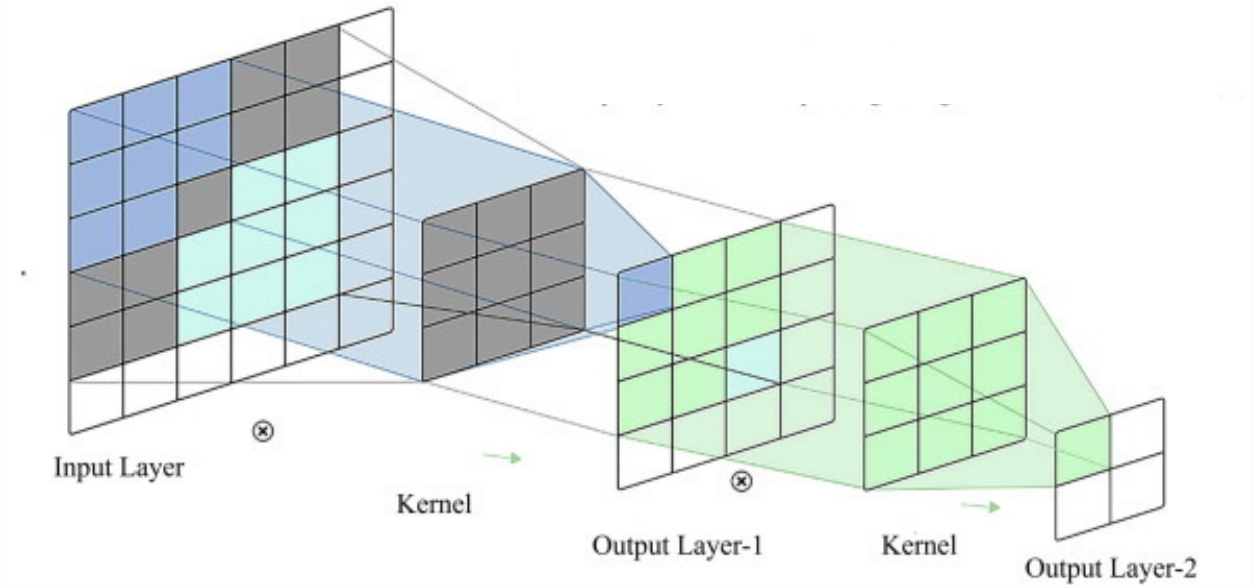


Spatial relationships matter

- Fully connected networks (on images) ignore *spatial* relationships
- One neuron per pixel cannot see structure

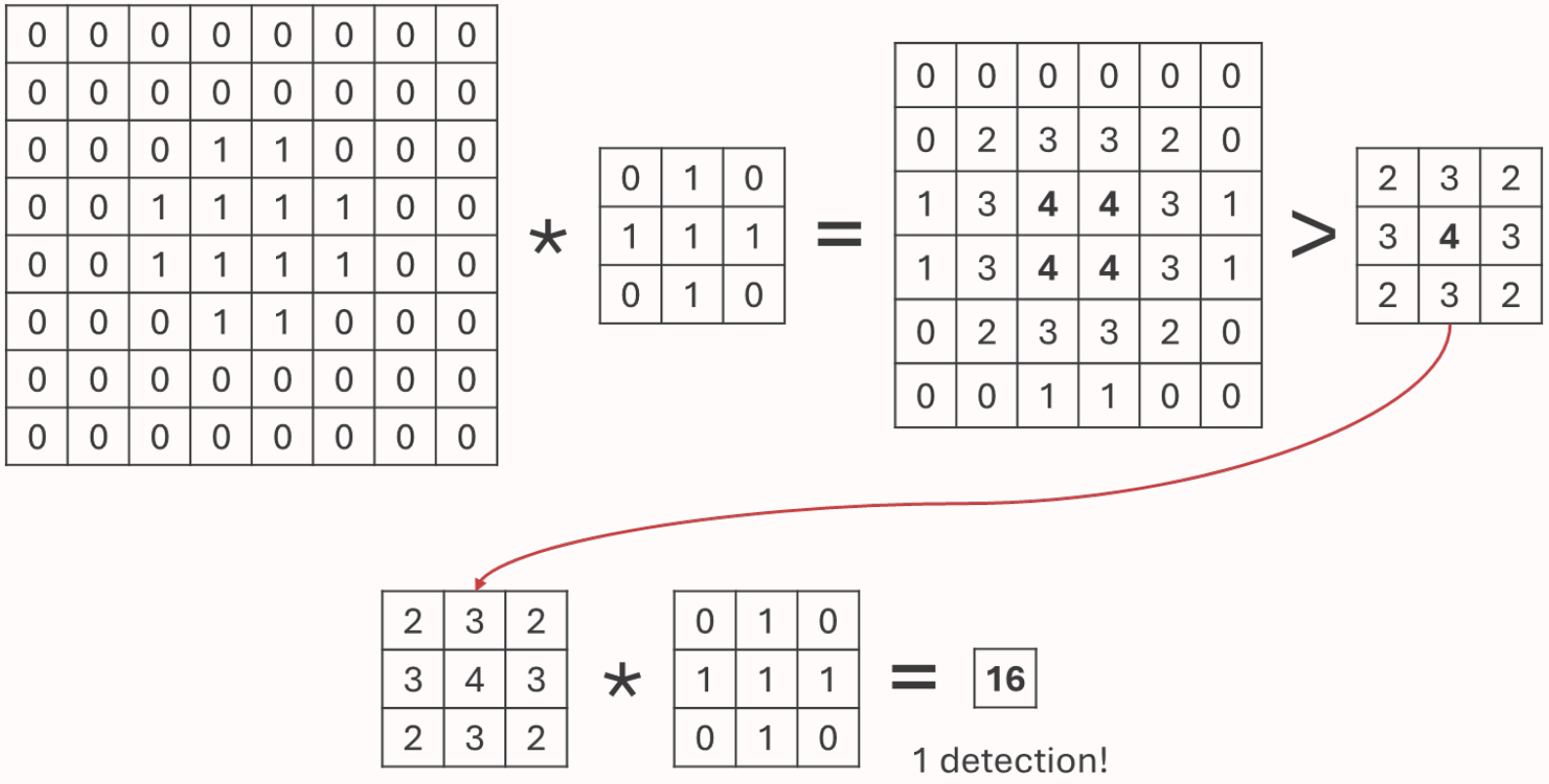


- Later layers "see" more of the input image
- Local to global context
- Enables *spatial* relationships



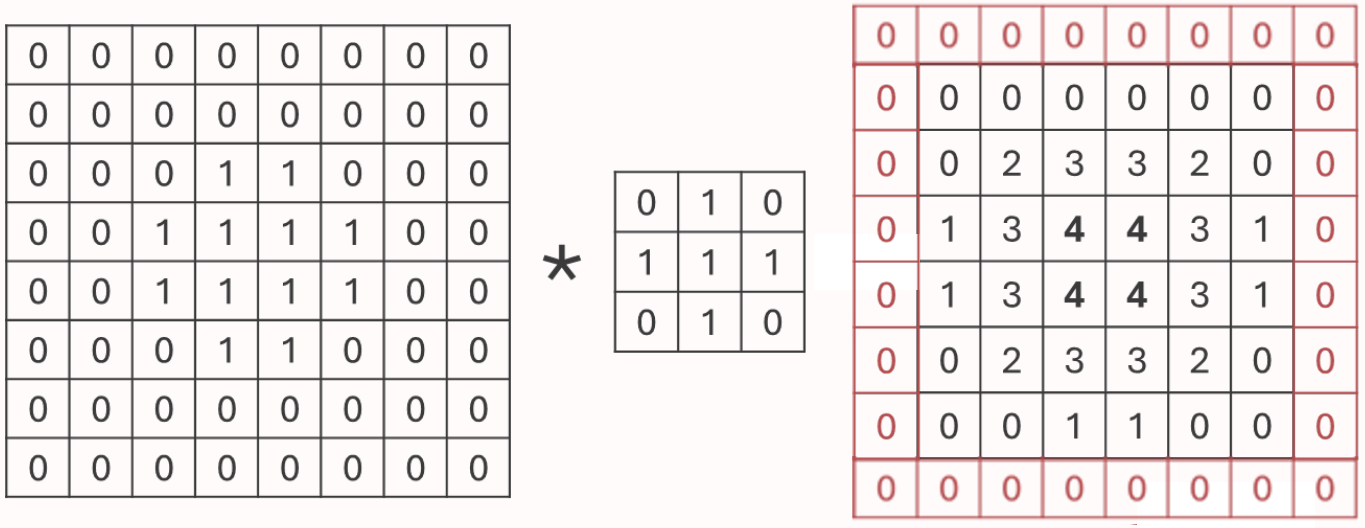
One loose end

- Image/matrix gets smaller and smaller!



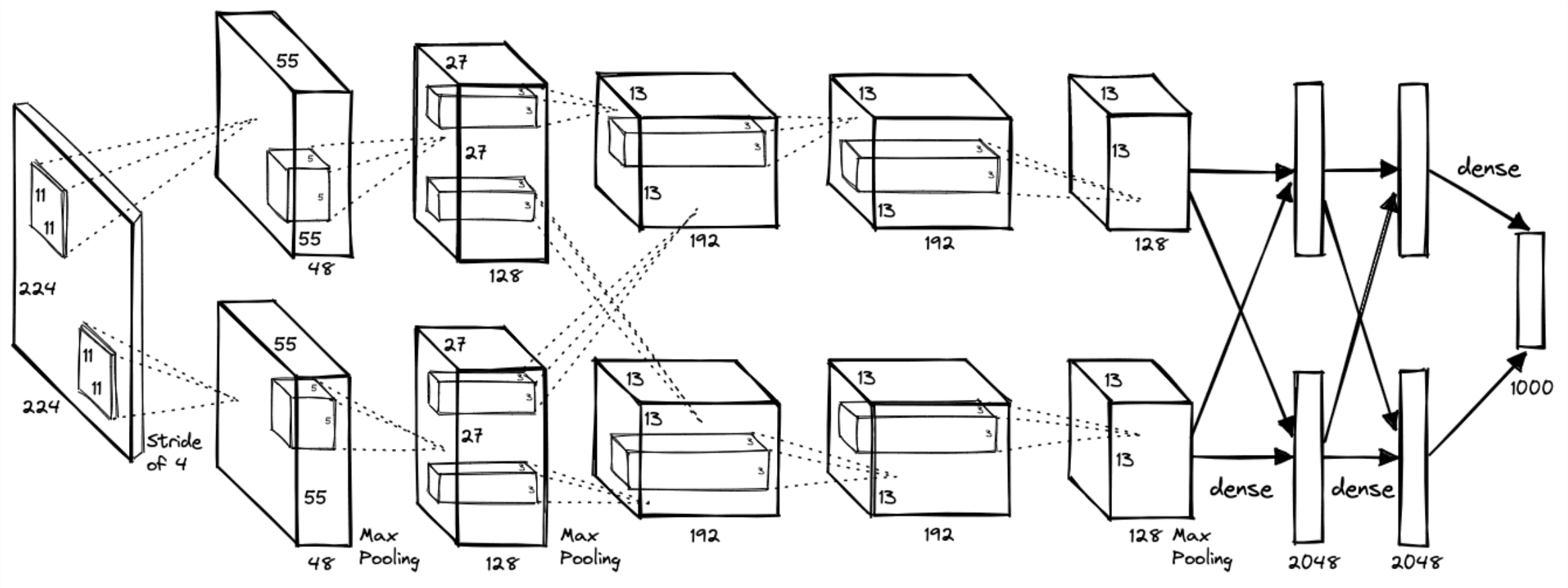
One loose end

- Image/matrix gets smaller and smaller!
  - Padding to solve



- Can be 0's, reflection of image, etc.

AlexNet!



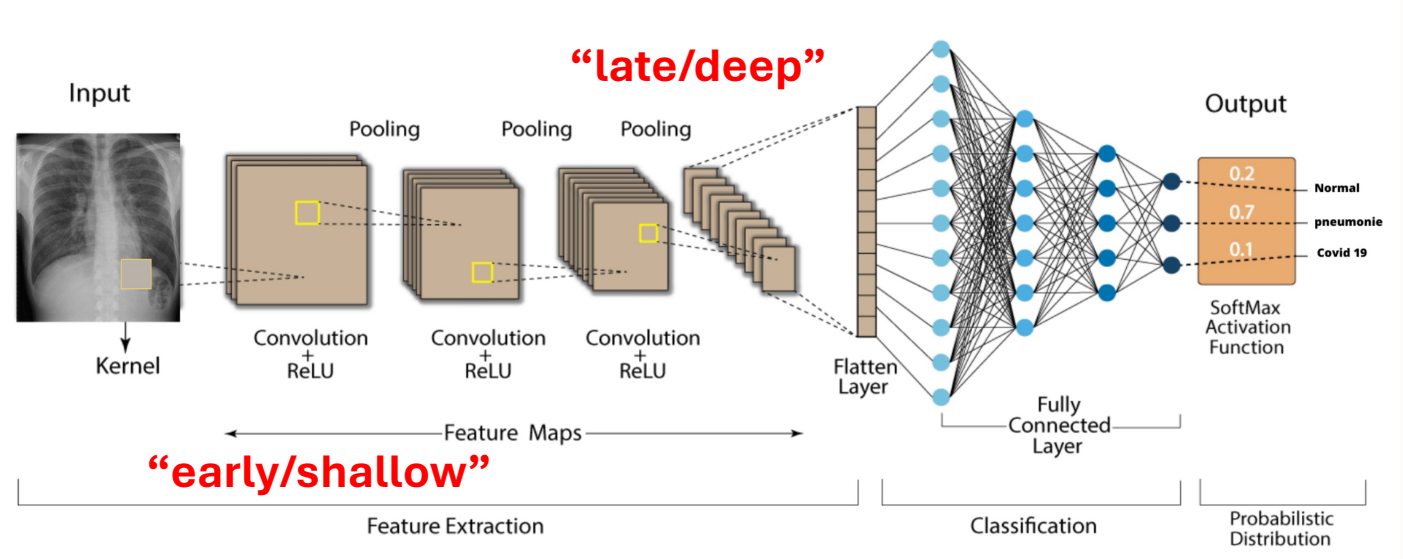
Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012).

Feature maps

What the model is ‘seeing’

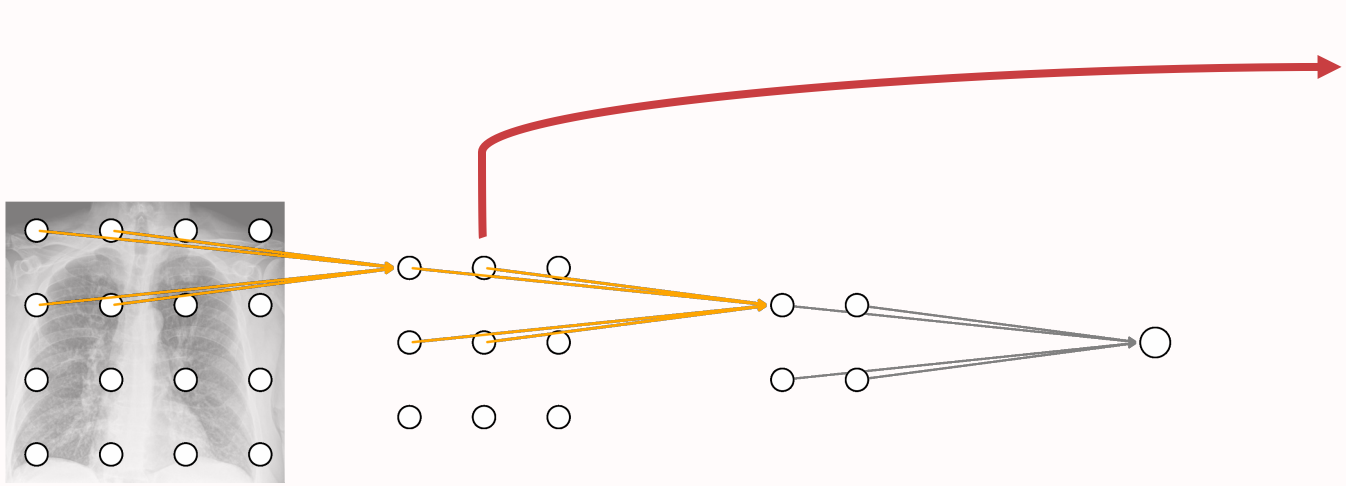
What does the model see?

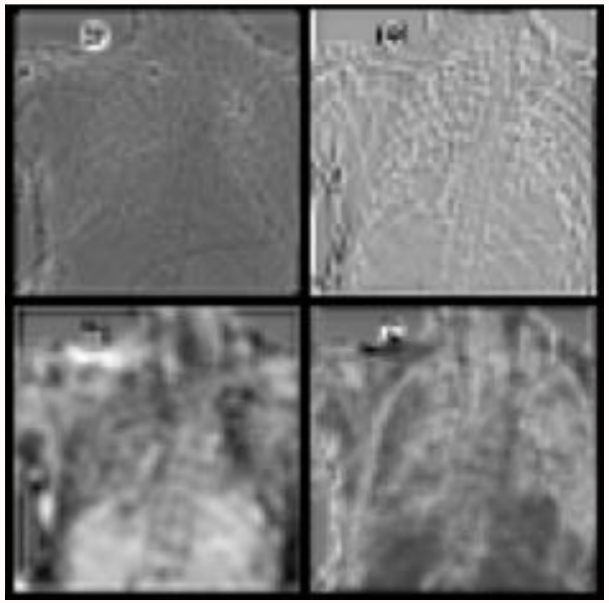
- Early layers detect simple features:
  - Edges
  - Lines
  - Brightness
- Later layers flexibly combine features into abstract concepts
  - Circles, triangles, textures…
  - Lesions, nodules, organs…
- Analogy:
  - Early features ⟶ Lego blocks
  - Later features ⟶ Lego cars and buildings
- Construct complex features using limited building blocks.



Shallow/early feature maps

- Also see which neurons activated
- Neuron activated → corresponding feature present
- Shows which part of X-ray has edges/blobs.

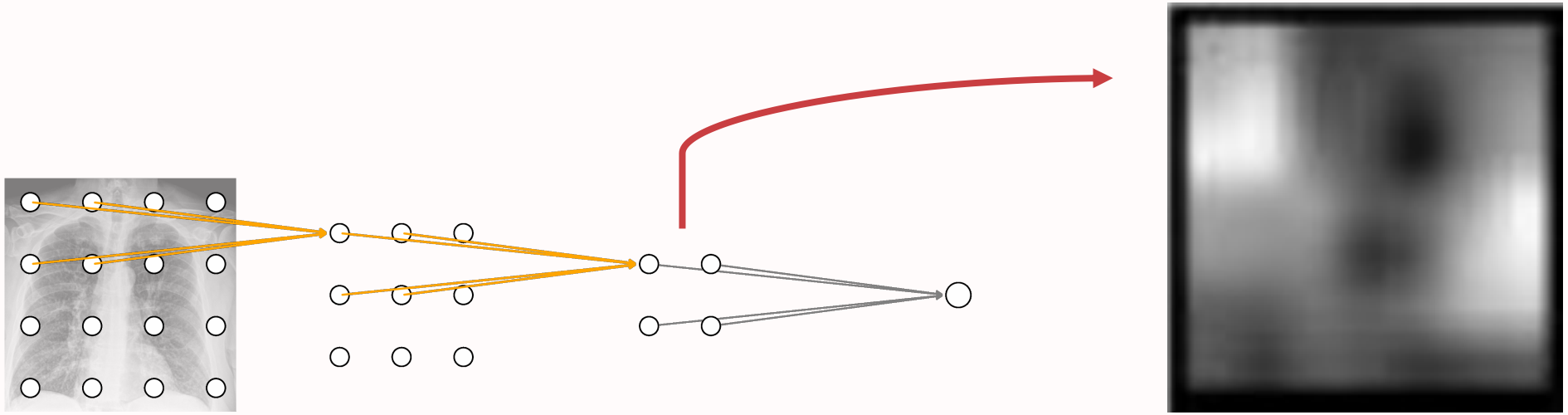




Dalton Griner 2023, The Development and Optimization of a Deep-Learning Strategy for COVID-19 Classification in Chest X-ray Radiography, Ph.D. thesis, University of Wisconsin-Madison

Late/deep feature maps

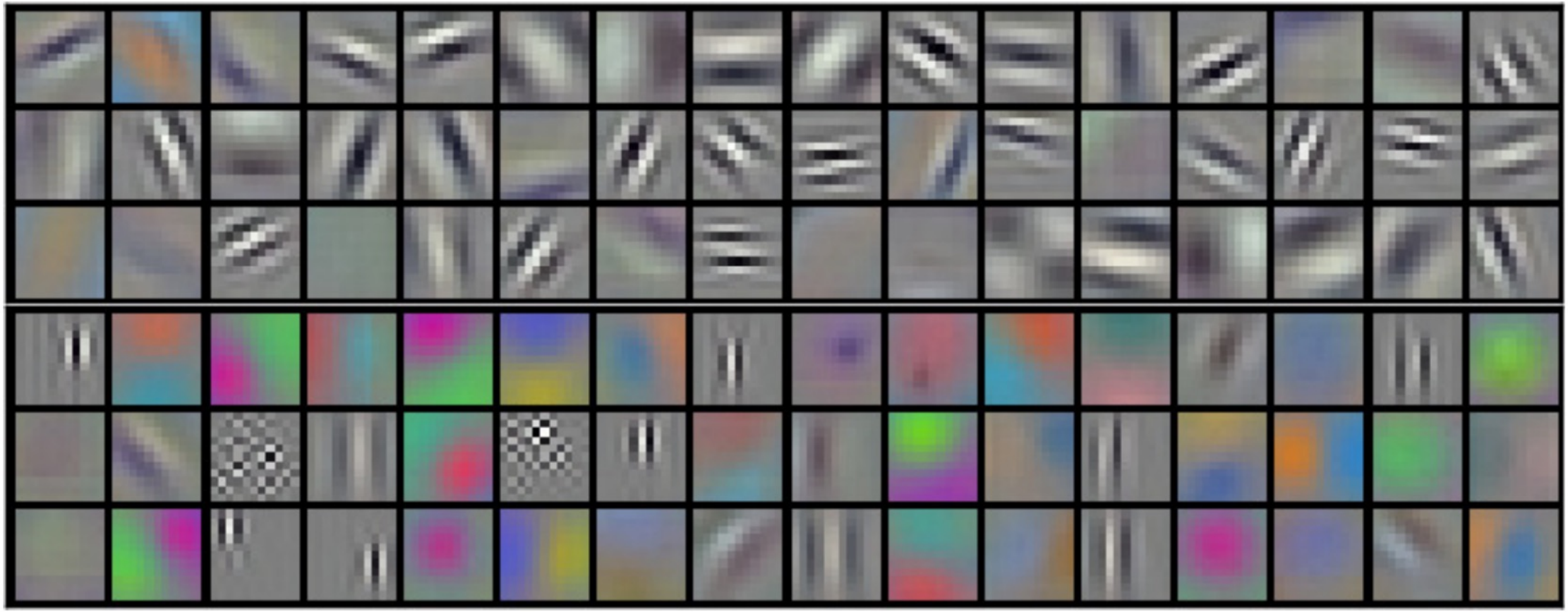
- What about deeper layer activations?
- Fewer neurons, more abstract
- Much more difficult to interpret



Dalton Griner 2023, The Development and Optimization of a Deep-Learning Strategy for COVID-19 Classification in Chest X-ray Radiography, Ph.D. thesis, University of Wisconsin-Madison

Courtesy of Dr. Foong

General example



Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012).

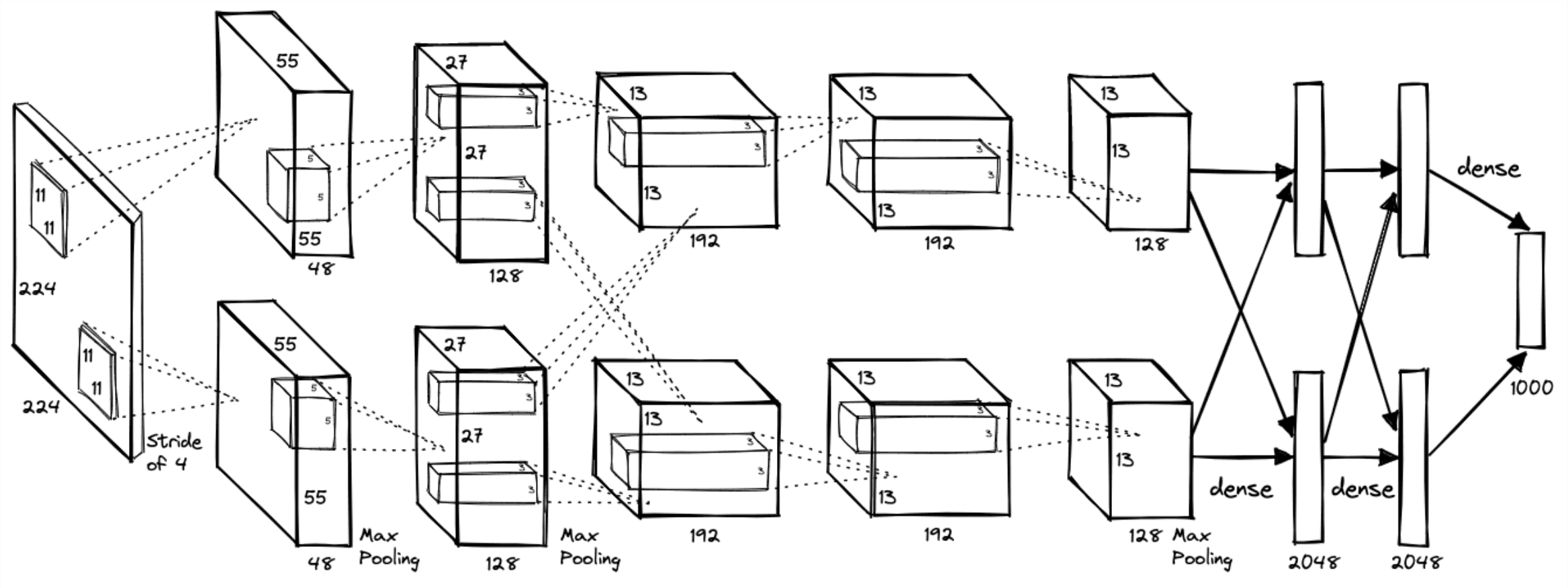
General example

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012).

Milestones for CNNs

Subtle changes, big impact

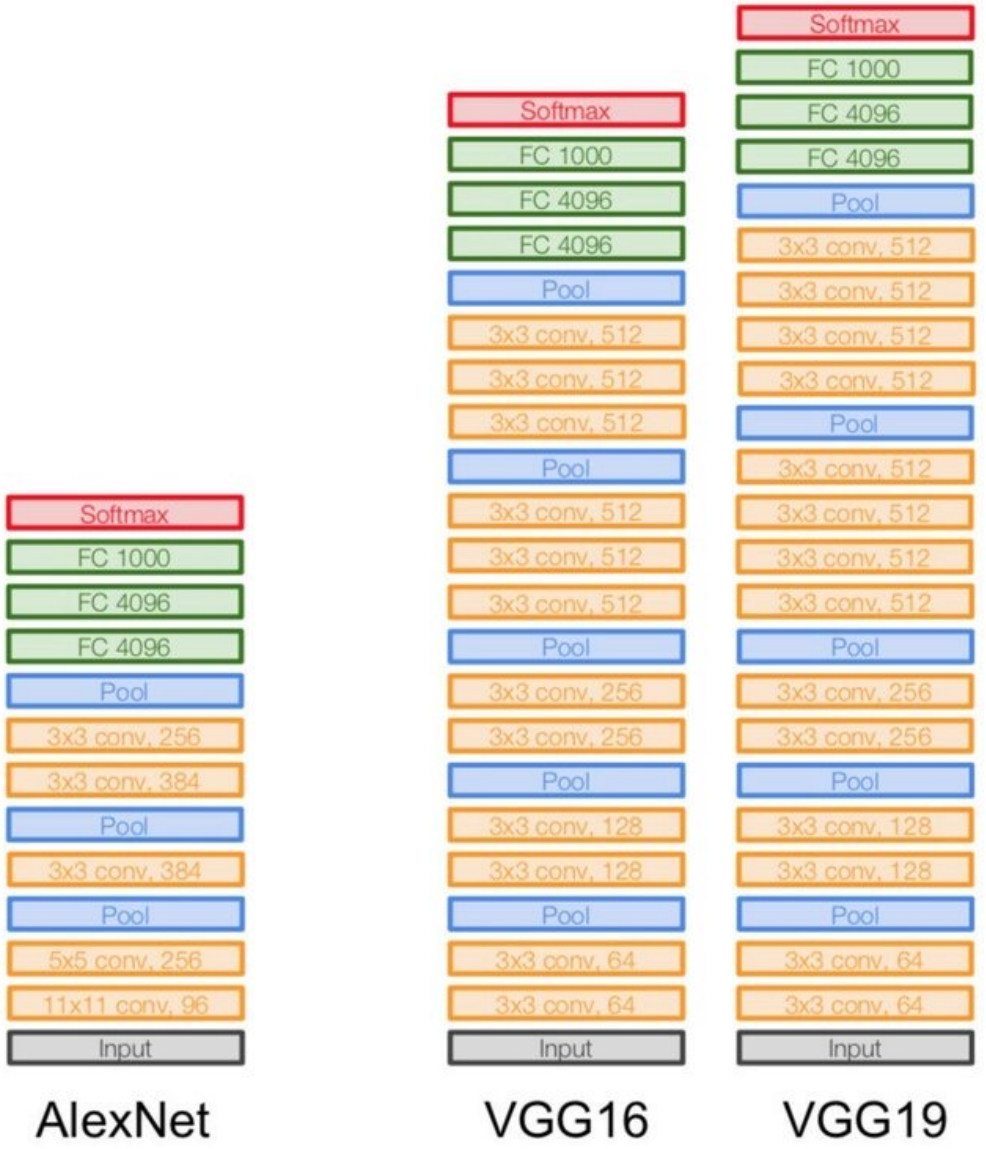
AlexNet ~184k citations



Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012).

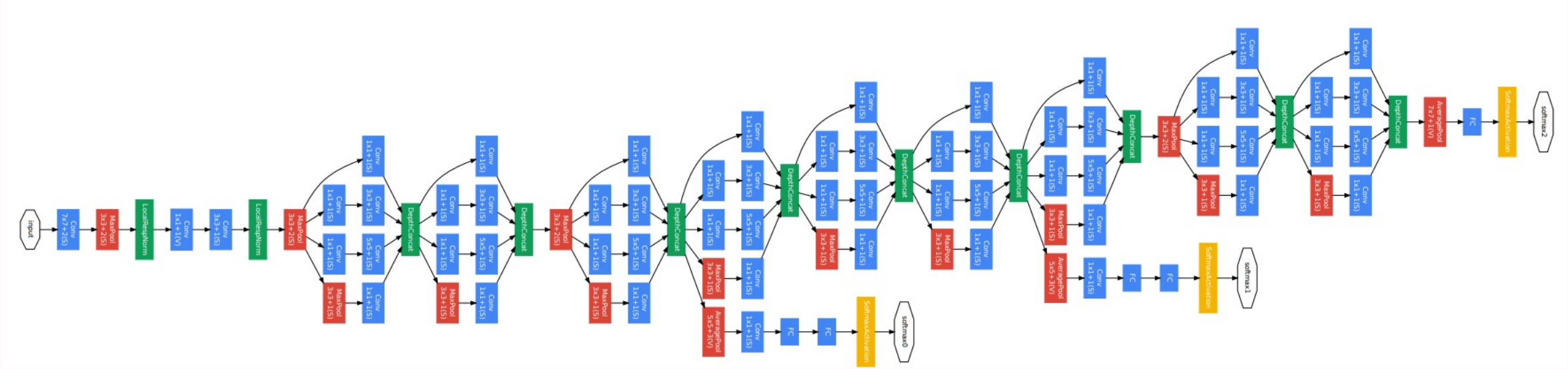
VGG ~145k citations

- Deeper than AlexNet
- Smaller convolutional filters
  - Only 3x3
- Receptive field
  - Achieved through successive convolutions
    - Rather than big filters



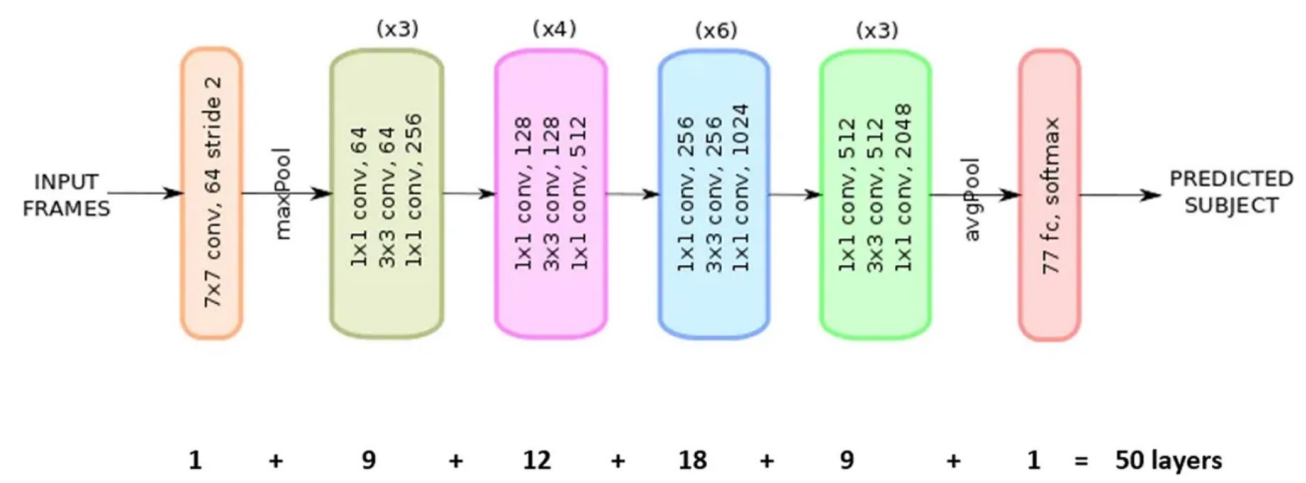
Inception ~94k citations

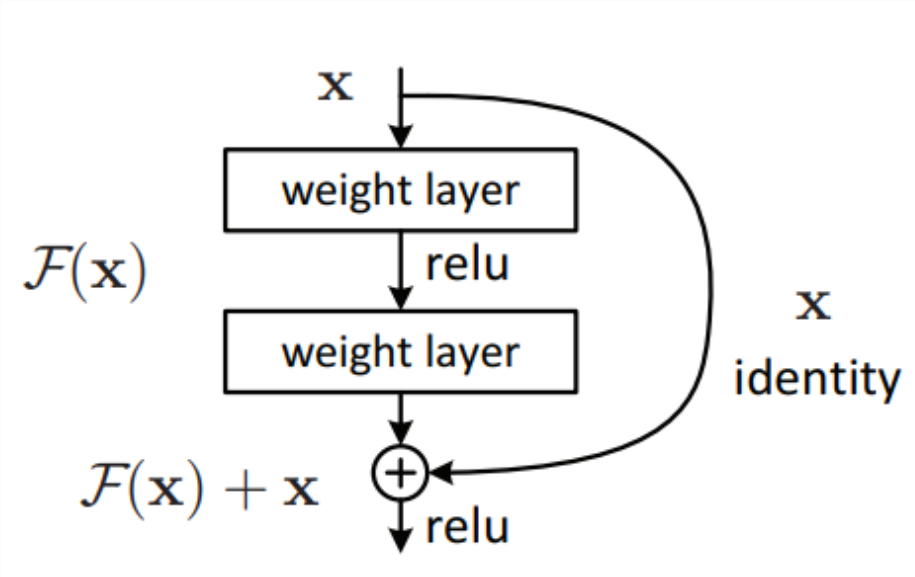
- Multi-scale processing
  - Multiple different filter sizes per layer
  - These are concatenated
- Auxiliary classifiers
  - Helps with vanishing gradient
- Global average pooling
  - Reduced FCN size at end of network



ResNet ~260k citations

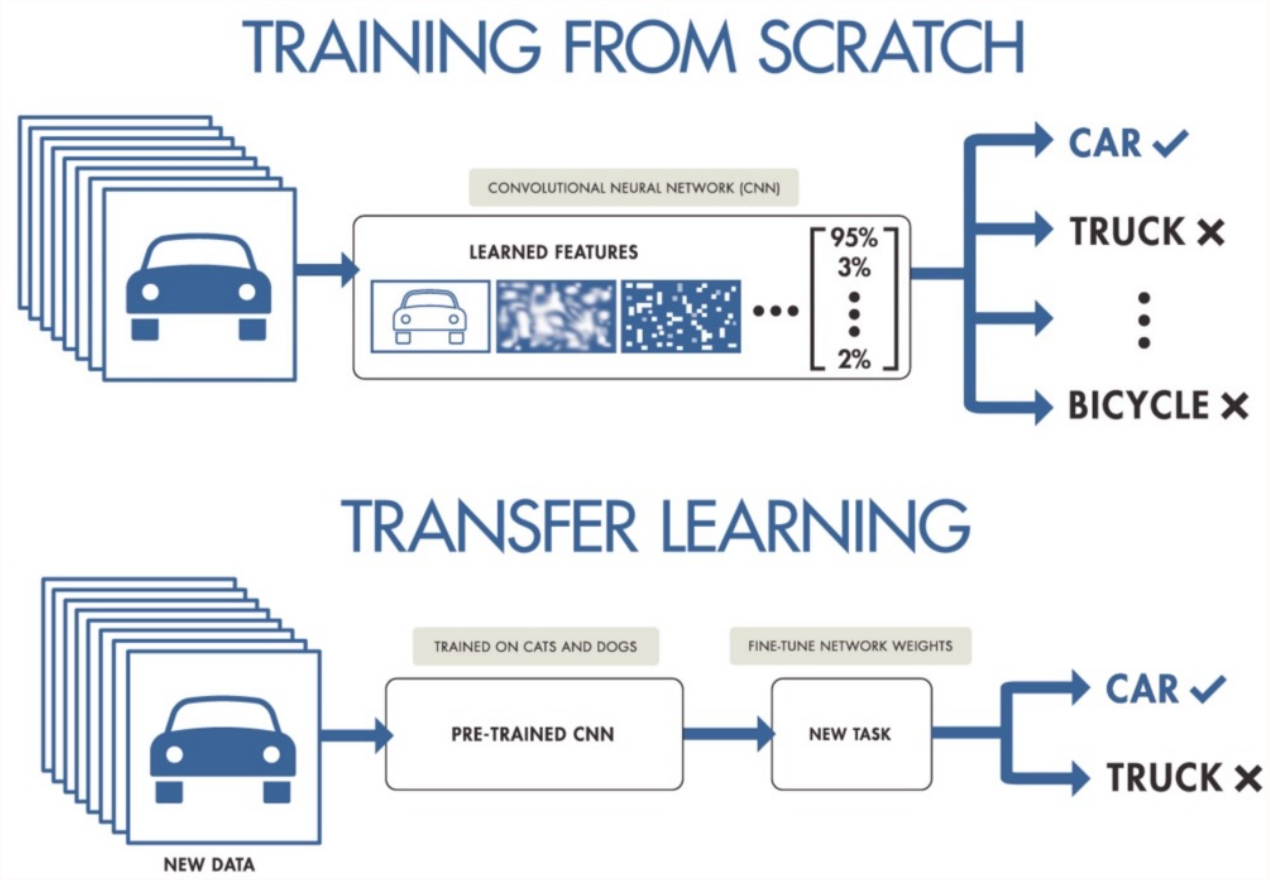
- Deeper (again)
- Residual connections
  - Solved vanishing gradient





Transfer learning

- Reusing a pre-trained model on a new, related task instead of building a model from scratch
- Saves significant time, computational resources, and data, especially when new datasets are small
- A model that has learned features on a large, general dataset can apply that "knowledge" to a specific problem



Many more

- R-CNN (2014)
- Efficiency gains
  - SqueezeNet (2016)
  - MobileNet (2017)
- DenseNet (2017)
- EfficientNet (2019)

Visualization

How to explain predictions

Feature maps

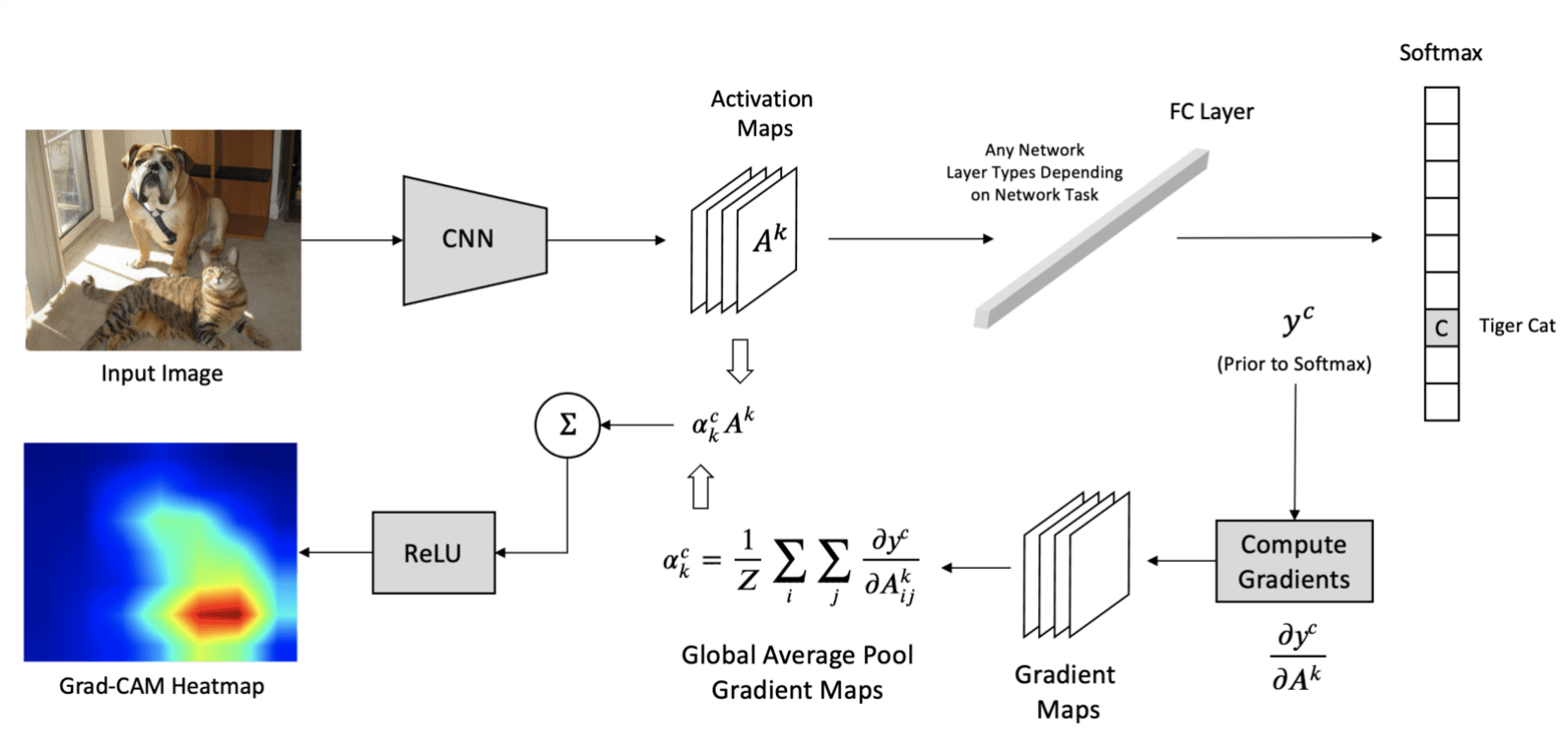
- Feature maps are good, but
  - They don’t tell us *why* a prediction is made
  - i.e. we want to make models *explainable*

Feature maps

- Feature maps are good, but
  - They don't tell us *why* a prediction is made
  - i.e. we want to make models *explainable*
- Gradient-based methods
- Perturbation-based methods

Gradient-based methods

- GradCAM
  - Class activation maps
  - Computes the gradients of the target class probability
  - Calculate "importance weights" corresponding to the last layer's feature map

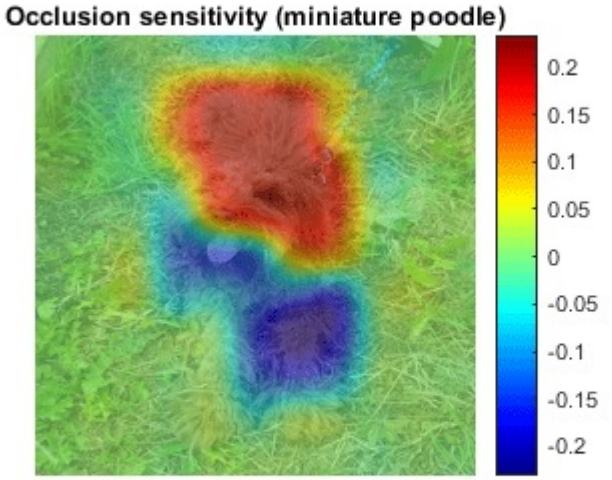


Gradient-based methods

- GradCAM
  - Class activation maps
  - Computes the gradients of the target class probability
  - Calculate "importance weights" corresponding to the last layer's feature map
- Grad-CAM++
- Score-CAM
- Ablation-CAM
- Eigen-CAM
- Many more

Perturbation-based methods

- Occludes or hides different parts of the input image
- Measures the impact on the classification
- Regions that cause a significant drop in confidence are considered important

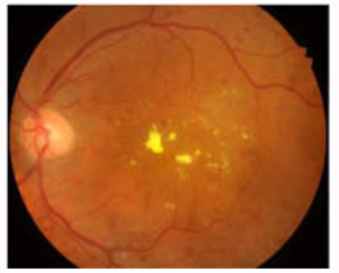


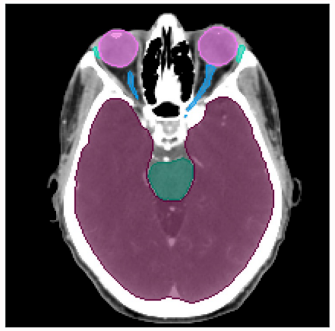
More applications!

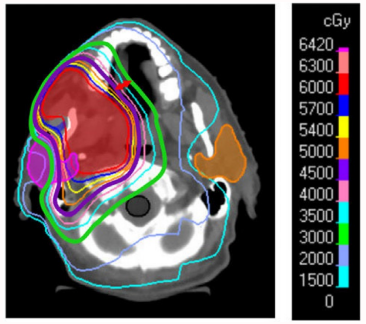
Thousands more

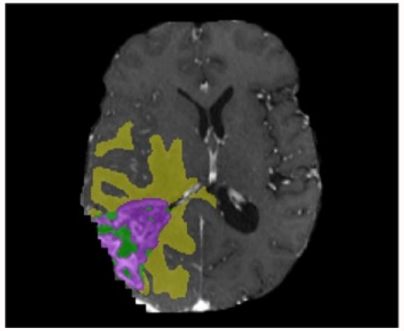
CNN applications

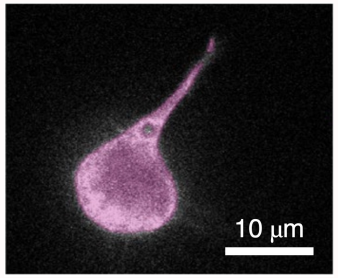
1. Image to single label:
   - Skin lesions.
   - Diabetic retinopathy.
   - Etc.
2. Image to image:
   - Organs in CT (RadOnc Google model).
   - Optimal dose prediction.
   - Edema, tumor, necrosis in brain MRI.
   - Cancer cells in fluorescence microscopy.
   - Synaptic clefts in electron microscopy.
   - Etc.

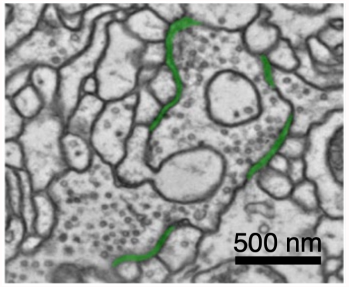










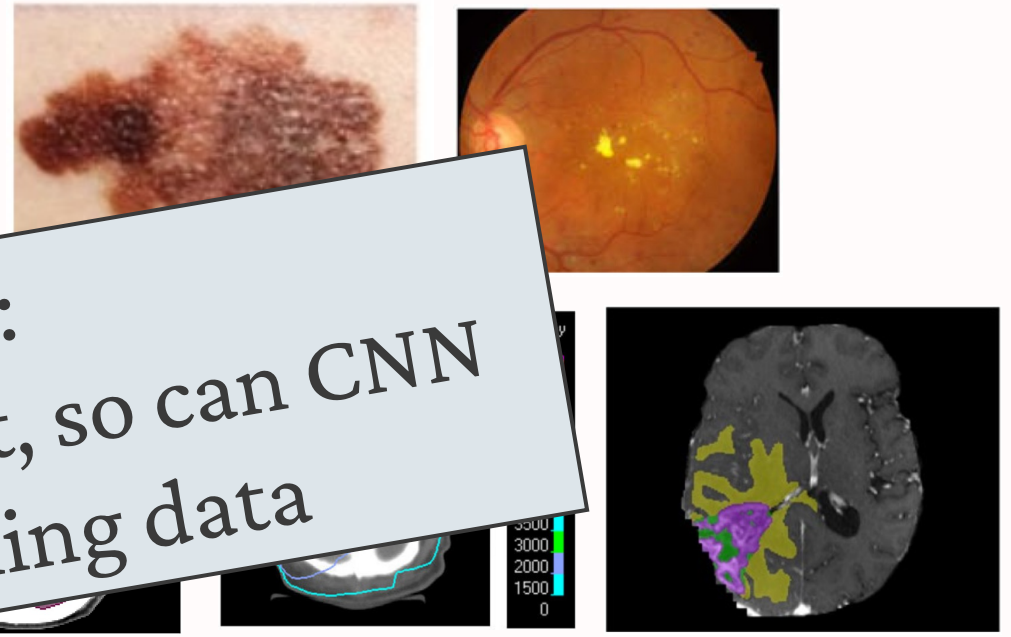


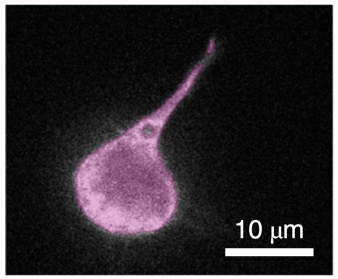
Esteva et al. 2017, Dermatologist-level classification of skin cancer with deep neural networks, Nature.
Nikolov et al. 2021, Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy, arXiv preprint.
Gronberg et al. 2023, Deep Learning–Based Dose Prediction for Automated, Individualized Quality Assurance of Head and Neck Radiation Therapy Plans, Pract Radiat Oncol.
Dai et al. 2021, A deep learning system for detecting diabetic retinopathy across the disease spectrum, Nature Communications.
Isensee et al. 2020, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nature Methods.

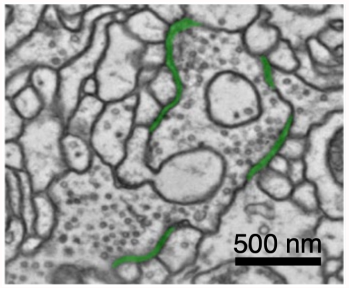
CNN applications

1. Image to single label:
   - Skin lesions.
   - Diabetic retinopathy.
   - Etc.
2. Image to image:
   - Organs in
   - Optimal d
   - Edema, tur
   - Cancer cells
   - Synaptic cle
   - Etc.

Rule of thumb:
If human experts can do it, so can CNN with "enough" training data



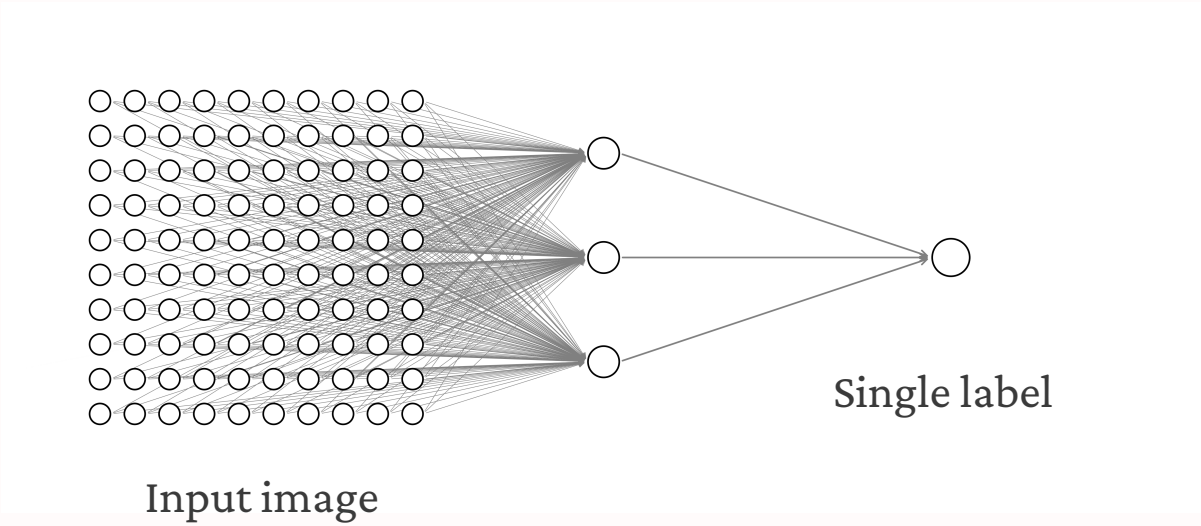




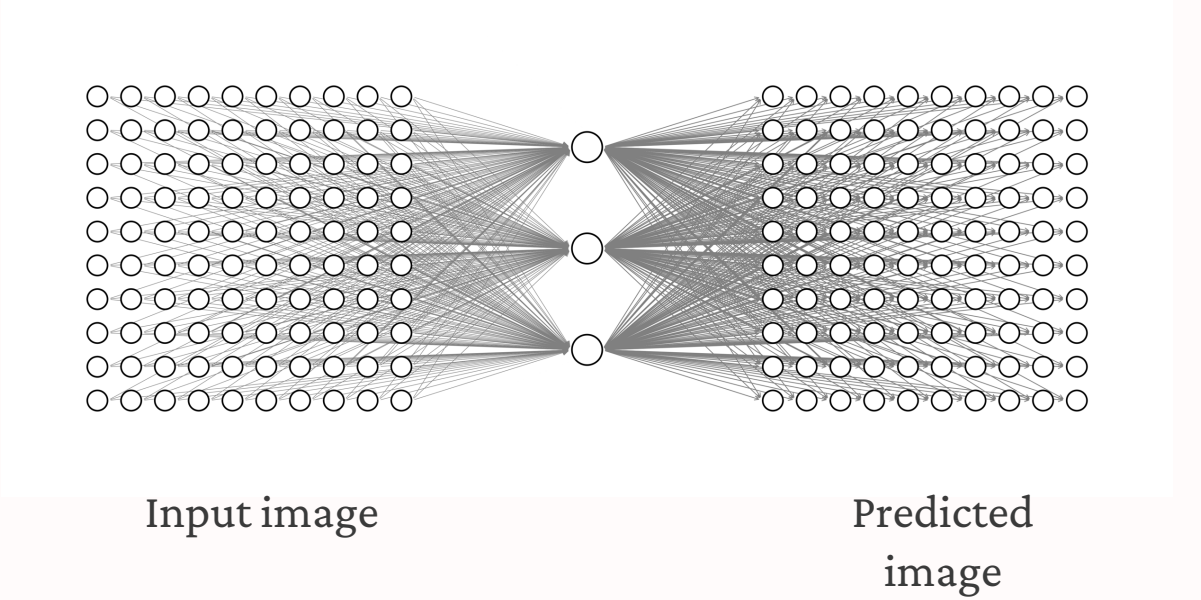
Esteva et al. 2017, Dermatologist-level classification of skin cancer with deep neural networks, Nature.
Nikolov et al. 2021, Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy, arXiv preprint.
Gronberg et al. 2023, Deep Learning–Based Dose Prediction for Automated, Individualized Quality Assurance of Head and Neck Radiation Therapy Plans, Pract Radiat Oncol.
Dai et al. 2021, A deep learning system for detecting diabetic retinopathy across the disease spectrum, Nature Communications.
Isensee et al. 2020, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nature Methods.

CNN applications

1. Image to single label:



2. Image to image:



Question & Answer