Convolutional Neural Networks: What Even Are They?

By Parker

Overview

- Used with grid-like data
  - Images (2D)
  - Time Series (1D)
- Special case of standard feedforward network w/ convolution instead of general matrix multiplication in 1+ layer(s)

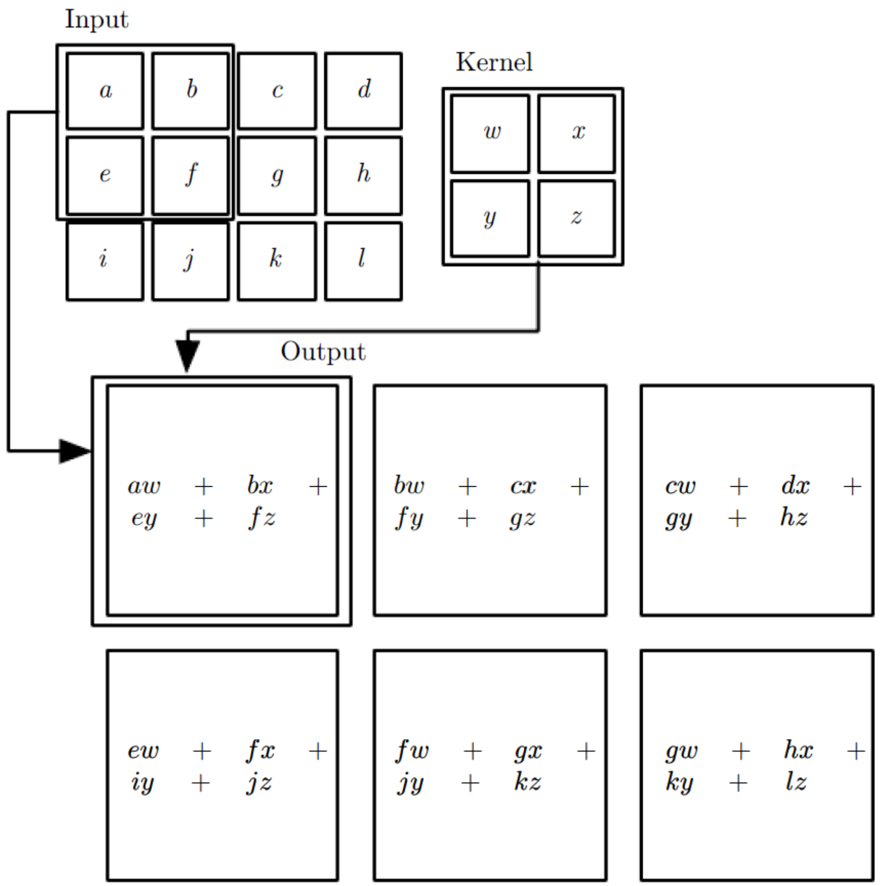
Convolution

- For real-valued convolution we have:
  - $(x * w)(t) = \int x(a)w(t-a)da$
  - *x* is the input, *w* is the kernel, output is the feature map
- For discrete contexts:
  - $(x * w)(t) = \sum_{a=-\infty}^{\infty} x(a)w(t-a)$

Convolution, 2D

- In 2D discrete contexts, we want a 2D kernel K and image I
  - $S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m,j-n)$
- More straightforward:
  - $S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i-m,j-n)K(m,n)$
- Cross-correlation (usual implementation):
  - $S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i+m,j+n)K(m,n)$
- Expressible as matrix multiplication, but with many entries identical (Toeplitz in 1D, doubly block circulant in 2D)

Convolution, 2D Example

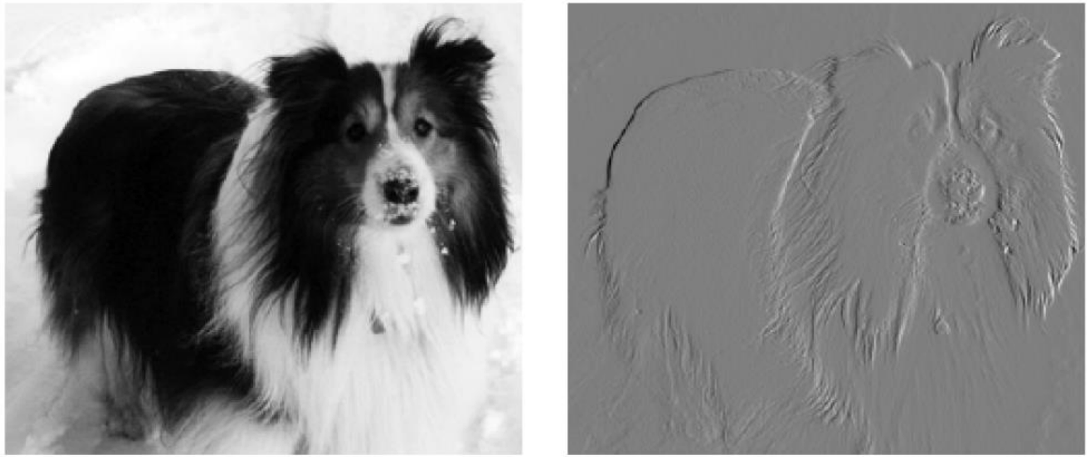


Motivation for CNNs: Why Fake News?

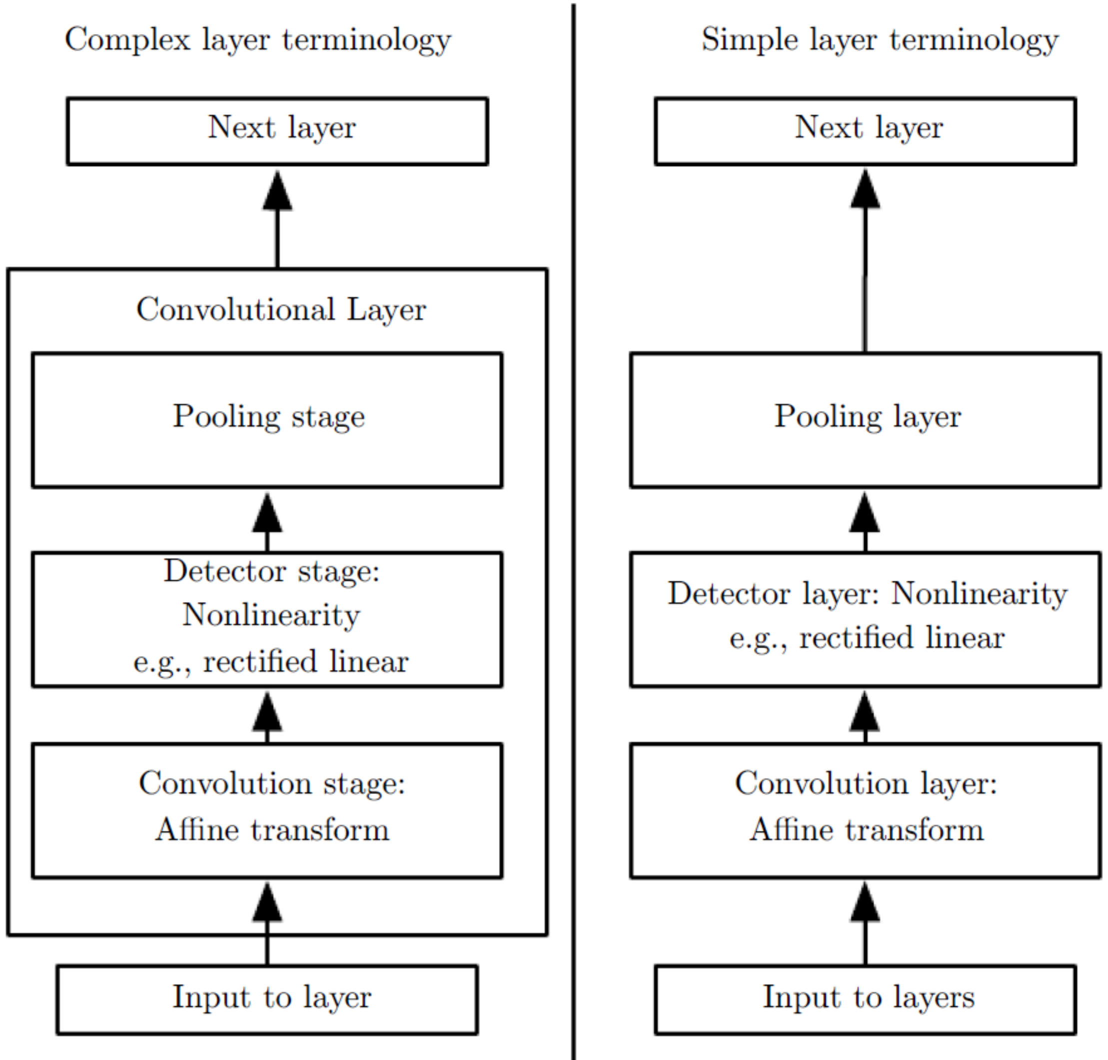
- Sparse interactions
  - Kernel usually much smaller than input
  - Significant reduction in memory and time requirements
- Parameter sharing
  - Same kernel weights used at every image position (expect boundary)
  - Reduces memory/storage requirement
- Equivariant to translation
  - Shifting image shifts output in same way

Efficiency Example

- Each pixel value had left neighbor subtracted
- 280x320 -> 280x319
- As convolution, 319*280*3 = 267,960 float ops (two mult, one add)
- Naïve matrix: over 8 billion ops
- Sparse matrix: same ops, but 178,640 entries (vs. 2 for convolution)



Network Overview



Pooling

- Produces at each position a summary of the nearby outputs from previous layer/stage
  - Max pooling, average, L2 norm, weighted average w/ center distance
- Approximates invariance to small translation
  - Useful if context is identification, not precise location
- Pooling over multiple features allows learned invariance
  - Example: rotation invariance for handwriting (pool over features detecting digit in various orientations)
- Spacing out pooling windows (stride) reduces network width in later layers
  - Further reduced parameter count
- Variable pooling window allows for variable network input size but constant input size to later layers
  - Summarize each quadrant -> 4 values
  - Enables using CNNs for inputs that cannot be processed by traditional matrix-multiplication networks

Other Details

- Multiple convolutions in parallel (multiple feature maps)
- Zero-pad input
  - Otherwise, each convolution shrinks width by 1 less than kernel width
  - Valid, same, full
- Channels
  - Images often have 3 channels (RGB); can be processed independently or together
- Variant: unshared convolution
  - Different kernel weights learned at each location
- Variant: tiled convolution
  - Finite set of kernels, “tiled” across input

Output

- For images, CNNs can provide per-pixel evaluations
  - Example: probability of pixel being part of object X
- Commonly, last convolutional layer's output fed to fully-connected layer(s)

Optimization

- If *d*-dimensional kernel equivalent to outer product of *d* 1D kernels, naïve convolution is inefficient (O(w^d) time and space; *w* is width of every kernel dimension)
    - Composing *d* 1D convolutions takes O($w \times d$) time and space
- Active research area

Untrained Kernels

- Hand-designed
  - Edge detection is easy
- Random initialization
  - Empirically better than Bogosort when fed into trainable fully-connected layer(s)
  - Enables easier search over architectures
- Clustering on small patches
  - Popular 2007-2013 (smaller datasets, less computational power)

Training

Neuroscientific Basis

- Work in the 60s studied individual neuron activity in cats when looking at certain images
  - Observed inspiration for low-level feature detectors
- CNNs modeled after V1 (primary visual cortex)
  - V1 has 2D orientation corresponding to retina
  - Simple cells: linear function of small receptive field
  - Complex cells: like simple cells, but have some invariance to small translation (pooling) and lighting (cross-channel pooling)

Neuroscientific Basis Cont’d

- Later convolutional layers correspond to “grandmother cells”
  - Individual neurons that activate when seeing one’s grandmother
  - Verified for recognizing many famous individuals (“Halle Barry neuron”)
    - More sophisticated than CNNs: also trigger on drawings and text
- NOT backpropagation
  - Neuroscience has given little on how to train
- Simple cells act roughly like Gabor functions
  - Lower layers seem to be edge detectors
  - Initial layers in most deep learning methods on images learn this behavior

Gabor Functions