STAT562 Lecture 18 Convolutional Neural Network

Beidi Qiang

SIUE

Convolutional neural network (CNN) is an advanced architectures of artificial neural network (ANN) and it is specifically used in image classification.

- Similar to ordinary Neural Networks: they are made up of hidden layers and units that have learnable weights.
- Convolutional neural network has two specialized types of hidden layers, called convolution layers and pooling layers.
- Convolution layers search for instances of small patterns in the image
- Pooling layers downsample these instances of patterns to select a prominent subset.

Convolution layer

A convolution layer is made up of a large number of convolution filters (activation units).

- A filter is a $l_1 \times l_2$ array of weights that is operated to the original image data.
- The operation is called convolution, which basically repeatedly multiplying matrix elements to every $l_1 \times l_2$ submatrix of the original image and then adding up.

$$\text{Original Image} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \\ j & k & l \end{bmatrix}.$$

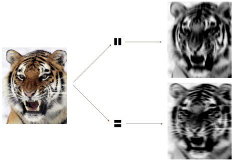
Now consider a 2×2 filter of the form

$$\text{Convolution Filter} = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}.$$

When we *convolve* the image with the filter, we get the result[8]

$$\text{Convolved Image} = \begin{bmatrix} a\alpha + b\beta + d\gamma + e\delta & b\alpha + c\beta + e\gamma + f\delta \\ d\alpha + e\beta + g\gamma + h\delta & e\alpha + f\beta + h\gamma + i\delta \\ g\alpha + h\beta + j\gamma + k\delta & h\alpha + i\beta + k\gamma + l\delta \end{bmatrix}.$$

- If a sub-matrix of the original image resembles the convolution filter, then it will have a large value in the convolved image; otherwise, it will have a small value.
- Thus, the convolved image highlights regions of the original image that resemble the convolution filter.
- We can think of the original image as the input layer in a convolutional neural network, and the convolved images as the units in the first hidden layer.



More details on colored Image Data

- Format of input data: W x H colored (RGB) pixels per image. Each color channel (red, green, and blue), is a 2-d (W x H) array (called feature map). The dimension of input image is W x H x 3.
- A single convolution filter will also have three channels, one per color, with potentially different filter weights.
- The results of the three convolutions are summed to form a 2-d output feature map (color is not passed on to subsequent layers).
- The dimension of output activations in the 1st convolution layer is W_1 x H_1 x K, where K is the number of filters.

Depth, stride and zero-padding

Three hyperparameters control the size of the output in each convolution layer: depth, stride and zero-padding.

- The depth (K) of the output volume is the number of filters we would like to use.
- The stride (usually $S = 1$ or 2) defined how we slide the filter. When the stride is 1 then we move the filters one pixel at a time. When the stride is 2 then the filters jump 2 pixels at a time.
- Zero-padding (P) is the amount of 0 added to the outside of the feature map, which allows us to control or preserve the spatial size of the input volume.

Detector layer

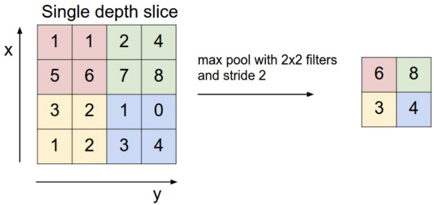
We apply the ReLU activation function to the convolved image from the convolution layer.

- This step is often viewed as a separate layer in CNN, called detector layer or ReLU layer.
- This layer simply threshold at zero and dimension of the information is unchanged.

Pooling Layers

A pooling layer condense a large image into a smaller summary image. The layer reduce the amount of parameters and computation in the network, and hence to also control overfitting.

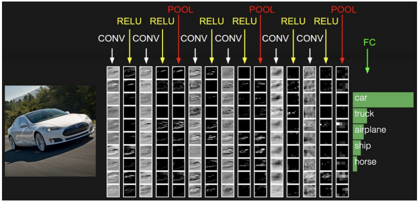
- The Pooling layer operates independently on every slice of the input, using the MAX operation.
- The most common form is a max pooling of size 2 x 2, taking a max over 4 numbers (little 2 x 2 region).
- This reduces the size of the image by a factor of two in each direction, discarding 75%.



Architecture of a Convolutional Neural Network

In a convolutional Neural Network, the convolve-then-pool sequence is often repeated several times

- Each subsequent convolutional layer is similar to the first, with number of channels same as the number of filters from previous convolution layer.
- Since the feature maps are reduced in size after each pool layer, we usually increase the number of filters in the next convolve layer to compensate.
- Sometimes we repeat several convolve layers before a pool layer.



Fully-connected layer

After a set of convolve-then-pool, each channel feature map down to just a few pixels in each dimension.

- We "flattened" the feature maps by treating each of the pixels as separate units.
- We then feed those into a fully connected layer just as one in a regular ANN.
- The output is fed into the output layer, which is has softmax activation for image classes.

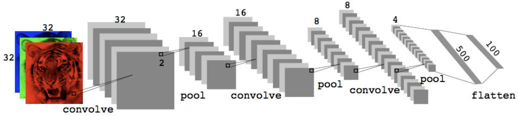


FIGURE 10.8. *Architecture of a deep CNN for the* CIFAR100 *classification task. Convolution layers are interspersed with* 2×2 *max-pool layers, which reduce the size by a factor of 2 in both dimensions.*

Many data sources are sequential in nature:

- Documents such as book, articles, and tweets. The sequence of words in a document needs to be exploited in tasks such as topic classification, and language translation.
- Time series of temperature, rainfall, wind speed, air quality, and so on.
- Financial time series, where we track market indices, trading volumes, stock and bond prices
- Recorded speech, musical recordings, and other sound recordings.

- Input: a sequence $X = \{X_1 \cdots X_L\}$
- Sequence of hidden layers: $A = \{A_1 \cdots A_L\}$
- The network processes the input sequence X sequentially.
- At each step, the network updates the activations A_l in the hidden layer, taking as input the vector X_l and the activation vector A_{l-1} from the previous step.
- Then A_l feeds into the output layer and produces a prediction O_l.

Note here O_L is the final output. Intermediate outputs O_l are not used.