1 Definitions Ahoy!

1.1 MLE/MAP

1. **MLE:** Finds the best parameters for a specific dataset, $\mathcal{D}$. Specifically, we want to find the parameters $\hat{\theta}_{MLE}$ that maximize the likelihood for $\mathcal{D}$.

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}\, p(\mathcal{D} \mid \theta)$$

2. **MAP:** Finds the best parameters given $\mathcal{D}$ and a prior belief about the parameters. Specifically, we want to find the parameters $\hat{\theta}_{MAP}$ that maximize the posterior distribution $p(\theta \mid \mathcal{D})$ of parameters θ.

$$\begin{aligned}
\hat{\theta}_{MAP} &= \underset{\theta}{\operatorname{argmax}}\, p(\theta \mid \mathcal{D}) \\
&= \underset{\theta}{\operatorname{argmax}}\, \frac{p(\mathcal{D} \mid \theta)p(\theta)}{\text{Normalizing Constant}} \\
&= \underset{\theta}{\operatorname{argmax}}\, p(\mathcal{D} \mid \theta)p(\theta) \\
&= \underset{\theta}{\operatorname{argmin}} - \log\left(p(\mathcal{D} \mid \theta)p(\theta)\right) \\
&= \underset{\theta}{\operatorname{argmin}} - \log p(\mathcal{D} \mid \theta) - \log p(\theta)
\end{aligned}$$

3. **MLE and MAP for conditional likelihood:** When we want to predict the output y given the input x using our supervised dataset, we have to reformulate the MLE and MAP optimizations to use the conditional likelihood (and conditional posterior) instead:

$$\begin{aligned}
\hat{\theta}_{MLE} &= \underset{\theta}{\operatorname{argmax}}\, p(\mathcal{D} \mid \theta) \\
&= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} p\left(y^{(i)} \mid x^{(i)}, \theta\right) \\
&= \underset{\theta}{\operatorname{argmin}} - \log \prod_{i=1}^{N} p\left(y^{(i)} \mid x^{(i)}, \theta\right) \\
&= \underset{\theta}{\operatorname{argmin}} - \sum_{i=1}^{N} \log p\left(y^{(i)} \mid x^{(i)}, \theta\right)
\end{aligned}$$

$$\begin{aligned}
\hat{\theta}_{MAP} &= \underset{\theta}{\operatorname{argmax}}\, p(\theta \mid \mathcal{D}) \\
&= \underset{\theta}{\operatorname{argmax}} \left(\prod_{i=1}^{N} p\left(y^{(i)} \mid x^{(i)}, \theta\right)\right) p(\theta) \\
&= \underset{\theta}{\operatorname{argmin}} - \log \prod_{i=1}^{N} p\left(y^{(i)} \mid x^{(i)}, \theta\right) - \log p(\theta) \\
&= \underset{\theta}{\operatorname{argmin}} - \sum_{i=1}^{N} \log p\left(y^{(i)} \mid x^{(i)}, \theta\right) - \log p(\theta)
\end{aligned}$$

1.2 CNNs

Convolution Neural Network: A type of neural network that is particularly well-suited for image and video processing tasks. CNNs consist of the following building blocks:

1. **Convolutional Layer:** Involves sliding a small window of fixed size across the input data and computing the dot product between the kernel and the overlapping portion of the input at each position. The result of each dot product is then summed to produce a single output value, which is stored in a new output.
   (a) **Kernel/Filter:** A small multidimensional array of weights that is used to scan over the input data during the convolution operation
   (b) **Stride:** The number of pixels or steps by which a filter is moved across the layer input data
   (c) **Channel:** An additional dimension within a layer of the network that represents a specific feature or attribute of the data. For example, an RGB image has three channels that correspond to the red, green, and blue color channels. The output of a convolutional layer usually has many different channels to allow the network to learn to detect different patterns in each channel. For example, the output of the first convolutional layer may learn to detect edges at different angles in each channel.
   (d) **Padding:** Extra space added to the outsides of the layer input (filled with zeros or a copy of the data on the border) to make the output dimensions have the desired size.
2. **Pooling Layer:** Used to reduce the spatial dimensions of the data. Two popular types of pooling are max pooling and average pooling.
   (a) **Max pooling:** Similar to a convolution, but in this case, the maximum value at each kernel location is selected. Analogous to picking the brightest pixel in a grid of small regions of the image and tossing the rest.
   (b) **Average pooling:** The average value of all the pixels at each kernel location is selected. Analogous to smoothing out an image while reducing its size.
3. **Fully-connected layer:** Typically a CNN will end with one or more fully-connected layers. These layers effectively perform classification or regression based on the features coming from the convolutional layers.

2 CNNs and not the kind on TV

2.1 Conceptual Questions

1. What are some benefits of CNNs over fully connected (also called dense or linear) layers?

   CNNs are good for image-related machine learning tasks because they learn the kernels that do feature engineering. Additionally, CNNs, through the use of convolutions and pooling, take advantage of local spatial coherence. This refers to the tendency of neighboring pixels in an image to have similar values or characteristics. CNNs are also parameter efficient.

2. How are the weights in a CNN updated during training?

   Backpropogation, the same as fully connected layers.

3. What are the parameters of a convolutional layer?

   The values in the kernels (a.k.a. weights or filters) and the bias term per output channel. The kernel sizes are num_input_channels by kernel_width by kernel_height for each output channel.

4. What are the hyperparameters of a convolutional layer?

   Stride size, padding size, filter size.

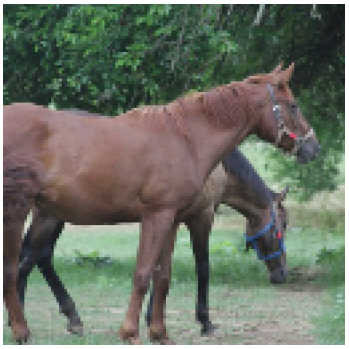
5. Can convolutions only be applied to 2-D shapes like images?

   Nope. Convolutions are very common on 1-D signals like sound or a stock market time series. You can also combine 2-D space with time and do convolutions for video. Or even higher dimensional data like 4-D cardiac MRI data.

2.2 Shapes and Parameters

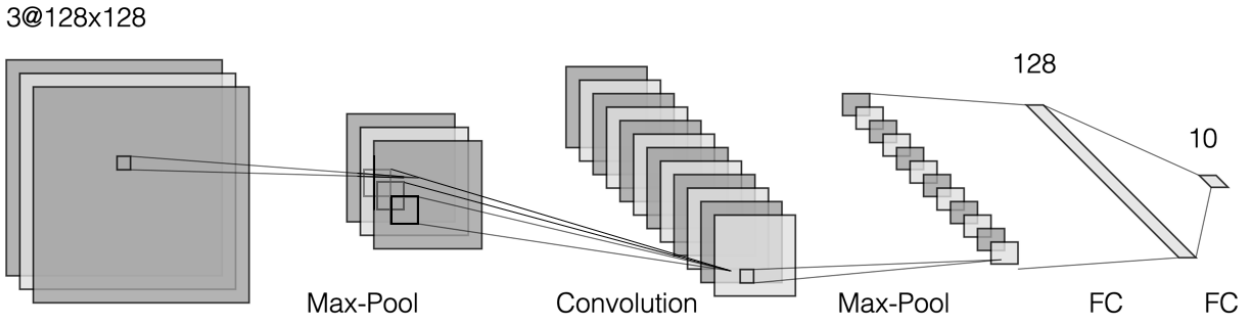
Torch the Triceratops is working on his 10-315 final project and is looking into the MS-COCO dataset. His goal is to classify images that belong to one of ten possible classes (i.e. `[cat, dog, bird, turtle, ..., horse]`). The images come in RGB format (one channel for each color) and are downsampled to dimension 128×128.

The following illustrates one such image from the MS-COCO dataset.



Torch constructs a convolutional neural network that has the following structure: the input is first max-pooled (not common) with a 2×2 filter with stride 2. The results are then sent to a convolutional layer that uses a 17×17 filter (uncommonly large) with stride 1 and 12 output channels. Those values are then passed through a max-pooling layer with a 3×3 filter with stride 3. The result is then flattened and passed through a fully-connected layer (with ReLU activation, not shown) with 128 hidden units, followed by a fully connected layer (softmax activation, not shown) with 10 hidden units. The final 10 hidden units represent the categorical probability for each of the ten classes. With enough labeled data, Torch plans on using some optimizer like SGD to train this model using backpropagation.

Note: Assume we have bias terms in all neural network layers unless explicitly stated otherwise.



1. What is the shape of the output tensor after max pooling?

   With a stride of two the size is cut in half in each direction. Unlike convolution, pooling is applied independently per channel.
   $3 \times 64 \times 64$

2. What is the shape of the output tensor after the convolutional layer?

   Thinking in just one dimension, we can fit a window of width 17 within the larger width of 64 and then shift it by one 47 more times before it hits the right edge.
   $12 \times 48 \times 48$

3. What is the shape of the output tensor after the second max pooling layer?

   $12 \times 16 \times 16$

4. What is the general equation for the shape of the output tensor if a filter of size $F \times F$ with stride S and padding P is applied to a tensor with width W and height H?

   output width $= \frac{W+2P-F}{S} + 1$

   output height $= \frac{H+2P-F}{S} + 1$

5. How many parameters are in this network for the first max pooling layer?

   Zero. Max-pooling layers do not have any parameters, since a fixed operation is performed on the input.

6. How many parameters are in this network for the convolutional layer?

   $12 \times (3 \times 17 \times 17) + 12 = 10416$

7. How many parameters are in this network for both fully connected layers?

   $(3072 \times 128 + 128) + (128 \times 10 + 10) = 394634$

8. From these parameter calculations, what can you say about convolutional layers and fully connected layers in terms of parameter efficiency (the ratio between the number of parameters from some layer type and the total number of parameters)? Why do you think this is the case?

Total number of parameters = 10416 + 394634 = 405050

Percent from convolutional components = $\frac{10416}{405050} = 2.57\%$

Percent from fully connected components = $\frac{394634}{405050} = 97.43\%$

Convolutional layers are much more parameter efficient, mainly because we are reusing the convolutional filter repeatedly for each convolutional layer (we only need to train one kernel per channel per layer). In comparison, the fully connected layer requires all nodes between two layers to be fully connected.

3 Anybody have a MAP?

Imagine you are a data scientist working for an advertising company. The advertising company has recently run an ad and they want you to estimate its performance. The ad was shown to N people. $y^{(i)} = 1$ if person i clicked on the ad and 0 otherwise. Thus $\sum_i^N y^{(i)} = N_1$ people decided to click on the ad. Assume that the probability that the i-th person clicks on the ad is ϕ and the probability that the i-th person does not click on the ad is $1 - \phi$.
Note

$$p(\mathcal{D} \mid \theta) = p\left(y^{(1)}, y^{(2)}, ..., y^{(N)} \mid \phi\right) = \prod_{i=1}^{N} p\left(y^{(i)} \mid \phi\right) = \phi^{N_1}(1-\phi)^{N-N_1}$$

1. Calculate $\hat{\phi}_{MLE}$.

$$\begin{aligned}
\hat{\phi}_{MLE} &= \underset{\phi}{\text{argmin}} - \log\left(p(\mathcal{D} \mid \phi)\right) \\
&= \underset{\phi}{\text{argmin}} - \log\left(\phi^{N_1}(1-\phi)^{N-N_1}\right) \\
&= \underset{\phi}{\text{argmin}} - N_1 \log(\phi) - (N - N_1)\log(1-\phi)
\end{aligned}$$

Setting the derivative equal to zero yields

$$\begin{aligned}
0 &= \frac{-N_1}{\phi} + \frac{(N - N_1)}{1-\phi} \\
\implies \phi_{MLE} &= \frac{N_1}{N}
\end{aligned}$$

2. Your coworker tells you that $\phi \sim \text{Beta}(\alpha, \beta)$. That is:

$$p(\phi) = \frac{\phi^{\alpha-1}(1-\phi)^{\beta-1}}{B(\alpha, \beta)}$$

Note that $B(\alpha, \beta)$ is not a function of ϕ and can be treated as a constant. Formulate the optimization of the log posterior, $\text{argmin}_\phi - \log p(\phi \mid \mathcal{D})$, in terms of N, N_1, ϕ, α, and β.

$$\begin{aligned}
&\underset{\phi}{\text{argmin}} - \log p(\phi \mid \mathcal{D}) \\
&= \underset{\phi}{\text{argmin}} - \log\left(p(\mathcal{D} \mid \phi)p(\phi)\right) \qquad \text{Drop } 1/p(\mathcal{D}) \\
&= \underset{\phi}{\text{argmin}} - \log p(\mathcal{D} \mid \phi) - \log p(\phi) \\
&= \underset{\phi}{\text{argmin}} - \log\left(\phi^{N_1}(1-\phi)^{N-N_1}\right) - \log \frac{\phi^{\alpha-1}(1-\phi)^{\beta-1}}{B(\alpha, \beta)} \\
&= \underset{\phi}{\text{argmin}} - \log \phi^{N_1} - \log(1-\phi)^{N-N_1} - \log \phi^{\alpha-1} - \log(1-\phi)^{\beta-1} \qquad \text{Drop } B(\alpha, \beta) \\
&= \underset{\phi}{\text{argmin}} - N_1 \log \phi - (N - N_1)\log(1-\phi) - (\alpha - 1)\log \phi - (\beta - 1)\log(1-\phi) \\
&= \underset{\phi}{\text{argmin}} - (N_1 + \alpha - 1)\log \phi - (N - N_1 + \beta - 1)\log(1-\phi)
\end{aligned}$$

3. Now, calculate $\hat{\phi}_{MAP}$.

$$\begin{aligned}\hat{\phi}_{MAP} &= \text{argmin}_\phi - \log\left(p(\mathcal{D} \mid \phi)p(\phi)\right) \\ &= \text{argmin}_\phi - (N_1 + \alpha - 1)\log(\phi) - (N - N_1 + \beta - 1)\log(1-\phi)\end{aligned}$$

Setting the derivative equal to zero yields

$$\begin{aligned}0 &= \frac{-N_1 - \alpha + 1}{\phi} + \frac{(N - N_1 + \beta - 1)}{1-\phi} \\ \implies \phi_{MAP} &= \frac{N_1 + \alpha - 1}{N + \alpha + \beta - 2}\end{aligned}$$

4. Suppose $N = 100$ and $N_1 = 10$. Furthermore, you believe that in general people click on ads about 6 percent of the time, so you, somewhat naively, decide to set $\alpha = 6 + 1 = 7$, and $\beta = 100 - 6 + 1 = 95$. calculate $\hat{\phi}_{MAP}$

$\hat{\phi}_{MAP} = \frac{N_1+\alpha-1}{N+\alpha+\beta-2} = \frac{10+7-1}{100+102-2} = \frac{16}{200} = 0.08$

5. How do $\hat{\phi}_{MLE}$ and $\hat{\phi}_{MAP}$ differ? Argue which estimate you think is better.

Both estimates are reasonable given the available information. If you believe that this advertisement is similar to those advertisements that averaged a 6 percent click rate, then $\hat{\phi}_{MAP}$ may be a superior estimate, but if the circumstances under which the advertisement was shown were different from the usual, then $\hat{\phi}_{MLE}$ might be a better choice.