Week 7 Glossary

Activation function

A function which transforms linear inputs into nonlinear forms to help the network learn complex patterns in the data. It is applied after the inputs to the network have been linearly transformed by its weights. To make neural networks nonlinear, they must use activation functions. Activation functions are also referred to as transition functions.

Backpropagation

Backpropagation is short for "backward propagation of the gradient of loss." It is an algorithm that updates the weights of a neural network iteratively from the last layer to the first, using the chain rule. It is an efficient way to train deep neural networks when using (stochastic) gradient descent

Batch

The set of examples used in one iteration (that is, one gradient update) of model training.

Batch size

The number of examples in a batch. For example, the batch size of SGD is 1, while the batch size of a mini-batch is usually between 10 and 1000. Batch size is usually fixed during training and inference; however, TensorFlow does permit dynamic batch sizes.

Batch normalization

Batch normalization normalizes the inputs to a layer to have zero mean and unit variance. Because it would be too slow to calculate the mean and variance for every point in very large training datasets, the normalization occurs across the current "mini-batch" instead of the whole training set.

Computer Vision (CV)

Computer vision is a branch of AI that enables computers to identify and understand information from images and videos. Some examples of common CV applications are facial recognition, object detection for self-driving cars and medical imaging analysis, and image analysis for security and surveillance.

Convolutional neural network (CNN)

A deep-learning neural network that makes use of convolutional layers. It learns convolutional filters that process the input image to a particular layer and produce a new image (with many channels). CNNs are particularly useful for processing image data but can also be used for audio signals or videos.

Cosine similarity

A function that computes the similarity between two vectors (or sequences of numbers).

Cross-Entropy

A loss function that measures the disparity between predicted probability distribution and the true label by calculating the average log loss over all classes.

Deep learning

A subset of machine learning that uses neural networks with several hidden layers. In deep learning algorithms, such as are used for recognizing images, it is sometimes the case that different layers are able to learn different types of features in the input data so that by the time the output layer is reached, a comprehensive characterization of the input data is achieved. Deep learning methods are supported by many different software packages, which typically provide high-level support for specifying network architectures and rules for firing.

Deep neural network

A neural network with many more than two hidden layers, each with many more than two nodes. Traditional networks have a few hidden layers, whereas deep learning models can have hundreds of hidden layers. Empirically deep neural networks perform better on certain machine learning tasks, such as Computer Vision tasks.

Feedforward neural network

A type of artificial neural network in which nodes' connections do not form a loop.

Gradient descent

An optimization algorithm that searches for the minimum of a loss function by slightly changing the parameters of a classifier in the direction of the greatest negative gradient of the loss. It is often too hard to calculate the gradient of the loss for all training examples; for this reason, we usually use stochastic gradient descent instead.

Hidden layer

A layer of a neural network located between the input and output layers which applies a linear transformation followed by a nonlinear transformation, called the activation function, to its input values. All neural networks have at least one hidden layer; deep neural networks have many hidden layers.

Hinge

A hinge, in the context of a piecewise activation function like ReLu, refers to the connection point between two linear pieces. Pictorially, they are "kinks" in the decision boundary.

Keras

A high level application programming interface (API) that lets you access the capabilities in TensorFlow with simple syntax.

Mini-batch

A small, randomly selected subset of the entire batch of examples run together in a single iteration of training or inference. The batch size of a mini-batch is usually between 10 and 1,000. It is much more efficient to calculate the loss on a mini batch than on the full training data.

Multi-layer perceptron

A multi-layer Perceptron is a neural network. Each node in a neural network is a Perceptron, and each layer of Perceptrons must have a nonlinear activation function.

Neural network

A supervised learning algorithm designed to solve complex, real-world problems. It can recognize complex patterns and nonlinear relationships between features and labels. Neural networks are often used in the Computer Vision field.

Output layer

In neural networks, the last layer that outputs a prediction.

Rectified linear unit (ReLu)

The rectified linear unit is a popular activation function that applies $max(0, x)$ to an input x. In other words, it clamps x up to 0 when it is negative.

Sigmoid

The sigmoid is a popular activation function with a very easy-to-use derivative. It is defined by $\sigma(x) = (1+e^{-x})^{-1}$.

Softmax

Softmax is a function that can be applied to a vector that forces all the elements of the vector to be between 0 and 1, and sum up to 1. In other words, it turns the vector into a valid probabilty distribution. It is defined as

$$s(\mathrm{x})_k = \frac{\mathrm{e}^{x_k}}{\sum_{I=1}^{n} \mathrm{e}^{x_i}}$$

where $\mathbf{e}$ is the unit vector. It is called softmax because the entry with the largest value in the output is the maximum value in the input, and typically this output value is even larger relative to the other entries (i.e., it dominates the output vector.)

Stochastic gradient descent (SGD)

An approximation of gradient descent. The gradient of the loss function is applied to a subset of all the training examples instead of the whole set, which is much faster to compute.

This stochastic sub-sampling of training samples introduces a lot of noise, which is actually helpful in preventing the algorithm from getting stuck in narrow local minima. Recall that especially for neural networks, we prefer wider local minima.

TensorFlow

An open source machine learning software developed by Google. It provides a very sophisticated set of tools for estimating, analyzing, and deploying deep learning models.

Weights

The parameters in the linear transformations at each hidden layer in a neural network.