

openSAP

Enterprise Deep Learning with TensorFlow

GLOSSARY

Accuracy is the fraction of predictions made by the classification model that are correct, calculated as

$$Accuracy = \frac{Correct\ Predictions}{Total\ Number\ of\ Examples}$$

for multi-class classification problems and

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Number\ of\ Examples}$$

for binary classification problems.

Activation function is a function that takes in the weighted sum of all inputs from a previous layer to produce an output value (typically non-linear) that is passed to the next layer; ReLU and Sigmoid are examples of activation functions.

Adam is a gradient-based optimization technique that computes an adaptive learning rate for each parameter, storing an exponentially decaying average of past squared gradients and an exponentially decaying average of past gradients, similar to momentum.

Anaconda is a package management and deployment tool for scientific computing.

Backward pass calculates and caches output values of each node in the network using a forward pass and then the partial derivative of the error with respect to each parameter to modify the weights of the neural network.

Batch normalization is a technique that normalizes layer input per minibatch during the training, speeding up training.


Bayes error is the lowest possible error that any model can achieve.

Bias is the assumptions made by a model in making the target function easier to learn; a low bias value suggests fewer assumptions made about the target function while a high bias value suggests more assumptions (also known as Underfitting).

Bias-variance trade-off is the goal of achieving low bias and low variance in order to get good generalization while at the same time getting good prediction performance, achieving the best fit for training data but not overfitting to it.

Bucketing is the process of converting a feature (usually continuous) into multiple binary features that are called buckets, based on the value range.

CBoW is an architecture developed to learn the relationships between pairs of words by representing a target word through feeding multiple words as context.



Channel is a dimension of the input data which can consist of multiple channels, examples being images that have *R*, *G*, and *B* channels and text that has different types of embeddings along different channels.

Class is one among a set of enumerated target values for a label.

CNN is a type of neural network that exploits spatial information by enforcing local connectivity patterns between neurons of adjacent layers.

Computation graph is a directed graph where the nodes are functions or computations and the edges are numbers, matrices, or tensors.

Computer vision is an interdisciplinary field which focuses on how computers can gain high-level understanding of the real-world through digital images or videos.

Cosine similarity is a measure of similarity between two non-zero vectors, calculated as

$$\text{similarity}[\cos(\theta)] = \frac{a \cdot b}{||a||_2 ||b||_2}$$

where *a* and *b* are non-zero vectors.

Cost function is a measure using the loss functions that is defined over all training data. Mean squared error is an example of a cost function.

Cross-entropy loss quantifies the difference between two probability distributions; a generalization of log loss to multi-class classification problems.

DataFrame is a two-dimensional, heterogenous, tabular representation of data with labeled axes (both rows and columns) used by Pandas; similar to that of a spreadsheet or a SQL table.

Deep learning is a sub-field of AI, machine learning, and neural networks, describing neural network architectures with many layers of neurons to solve a task.

Deep neural network is a neural network that typically has several hidden layers.

Dropout is a regularization method used in training neural networks wherein a fixed number of units in a neural network layer are randomly removed during a single step of training.


Early stopping is a method for regularization that involves stopping the training of the model even before the training loss finishes decreasing; generally done by monitoring the validation error where training is stopped when the validation error begins to increase, showing signs of worsening generalization performance.

Embeddings map an input representation, such as a word or sentence, into a vector.

Epoch is a forward and backward pass of all the examples available in the training data.

Estimators are a high-level API in TensorFlow that simplifies the development of machine learning applications, encapsulating methods for training, evaluation, prediction, and export for serving.

Exploding gradient occurs in deep neural networks, typically in recurrent neural networks, that use activation functions whose gradients are large, which are multiplied during backpropagation, resulting in a situation where the gradient explodes (becomes extremely large); a scenario opposite to vanishing gradient.



Facets is a library for data exploration and visualization of machine learning data.

False negative (FN) is an example that the model mistakenly predicts to belong to the negative class.

False positive (FP) is an example that the model mistakenly predicts to belong to the positive class.

Feature is an input variable used in making output predictions.

Feature crossing is a synthetic feature that is calculated by crossing (multiplying or calculating a Cartesian product of) individual features, providing good representations for non-linear relations in data.

Feed dictionary is used to override tensor values in a graph, accepts a dictionary as input whose keys are handles to tensor objects that should be overridden, while the values can be numbers, strings, lists, or NumPy arrays.

Fetches accepts a graph element (either an operation or tensor object), which specifies which object would be executed. If the requested object is a tensor, the output of the fetch operation is a NumPy array, or None if the requested object is an operation.

Forward pass computes values from inputs to the output nodes of the neural network.

Frobenius norm also referred to as Euclidean norm is the matrix norm of an $m \times n$ matrix X defined as the square root of the sum of absolute squares of the elements of the matrix,

$$\|X\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |x_{ij}|^2}$$

Fully-connected layer is a layer in a neural network where each neuron in the previous layer is connected to each neuron in the current layer, typically of the form $y = f(Wx + b)$ where x is the layer inputs, W the parameters, b the bias vector, and f a non-linear activation function.

Generalization is the ability of a model to make correct predictions on unseen data.

Git is a source code management software used by GitHub.


GitHub is a version control platform for hosting code and managing projects.

GPU is a graphics processing unit, a specialized hardware designed to perform highly parallel computations for processing large blocks of data efficiently.

Gradient clipping is a technique to prevent exploding gradient problems in deep neural networks, generally done by normalizing the gradients of a parameter vector when the L2 norm of the vector exceeds a threshold.

Gradient descent is a technique that iteratively adjusts parameters of the neural network to find the best combination of weight and bias values in order to minimize loss.

GRU is a simplified LSTM unit with a lesser number of parameters, consisting of a reset and update gate that determine which part of the old memory to retain, and a gating mechanism that allows efficient learning of long-range dependencies.



Hidden layer is a layer in a neural network, between the input (features) and output (prediction) layers, that transforms inputs into useful features for the output layer.

Hinge loss is a family of loss functions used in classification tasks, designed to find a decision boundary that is as distant as possible from each training example, which thus maximizes the margin between examples and the boundary; defined for binary classification as

$$loss = \max(0, 1 - (y' * y))$$

where y' is the output of the classifier model and y is the true label.

Hyperparameters are the knobs that are tweaked based on different training runs of a model.

Inference is the process of making predictions using a trained model on unlabeled examples.

Intelligent Enterprise is a vision for a business organization where machine learning and AI drive core business processes and machines do mundane, repetitive tasks while humans focus on exceptions and higher-value work.

Iteration is a single pass (forward and backward) along the neural network using *batch size* examples.

Jupyter is a browser-based development environment for data science and scientific computing.

Label is the answer part of an example that describes the category to which the data belongs.

Layer consists of one or more nodes; a layer being either an input, output, or a hidden layer, capable of processing data.

Learning rate is a scalar value used to train a model using gradient descent, wherein the gradient descent algorithm multiplies the learning rate by the gradient.

Linear regression is a type of regression model which calculates continuous output values from a linear combination of input features.

Log loss is an information-theoretic measure of the performance of a classification model, used when the model outputs a probability for each class, measured as

$$loss = -(y \log(p) + (1 - y) \log(1 - p))$$


for a binary classification problem.

Logistic regression is a model that generates a probability score for each discrete label value.

Logits is a layer that produces raw values that can be used for predictions; maps input features that are n -dimensional into a tensor that is k -dimensional, where k is the number of classes.

Loss function is a function that is defined on a data point, its corresponding label, and a prediction, to measure the error made by the model; squared loss and hinge loss are examples of loss functions.

LSTM is a special type of RNN that prevents the vanishing gradient problem in RNNs by using a memory gating mechanism.



Mean squared error is calculated as the total squared loss divided by the number of examples.

Mini-batch is a small, randomly selected subset of training examples that are run together in a single iteration of training or during inference.

Mini-batch SGD is a gradient descent algorithm that uses mini-batches where the gradients are estimated based on a small subset of training data.

Model capacity is the ability of a neural network to model any given function; typically, the amount of information that can be stored in the network along with the complexity of the network.

Momentum is a type of gradient descent algorithm where the learning step depends not only on the derivatives calculated in the current step, but also on the derivatives of the step(s) immediately preceding it.

N-gram is a sequence of n contiguous tokens in a given sequence of text.

Neural network is a program that learns from data, inspired by the human brain.

Neuron is a simple compute unit in a neural network.

NLP is a field in computer science which focuses on how computers can understand human languages.

One-hot encoding is a method to encode categorical integer features into a sparse vector, where each column corresponds to a possible value in a feature.

Overfitting is when a model matches the trained data so closely that it is unable to make correct predictions on new unseen data.

Pandas is a column-oriented data analysis tool providing high-performance, easy-to-use data structures.

Parametric Rectified Linear Unit (PReLU) is an activation function where a parameter is learned along with other neural network parameters, equivalent to $\max(x, ax)$

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ ax, & \text{if } x \leq 0 \end{cases}$$

PCA is a technique used for dimensionality reduction of high-dimensional data, using singular value decomposition to project the data to a lower dimension.

Pooling is an operation that helps reduce the dimensionality of a representation by keeping the most important information; common pooling approaches being max-pooling and average-pooling.

Rectified Linear Unit (ReLU) is an activation function where the output is zero if input is negative, equal to input if input is positive, equivalent to $\max(0, x)$

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0 \end{cases}$$

Regularization is the penalty on the model complexity, used to reduce overfitting.

RMS Prop is a gradient-based optimization technique that uses a moving average of squared gradients to normalize the gradient, balancing the step size - reducing the step size to avoid gradient explosion or increasing the step size for small gradients to avoid vanishing gradients.

RNN is a type of neural network that models sequential data through a hidden state, calculating a new hidden state based on the current input and the previous hidden state.

Scope A TensorFlow scope can either be a variable scope or a name scope; a variable scope acts as a context manager, for defining operations, that creates variables while a name scope acts as a context manager for use when defining a Python operation.

Session is an object responsible for graph execution in TensorFlow.

SGD is a gradient descent algorithm which uses a single example chosen at random from the train data to estimate the model parameters at each step, cycling over the entire train data.

Shallow neural network is a neural network that typically has just a single hidden layer.

Sigmoid is an 'S'-shaped function that maps inputs to values in the range [0, 1], which can be interpreted as probabilities, with the formula

$$y = \frac{1}{1 + e^{-\sigma}}$$

where σ in logistic regression problems can be calculated as $\sigma = b + w_1x_1 + \dots w_nx_n$.

Skip-gram is an architecture developed to learn the relationships between pairs of words by representing a set of context words through feeding a target word as input.

Softmax is a function that outputs probabilities for each possible class, the sum of which adds up to exactly 1.0. The function is given by

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}}$$

where z is a vector of the inputs and $j = 0, 1, 2, \dots K$; generalizes the sigmoid to multiple classes.

Squared loss is a loss function used in linear regression (also called L_2 loss) which calculates the square between the model's predicted value and the actual value of the label.


t-SNE is a technique used for dimensionality reduction of high-dimensional data by converting similarity between data points to joint probabilities, minimizing the KL divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data.

Tensor is the primary data structure (N-dimensional array) used to represent data in TensorFlow programs.

TensorBoard is a dashboard that displays summary information from the execution of a TensorFlow program.

TensorFlow is an open-source software library for numerical computation and development of deep learning; its development is led by Google.

TensorFlow Playground is a lightweight software library for interactive visualization of neural networks, capable of running on modern browsers.



TensorFlow Serving is a flexible, high-performance serving system for machine learning models designed for production environments, which makes it easy to deploy new algorithms and experiments while keeping the same server architecture and APIs.

Test set is a subset of the dataset used for testing the model after evaluating the model using the validation set.

Train set or training set is a subset of the dataset used to train a model.

Training is a process for determining the ideal parameters for a given model.

Underfitting is when a model is not powerful enough even to fit the training data.

Validation set or development set or devset is a subset of the dataset used to validate the model performance and adjust the hyperparameters.

Vanishing gradient occurs in deep neural networks, typically in recurrent neural networks that use activation functions whose gradients are small, which are multiplied during backpropagation, resulting in a situation where the gradient vanishes (becomes zero), preventing the network from learning further; an opposite scenario to exploding gradient.

Variables provide the ability to maintain state across executions of the graph in TensorFlow.

Variance is the amount that the estimate of a target function will change when different training data is used; a low variance indicates small changes to the estimate of a target function when training data is changed, while a high variance indicates large changes to the estimate of the target function, modeling the random noise present in training data (also known as overfitting).

Vocabulary is a set of (word) tokens that have vector representations, using which a classifier is trained.

Weight is an edge in a deep neural network or a coefficient for a feature in a linear model, with a goal of determining the ideal weight for each feature.

Word2Vec is a group of models with a shallow, two-layer neural network architecture that are used to produce vector representations for words, or word embeddings.



ACRONYMS

AI Artificial Intelligence
API Application Programming Interface
AUC Area under the ROC Curve
CBoW Continuous Bag of Words
CNN Convolutional Neural Network
GPU Graphics Processing Unit
GRU Gated Recurrent Unit
LSTM Long Short Term Memory
ML Machine Learning
NLP Natural Language Processing
PCA Principal Components Analysis
RNN Recurrent Neural Network
ROC Receiver Operating Characteristic
SGD Stochastic Gradient Descent
t-SNE t-Distributed Stochastic Neighbor Embedding

www.sap.com/contactsap

© 2017 SAP SE or an SAP affiliate company. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP SE or an SAP affiliate company.

The information contained herein may be changed without prior notice. Some software products marketed by SAP SE and its distributors contain proprietary software components of other software vendors. National product specifications may vary.

These materials are provided by SAP SE or an SAP affiliate company for informational purposes only, without representation or warranty of any kind, and SAP or its affiliated companies shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP or SAP affiliate company products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

In particular, SAP SE or its affiliated companies have no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation, and SAP SE's or its affiliated companies' strategy and possible future developments, products, and/or platform directions and functionality are all subject to change and may be changed by SAP SE or its affiliated companies at any time for any reason without notice. The information in this document is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, and they should not be relied upon in making purchasing decisions.

SAP and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP SE (or an SAP affiliate company) in Germany and other countries. All other product and service names mentioned are the trademarks of their respective companies. See <http://www.sap.com/corporate-en/legal/copyright/index.asp> for additional trademark information and notices.

