Title: Data augmentation

URL: https://en.wikipedia.org/wiki/Data_augmentation

PageID: 51443362

Categories: Category:Machine learning

Supervised learning

Unsupervised learning

Semi-supervised learning

Self-supervised learning

Reinforcement learning

Meta-learning

Online learning

Batch learning

Curriculum learning

Rule-based learning

Neuro-symbolic AI

Neuromorphic engineering

Quantum machine learning

Classification

Generative modeling

Regression

Clustering

Dimensionality reduction

Density estimation

Anomaly detection

Data cleaning

AutoML

Association rules

Semantic analysis

Structured prediction

Feature engineering

Feature learning

Learning to rank

Grammar induction

Ontology learning

Multimodal learning

Apprenticeship learning

Decision trees

Ensembles Bagging Boosting Random forest

Bagging

Boosting

Random forest

k -NN

Linear regression

Naive Bayes

Artificial neural networks

Logistic regression

Perceptron

Relevance vector machine (RVM)

Support vector machine (SVM)

BIRCH

CURE

Hierarchical

k -means

Fuzzy

Expectation–maximization (EM)

DBSCAN

OPTICS

Mean shift

Factor analysis

CCA

ICA

LDA

NMF

PCA

PGD

t-SNE

SDL

Graphical models Bayes net Conditional random field Hidden Markov

Bayes net

Conditional random field

Hidden Markov

RANSAC

k -NN

Local outlier factor

Isolation forest

Autoencoder

Deep learning

Feedforward neural network

Recurrent neural network LSTM GRU ESN reservoir computing

LSTM

GRU

ESN

reservoir computing

Boltzmann machine Restricted

Restricted

GAN

Diffusion model

SOM

Convolutional neural network U-Net LeNet AlexNet DeepDream

U-Net

LeNet

AlexNet

DeepDream

Neural field Neural radiance field Physics-informed neural networks

Neural radiance field

Physics-informed neural networks

Transformer Vision

Vision

Mamba

Spiking neural network

Memtransistor

Electrochemical RAM (ECRAM)

Q-learning

Policy gradient

SARSA

Temporal difference (TD)

Multi-agent Self-play

Self-play

Active learning

Crowdsourcing

Human-in-the-loop

Mechanistic interpretability

RLHF

Coefficient of determination

Confusion matrix

Learning curve

ROC curve

Kernel machines

Bias–variance tradeoff

Computational learning theory

Empirical risk minimization

Occam learning

PAC learning

Statistical learning

VC theory

Topological deep learning

AAAI

ECML PKDD

NeurIPS

ICML

ICLR

IJCAI

ML

JMLR

Glossary of artificial intelligence

List of datasets for machine-learning research List of datasets in computer vision and image processing

List of datasets in computer vision and image processing

Outline of machine learning

v

t

e

Data augmentation is a statistical technique which allows maximum likelihood estimation from incomplete data. Data augmentation has important applications in Bayesian analysis , and the technique is widely used in machine learning to reduce overfitting when training machine learning models, achieved by training models on several slightly-modified copies of existing data.

Synthetic oversampling techniques for traditional machine learning

Synthetic Minority Over-sampling Technique (SMOTE) is a method used to address imbalanced datasets in machine learning. In such datasets, the number of samples in different classes varies significantly, leading to biased model performance. For example, in a medical diagnosis dataset with 90 samples representing healthy individuals and only 10 samples representing individuals with a particular disease, traditional algorithms may struggle to accurately classify the minority class.

SMOTE rebalances the dataset by generating synthetic samples for the minority class. For instance, if there are 100 samples in the majority class and 10 in the minority class, SMOTE can create synthetic samples by randomly selecting a minority class sample and its nearest neighbors, then generating new samples along the line segments joining these neighbors. This process helps increase the representation of the minority class, improving model performance.

Data augmentation for image classification

When convolutional neural networks grew larger in mid-1990s, there was a lack of data to use, especially considering that some part of the overall dataset should be spared for later testing. It was proposed to perturb existing data with affine transformations to create new examples with the same labels, which were complemented by so-called elastic distortions in 2003, and the technique was widely used as of 2010s. Data augmentation can enhance CNN performance and acts as a countermeasure against CNN profiling attacks.

Data augmentation has become fundamental in image classification, enriching training dataset diversity to improve model generalization and performance. The evolution of this practice has introduced a broad spectrum of techniques, including geometric transformations, color space adjustments, and noise injection.

Geometric Transformations

Geometric transformations alter the spatial properties of images to simulate different perspectives, orientations, and scales. Common techniques include:

Rotation: Rotating images by a specified degree to help models recognize objects at various angles.

Flipping: Reflecting images horizontally or vertically to introduce variability in orientation.

Cropping: Removing sections of the image to focus on particular features or simulate closer views.

Translation: Shifting images in different directions to teach models positional invariance.

Morphing within the same class : Generating new samples by applying morphing techniques between two images belonging to the same class, thereby increasing intra-class diversity.

Color Space Transformations

Color space transformations modify the color properties of images, addressing variations in lighting, color saturation , and contrast. Techniques include:

Brightness Adjustment: Varying the image's brightness to simulate different lighting conditions.

Contrast Adjustment: Changing the contrast to help models recognize objects under various clarity levels.

Saturation Adjustment: Altering saturation to prepare models for images with diverse color intensities.

Color Jittering: Randomly adjusting brightness, contrast, saturation, and hue to introduce color variability.

Noise Injection

Injecting noise into images simulates real-world imperfections, teaching models to ignore irrelevant variations. Techniques involve:

Gaussian Noise: Adding Gaussian noise mimics sensor noise or graininess.

Salt and Pepper Noise: Introducing black or white pixels at random simulates sensor dust or dead pixels .

Data augmentation for signal processing

Residual or block bootstrap can be used for time series augmentation.

Biological signals

Synthetic data augmentation is of paramount importance for machine learning classification, particularly for biological data, which tend to be high dimensional and scarce. The applications of robotic control and augmentation in disabled and able-bodied subjects still rely mainly on subject-specific analyses. Data scarcity is notable in signal processing problems such as for Parkinson's Disease Electromyography signals, which are difficult to source - Zanini, et al. noted that it is possible to use a generative adversarial network (in particular, a DCGAN) to perform style transfer in order to generate synthetic electromyographic signals that corresponded to those exhibited by sufferers of Parkinson's Disease.

The approaches are also important in electroencephalography (brainwaves). Wang, et al. explored the idea of using deep convolutional neural networks for EEG-Based Emotion Recognition, results show that emotion recognition was improved when data augmentation was used.

A common approach is to generate synthetic signals by re-arranging components of real data. Lotte proposed a method of "Artificial Trial Generation Based on Analogy" where three data examples x 1 , x 2 , x 3 {\displaystyle x_{1},x_{2},x_{3}} provide examples and an artificial x s y n t h e t i c {\displaystyle x_{synthetic}} is formed which is to x 3 {\displaystyle x_{3}} what x 2 {\displaystyle x_{2}} is to x 1 {\displaystyle x_{1}} . A transformation is applied to x 1 {\displaystyle x_{1}} to make it more similar to x 2 {\displaystyle x_{2}} , the same transformation is then applied to x 3 {\displaystyle x_{3}} which generates x s y n t h e t i c {\displaystyle x_{synthetic}} . This approach was shown to improve performance of a Linear Discriminant Analysis classifier on three different datasets.

Current research shows great impact can be derived from relatively simple techniques. For example, Freer observed that introducing noise into gathered data to form additional data points improved the learning ability of several models which otherwise performed relatively poorly. Tsinganos et al. studied the approaches of magnitude warping, wavelet decomposition, and synthetic surface EMG models (generative approaches) for hand gesture recognition, finding classification performance increases of up to +16% when augmented data was introduced during training. More recently, data augmentation studies have begun to focus on the field of deep learning, more specifically on the ability of generative models to create artificial data which is then introduced during the classification model training process. In 2018, Luo et al. observed that useful EEG signal data could be generated by Conditional Wasserstein Generative Adversarial Networks (GANs) which was then introduced to the training set in a classical train-test learning framework. The authors found classification performance was improved when such techniques were introduced.

Mechanical signals

The prediction of mechanical signals based on data augmentation brings a new generation of technological innovations, such as new energy dispatch, 5G communication field, and robotics control engineering. In 2022, Yang et al. integrate constraints, optimization and control into a deep network framework based on data augmentation and data pruning with spatio-temporal data correlation, and improve the interpretability, safety and controllability of deep learning in real industrial projects through explicit mathematical programming equations and analytical solutions.

See also

Oversampling and undersampling in data analysis

Surrogate data

Generative adversarial network

Variational autoencoder

Data pre-processing

Convolutional neural network

Regularization (mathematics)

Data preparation

Data fusion
References
v
t
e
Acquisition
Augmentation
Analysis
Anonymization
Archaeology
Big
Cleansing
Collection
Compression
Corruption
Curation
Deduplication
Degradation
De-identification
Ecosystem
Editing
Engineering
Erasure
ETL / ELT Extract Transform Load
Extract
Transform
Load
Ethics
Exhaust
Exploration
Farming
Format management
Fusion
Governance Cooperatives
Cooperatives
Infrastructure
Integration
Integrity

Library
Lineage
Loss
Management
Meta
Migration
Mining
Philanthropy
Pre-processing
Preservation
Processing
Protection (privacy)
Publishing Open data
Open data
Recovery
Reduction
Redundancy
Re-identification
Remanence
Rescue
Retention
Quality
Science
Scraping
Scrubbing
Security
Sharing
Stewardship
Storage
Structure
Synchronization
Topological data analysis
Type
Validation
Warehouse
Wrangling/munging
v
t

e
History timeline
timeline
Companies
Projects
Parameter Hyperparameter
Hyperparameter
Loss functions
Regression Bias–variance tradeoff Double descent Overfitting
Bias–variance tradeoff
Double descent
Overfitting
Clustering
Gradient descent SGD Quasi-Newton method Conjugate gradient method
SGD
Quasi-Newton method
Conjugate gradient method
Backpropagation
Attention
Convolution
Normalization Batchnorm
Batchnorm
Activation Softmax Sigmoid Rectifier
Softmax
Sigmoid
Rectifier
Gating
Weight initialization
Regularization
Datasets Augmentation
Augmentation
Prompt engineering
Reinforcement learning Q-learning SARSA Imitation Policy gradient
Q-learning
SARSA
Imitation
Policy gradient
Diffusion

Latent diffusion model
Autoregression
Adversary
RAG
Uncanny valley
RLHF
Self-supervised learning
Reflection
Recursive self-improvement
Hallucination
Word embedding
Vibe coding
Machine learning In-context learning
In-context learning
Artificial neural network Deep learning
Deep learning
Language model Large language model NMT
Large language model
NMT
Reasoning language model
Model Context Protocol
Intelligent agent
Artificial human companion
Humanity's Last Exam
Artificial general intelligence (AGI)
AlexNet
WaveNet
Human image synthesis
HWR
OCR
Computer vision
Speech synthesis 15.ai ElevenLabs
15.ai
ElevenLabs
Speech recognition Whisper
Whisper
Facial recognition
AlphaFold

Text-to-image models Aurora DALL-E Firefly Flux Ideogram Imagen Midjourney Recraft Stable Diffusion

Aurora

DALL-E

Firefly

Flux

Ideogram

Imagen

Midjourney

Recraft

Stable Diffusion

Text-to-video models Dream Machine Runway Gen Hailuo AI Kling Sora Veo

Dream Machine

Runway Gen

Hailuo AI

Kling

Sora

Veo

Music generation Riffusion Suno AI Udio

Riffusion

Suno AI

Udio

Word2vec

Seq2seq

GloVe

BERT

T5

Llama

Chinchilla AI

PaLM

GPT 1 2 3 J ChatGPT 4 4o o1 o3 4.5 4.1 o4-mini 5

1

2

3

J

ChatGPT

4

4o

o1

o3

4.5

4.1

o4-mini

5

Claude

Gemini Gemini (language model) Gemma

Gemini (language model)

Gemma

Grok

LaMDA

BLOOM

DBRX

Project Debater

IBM Watson

IBM Watsonx

Granite

PanGu-Σ

DeepSeek

Qwen

AlphaGo

AlphaZero

OpenAI Five

Self-driving car

MuZero

Action selection AutoGPT

AutoGPT

Robot control

Alan Turing

Warren Sturgis McCulloch

Walter Pitts

John von Neumann

Claude Shannon

Shun'ichi Amari

Kunihiko Fukushima

Takeo Kanade

Marvin Minsky

John McCarthy
Nathaniel Rochester
Allen Newell
Cliff Shaw
Herbert A. Simon
Oliver Selfridge
Frank Rosenblatt
Bernard Widrow
Joseph Weizenbaum
Seymour Papert
Seppo Linnainmaa
Paul Werbos
Geoffrey Hinton
John Hopfield
Jürgen Schmidhuber
Yann LeCun
Yoshua Bengio
Lotfi A. Zadeh
Stephen Grossberg
Alex Graves
James Goodnight
Andrew Ng
Fei-Fei Li
Alex Krizhevsky
Ilya Sutskever
Oriol Vinyals
Quoc V. Le
Ian Goodfellow
Demis Hassabis
David Silver
Andrej Karpathy
Ashish Vaswani
Noam Shazeer
Aidan Gomez
John Schulman
Mustafa Suleyman
Jan Leike
Daniel Kokotajlo

François Chollet
Neural Turing machine
Differentiable neural computer
Transformer Vision transformer (ViT)
Vision transformer (ViT)
Recurrent neural network (RNN)
Long short-term memory (LSTM)
Gated recurrent unit (GRU)
Echo state network
Multilayer perceptron (MLP)
Convolutional neural network (CNN)
Residual neural network (RNN)
Highway network
Mamba
Autoencoder
Variational autoencoder (VAE)
Generative adversarial network (GAN)
Graph neural network (GNN)
Category