Datasets For Training Neural Networks

Group 11

E/15/123 Wishma Herath e15123@eng.pdn.ac.lk
E/15/173 Dilshani Karunarathna e15173@eng.pdn.ac.lk
E/15/280 Pubudu Prematilaka pubudu.prematilaka@eng.pdn.ac.lk
E/15/316 Suneth Samarasinghe e15316@eng.pdn.ac.lk

Date: 05/03/2021

1. MNIST digit classification dataset

- MNIST stands for Modified National Institute of Standards and Technology database
- Subset of NIST database
- Dataset size: 70,000 images
  - Train : 60,000
  - Test : 10,000
- Number of Classes: 10
- Image size: 20x20

http://yann.lecun.com/exdb/mnist/

Papers

- ADAM (A Method for Stochastic Optimization)
  - L2-regularized multi-class logistic regression
- Training Very Deep Networks
  - To train
  - Test set classification accuracy for pilot experiments
- Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning
  - Compare the uncertainty obtained from different model architectures

2. IMDB movie review dataset

- A binary sentiment analysis dataset
- Consisting of 50,000 reviews
- Labeled as positive or negative
  - A negative review : score ≤ 4/10
  - A positive review : score ≥ 7/10
- For training : 25,000
- For testing : .25,000

http://nlpprogress.com/english/sentiment_analysis.htm

Papers

- ADAM (A Method for Stochastic Optimization)
  - IMDB movie review dataset from (Maas et al., 2011). - sparse feature problem

3. NIST Special Database 1/2

- NIST Structured Forms Database consists of 5,590 pages of binary, black-and-white images of synthesized documents
- The database has the following features:
  - 900 simulated tax submissions
  - 5,590 images of completed structured form faces
  - 5,590 text files containing entry field answers
  - 20 tables of entry field types and contexts

Papers

- Gradient-based learning applied to document recognition
  - Reviews various methods applied to handwritten character recognition and comparison

4. Toronto Face Database(TFB)

- Set of 32×32 grayscale images
- A small subset of faces have been labeled into seven categories.
  - Anger, Disgust, Fear, Happy, Sad, Surprise, Neutral
- 2925 labeled images for training and validation
- Each labeled face image has an identity

(a) Anger

(b) Disgust

Source: https://www.cs.toronto.edu/~urtasun/courses/CSC411/hw3-411.pdf

Papers

- Generative Adversarial Nets
  - Propose a new framework for estimating generative models via an adversarial process

5. ImageNet

Dataset size: 14.2 million annotated images

Since 2010 the dataset is used in the ILSVRC

ILSVRC annotations

(1) image-level annotation

(2) object-level annotation

Source: https://cs.stanford.edu/people/karpathy/cnnembed/

www.image-net.org

Papers

09. Compression of deep convolutional neural networks for fast and low power mobile applications

Authors: Kim Y.D., Park E., Yoo S., Choi T., Yang L. and Shin D.

Year: 2015

Simple and effective scheme to compress the entire CNN to deploy deep CNNs on mobile devices called “one-shot whole network compression”

Image Net 2012 dataset:

- in fine-tuning step to recover the accuracy loss of the compression scheme.
- for the validation

Papers

11. Imagenet classification with deep convolutional neural networks

Authors: Krizhevsky, A., Sutskever, I. and Hinton, G.E.

Year: 2017

A deep CNN for ImageNet LSVRC-2010 contest

Achieved top-1 and top-5 error rates that are considerably better than the previous state-of-the-art

ImageNet 2010 dataset used to train, validate and test

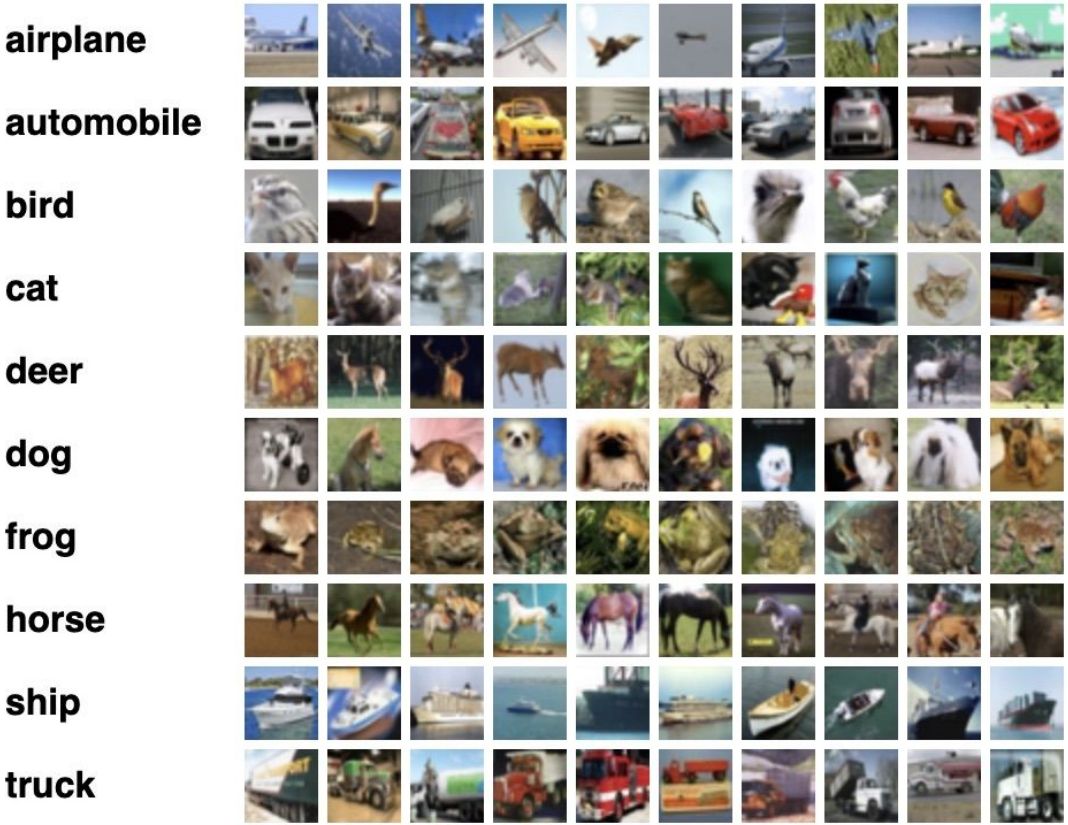
6. CIFAR-10

Dataset size: 60,000 images

Train	50000
Test	10000

Image size: 32x32 colour images

Number of Classes: 10 mutually exclusive



Source: https://www.cs.toronto.edu/~kriz/cifar.html.

CIFAR-100

Dataset size: 60,000 images

Train	50000
Test	10000

Number of Classes: 100 (completely mutually exclusive)
Image size: 32x32 colour images

Grouped into 20 superclasses
Each image has 2 labels

- Fine label : class
- Coarse label: superclass

https://www.cs.toronto.edu/~kriz/cifar.html

Papers

04. Training Very Deep Networks

Authors: Srivastava, R.K., Greff, K. and Schmidhuber, J.

Year: 2015

A new architecture designed to overcome inefficiencies of training when the depth increases.

Datasets:

- MNIST digit classification dataset: for pilot experiments
- **CIFAR-10** and **CIFAR-100**: object recognition experiments

Workshop on Statistical Machine Translation (WMT) 2014

- A collection of datasets used in shared tasks of the Ninth Workshop on Statistical Machine Translation.
- The primary objectives of WMT
  - to evaluate the state of the art in machine translation,
  - to disseminate common test sets and public training data
  - to refine evaluation and estimation methodologies
- The workshop featured four tasks:
  1. a news translation task,
  2. a quality estimation task,
  3. a metrics task,
  4. a medical text translation task.

Workshop on Statistical Machine Translation (WMT) 2014

- Available language pairs
  - French-English
  - Hindi-English
  - German-English
  - Czech-English
  - Russian-English

7. WMT 2014 English-French dataset

- Consisting of 36M sentences
  - split tokens into a 32000 word-piece vocabulary
- Sentence pairs were batched
- Training batch contained a set of sentence pairs containing approximately 25000 source tokens and 25000 target tokens.
- Size: 6.20 GiB

8. WMT 2014 English-German dataset

- 4.5 million sentence pairs.
- Sentences were encoded and has 37000 tokens.
- Training batch contained a set of sentence pairs
  - ~25000 source tokens
  - ~25000 target tokens.
- Size: 1.58 GiB

Tasks can be done using the datasets

- Investigate the applicability of current MT techniques
- Examine special challenges in translating between European languages
- Investigate the translation of low-resource, morphologically rich languages
- Create publicly available corpora for machine translation and machine translation evaluation
- Generate up-to-date performance numbers for European languages
- Offer newcomers a smooth start with hands-on experience in state-of-the-art statistical machine translation methods

Papers

13. Attention Is All You Need

Author: Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit,
Llion Jones,Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

Year: 2017

- A Transformer,
  - a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output.
  - Allows for significantly more parallelization.
  - Can be trained faster than architectures

Concluded results

(BLEU => bilingual evaluation understudy)

- Achieves 28.4 BLEU on the WMT 2014 English-to-German translation task,
- improving over the existing best results, including ensembles.

- On the WMT 2014 English-to-French translation task, the model establishes a new single-model state-of-the-art BLEU score of 41.0

What is VQA?

Visual Question Answering

A new dataset containing open-ended questions about images.

Includes

- COCO (Common Objects in Context) images and abstract scenes
- Questions (5.4 questions on average) per image
- Ground truth answers per question
- Plausible (but likely incorrect) answers per question
- Automatic evaluation metric

Datasets Versions

- October 2015: Full release (v1.0)

  Real Images
  - 204,721 COCO images (all of current train/val/test)
  - 614,163 questions
  - 6,141,630 ground truth answers
  - 1,842,489 plausible answers

  Abstract Scenes
  - 50,000 abstract scenes
  - 150,000 questions
  - 1,500,000 ground truth answers
  - 450,000 plausible answers
  - 250,000 captions

- April 2017: Full release (v2.0)

  Balanced Real Images
  - 204,721 COCO images
  - (all of current train/val/test)
  - 1,105,904 questions
  - 11,059,040 ground truth answers

Dataset Details (Statical)

- The dataset includes 614,163 questions
- 7,984,119 answers (including answers provided by workers with and without looking at the image)
- 204,721 images from the MS COCO dataset
- 150,000 questions with 1,950,000 answers for 50, 000 abstract scenes.

Datasets Details (descriptional)

- The MS COCO dataset has images depicting diverse and complex scenes
- Collected a new dataset of “realistic” abstract scenes to enable research focused only on the high level reasoning
- Three questions were collected for each image or scene.
- Each question was answered by ten subjects.
- Contains over 760K questions with around 10M answers.

VQA Dataset Collection

- Real Images
  - Used 123,287 training and validation images and 81,434 test images from Microsoft Common Objects in Context (MS COCO) dataset.
- Abstract Scenes
  - Created a new abstract scenes dataset containing 50K scenes to attract researchers interested in exploring the high-level reasoning required for VQA,
  - The dataset contains 20 “paperdoll” human models spanning genders, races, and ages with 8 different expressions.

VQA Dataset Collection

- Splits
  - For real images, follow the same train/val/test split strategy as the MC COCO dataset
  - For abstract scenes, Create standard splits, separating the scenes into 20K/10K/20K for train/val/test splits, respectively.
- Captions
  - The MS COCO dataset
  - Collected five single-captions for all abstract scenes

VQA Dataset Collection

- Questions
  - Many simple questions may only require low-level computer vision knowledge
  - Questions that require commonsense knowledge about the scene
  - dataset contains over ~0.76M questions
- Answers.
  - 10 answers are gathered for each question
  - For testing, two modalities were offered for answering the questions:
    1. open-answer
    2. multiple-choice.

Paper

15. VQA: Visual Question Answering

Author: Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh

Year: 2015

- Introduce the task of Visual Question Answering (VQA).
- Demonstrated
  - questions and answers in their dataset,
  - diverse set of AI capabilities in computer vision,
  - natural language processing
  - commonsense reasoning

10. Mauna Loa CO2 dataset

CONTRIBUTOR: C. D. KEELING SCRIPPS, INSTITUTION OF OCEANOGRAPHY UNIVERSITY OF CALIFORNIA LA JOLLA, CA 92093

Year: From March 1958 Upto now

- Contains monthly and annual atmospheric carbon dioxide (CO2) concentrations
- The longest continuous record of atmospheric CO2 concentrations available in the world.
- Monthly and annual average mole fractions of CO2 in water-vapor-free air are given except for a few interruptions.

Papers

16. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning

Author: Yarin Gal, Zoubin Ghahramani

Year: 2016

- Deep learning tools for regression and classification do not capture model uncertainty.
- Bayesian models offer a mathematically grounded framework.
- Develop a new theoretical framework casting dropout training in deep neural networks (NNs)

Datasets

- Used two regression datasets and model scalar functions
- Used a subset of the atmospheric CO2 concentrations dataset to evaluate model extrapolation.
- The datasets are fairly small
- Both datasets were centred and normalised.

Thank You