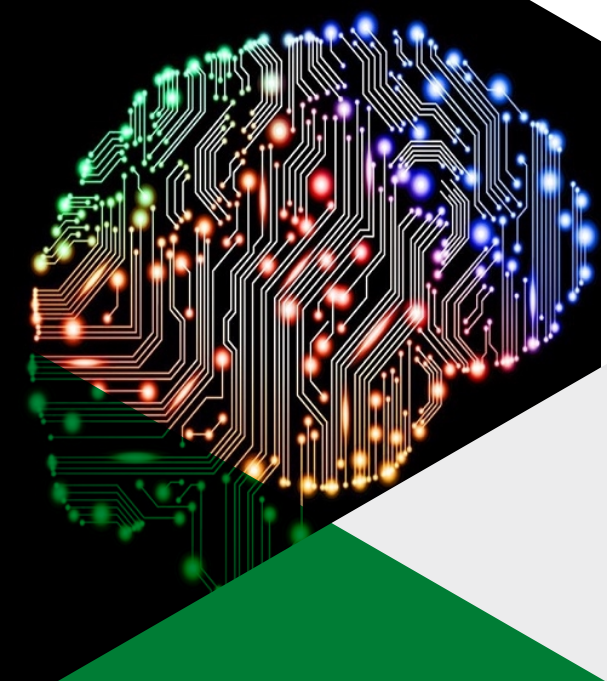


LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Final Presentation

Multi Modal Deep Learning

C. Marquardt

20.07.2022

Overview

Datasets

Datasets

(or some of the most common ones)

Dataset for the following modalities will be covered

- Natural Language Processing (NLP)
- Computer Vision (CV)
- Multi Modal (MM)

Even if the datasets might be completely different, the procedure to get them isn't

NLP Datasets

The Pile

- Rosset et al.(Rosset, 2020): Diversity in training datasets improves general cross-domain knowledge and downstream generalization
- large web scrapes and more targeted, higher-quality datasets (825 GB)
- 22 sub-datasets

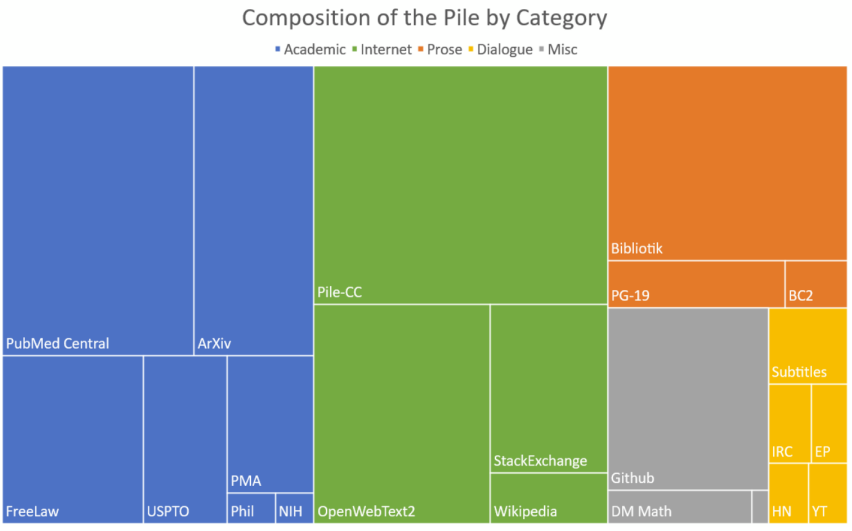


Figure 1: Composition of the Pile

NLP Datasets

BooksCorpus & Multi-Lingual Datasets

BooksCorpus

- unplished authors from smashwords.com
- 11k books from 16 different genres
- Some concerns:
  - copyright violations
  - duplicate books
  - problematic content

Multi-Lingual

- mC4 (+100 languages)
- CC-100 (+100 languages)
- BibleCorpus (+900 languages)
- Ted Talks (+50 languages)

Multi-lingual dataset might help low-resource languages

CV Datasets

ImageNet

- uses the hierarchical structure of WordNet
- amount of classes was unheard at this time point
- increased the resolution
- 1.3 Terabyte

ImageNet 1K

- 1.2 million training images
- 1k mutually exclusive labels

ImageNet 21K

- 14 million images
- almost 22k labels

CV Datasets

Google's CV Datasets

Joint-Foto-Tree (JFT)

- multiple versions
  - JFT 300M:
    300 million images, 375 million labels $\approx$ 1.26 labels per image
- 18k classes

Entity-Foto-Tree (EFT)

- 20 diversified verticals and consists of 100k classes
- rarely used by Google
  large model size and the slow training speed

- labels generated automatically
- no datasheet available

CV Datasets

Object 365 & MS-COCO

Object 365

- 11 super-categories and 365 object categories
- 2 million images
- collected images mainly from Flicker
- mainly used to train object detection and semantic segmentation
- $>$ 10M bounding boxes

Microsoft Common Objects in Context

- non-iconic image collection
- 11 super-categories and 91 categories
- 328k images
- five written captions and instance segmentation per image
- used for object detection and semantic segmentation

MM Datasets

LAION Datasets

LAION-400M *aka.crawling@home*

- The answer to Open AI collection 250 million text-images pairs
- 400 million image-text pairs
- used CLIP to compute embeddings

LAION-5B

- 14× bigger than LAION-400M
- 5.85 billion CLIP-filtered image-text pairs

- LAION-5B is the biggest openly accessible image-text dataset
- opens the road for everyone

MM Datasets

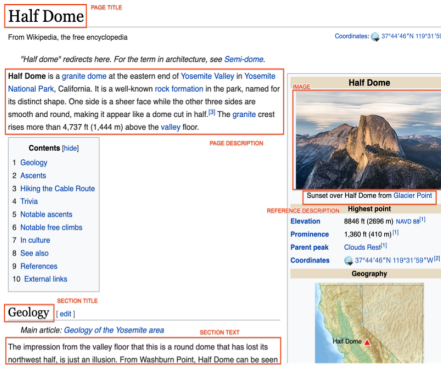
Localized Narratives

- 849k images
- contains COCO, ADE20K, Flickr30k & 32k datasets and 671k images of Open Images
- new form of connecting vision and language
- four synchronized modalities
- manual transcription

MM Datasets

Wikipedia Image Text (WIT)

- 37.6 million image-text sets and spans 11.5 million unique images
- 100K image-text pairs in 53 languages
- 12K image-text pairs in 108 languages
- Uses Wikipedia articles and Wikimedia image
  - ensure a high-quality bar
  - additionaly extensive human-annotation process



Datasets

NLP Datasets	CV Datasets	MM Datasets
Common Crawl	ImageNet	LAION-400M & 5B
The Pile	JFT & EFT	Localized Narratives
BooksCorpus	Object 365	WuDaoMM
Multi-Lingual	MS-COCO	Wikipedia Image Text

- collected data is often not public available
  - necessity due to the proprietary or sensitive nature of the data
  - not always the case
- effective dataset is a catalyst and accelerator for technological development (Yuan et al., 2022)
  - be a reason, why the big companies don't share their datasets

Pre-Training Tasks

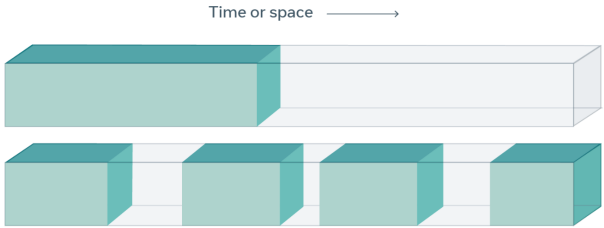
Supervised Learning (SL) gone and Self-SL future?

- New data easy to get, BUT
  - carefully labeled by humans
  - unlabeled data: richer representation (Self-Attention @ Last Layer)

(a) Supervised **(b)** Ground Truth **(c)** Self-SL (DINO)

- Supervised learning strategy is a bottleneck for future progress
- Self SL uses supervised learning algorithm, but no external supervisor
- Self SL biologically more plausible than supervised methods
- Feasibility of Self SL depends on modality

Self-Supervised Learning aka. Predictive Learning



NLP

- easy to implement
- uses contrastive methods

CV

- impossible to implement like in NLP
- contrastive methods
- non-contrastive methods

MM

- easy to implement compared to CV
- already have hard negatives

Benchmarks

Benchmarks

Comments on Benchmarks

- Important to create and use new benchmarks
- models need to be compared or to capture effect of different pre-trainings
- Split datasets into training, test and validation sets
- Tested on so called held-out data
  - held-out datasets are often not comprehensive, and contain the same biases
  - may overestimate the real-world performance
- Overlap of pre-training and down-tasks

NLP Benchmarks

(Super) General Language Understanding Evaluation (GLUE)

- created to benchmark General Language Understanding

GLUE

- nine different task datasets
  - Single-Sentence Tasks
  - Similarity and Paraphrase Tasks
  - Inference Tasks

Super-GLUE

- eight language understanding tasks
- more challenging tasks
- more diverse task formats

- both provide leaderboard with a human benchmark

Dataset	Description	Data example
CoLA	Is the sentence grammatical or ungrammatical?	"This building is than that one." = **Ungrammatical**
SST-2	Is the movie review positive, negative, or neutral?	"The movie is funny , smart , visually inventive , and most of all , alive ." = **.93056 (Very Positive)**
MRPC	Is the sentence B a paraphrase of sentence A?	A) "Yesterday , Taiwan reported 35 new infections , bringing the total number of B) "The island reported another 35 probable cases yesterday , taking its total to = **A Paraphrase**
STS-B	How similar are sentences A and B?	A) "Elephants are walking down a trail." B) "A herd of elephants are walking along a trail." = **4.6 (Very Similar)**
QQP	Are the two questions similar?	A) "How can I increase the speed of my internet connection while using a VPN?" B) "How can Internet speed be increased by hacking through DNS?" = **Not Similar**
MNLI-mm	Does sentence A entail or contradict sentence B?	A) "Tourist Information offices can be very helpful." B) "Tourist Information offices are never of any help." = **Contradiction**
QNLI	Does sentence B contain the answer to the question in sentence A?	A) "What is essential for the mating of the elements that create radio waves?" B) "Antennas are required by any radio receiver or transmitter to couple its elect to the electromagnetic field." = **Answerable**
RTE	Does sentence A entail sentence B?	A) "In 2003, Yunus brought the microcredit revolution to the streets of Banglade more than 50,000 beggars, whom the Grameen Bank respectfully calls Strugglі B) "Yunus supported more than 50,000 Struggling Members." = **Entailed**
WNLI	Sentence B replaces sentence A's ambiguous pronoun with one of the nouns - is this the correct noun?	A) "Lily spoke to Donna, breaking her concentration." B) "Lily spoke to Donna, breaking Lily's concentration." = **Incorrect Referent**

Figure 3: GLUE

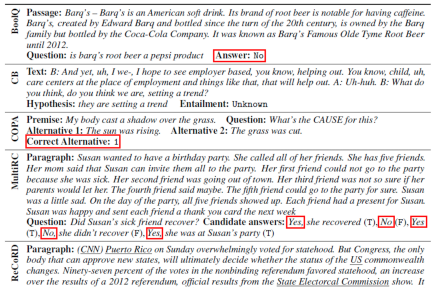


Figure 4: Super-GLUE

Is it really understanding?

NLP Benchmarks

Beyond the Imitation Game Benchmark (BIG-bench)

- collaborative living benchmark (200 tasks)
- current language-modeling benchmarks are insufficient
  - short useful lifespans “challenge-solve-and-replace”
  - others benachmarks are narrowly targeted
  - collected through non-experts and not by task authors

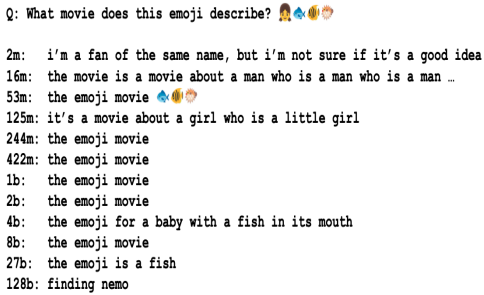


Figure 5: emoji_movie task

NLP Benchmarks

NLI broken? (Bowman & Dahl, 2021)

- systems already score so highly on standard benchmarks
- more work on dataset design and data collection
- much harder and/or much larger benchmarks
- use of auxiliary bias evaluation metrics

CheckList

- model-agnostic and task-agnostic methodology for testing NLP models
- three tasks:
  - **M**inimum **F**unctionality **T**est
  - **INV**ariance test
  - **DIR**ectional Expectation test
- software tool to generate a large and diverse test cases
- revealed critical bugs in commercial systems (BERT, RoBERTa)

	Test *TYPE* and Description
Vocab	***MFT:*** comparisons
	MFT: intensifiers to superlative: most/least

Figure 6: Machine Comprehension Task

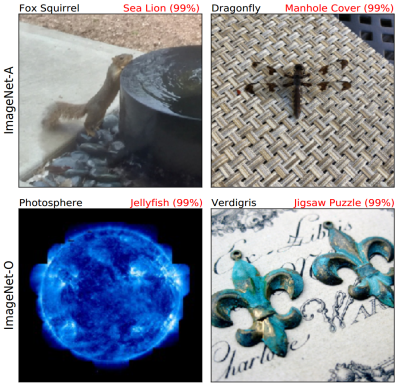
Example Test cases (with expected behavior and prediction)
C: Victoria is younger than Dylan. **Q:** Who is less young? **A:** Dylan : Victoria
C: Anna is worried about the project. Matthew is extremely worried about the project. **Q:** Who is least worried about the project? **A:** Anna : Matthew

Figure 7: Machine Comprehension Example

CV Benchmarks

ImageNet

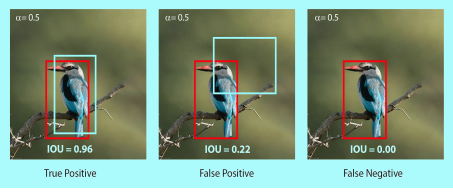
- ImageNet-R: non-natural images
- ImageNet-A: adversarial images
- ImageNet-O: Out-of-distribution
- ImageNet-V2
- ReaL ("Reassessed Labels")
  - new ImageNet validation set
  - mostly clear view on a single object
  - other images contain multiple objects
  - new metric uses set of labels (ReaL accuracy)

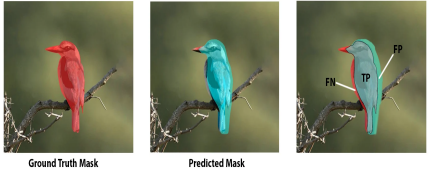


CV Benchmarks

MS-COCO & Object365 & ADE 20K

- object detection dataset
- Panoptic and instance segmentation often done on COCO
- semantic segmentation on ADE20K
  - 20,000 images (between 5 and 273 objects per image)
- precision is captured with Intersection over Union (IoU)
- object detection: mean Average Precision (mAP) at IoU
- semantic segmentation: calculation of IoU more intuitive

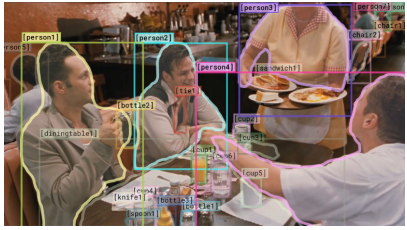




MM Benchmarks

Visual Commonsense Reasoning (VCR)

- 290k multiple choice QA problems derived from 110k movie scenes
- four-way multiple choice task
  1. Question Answering
  2. Answer Justification
- can be evaluated alone or combined
- Almost every answer and rationale is unique.

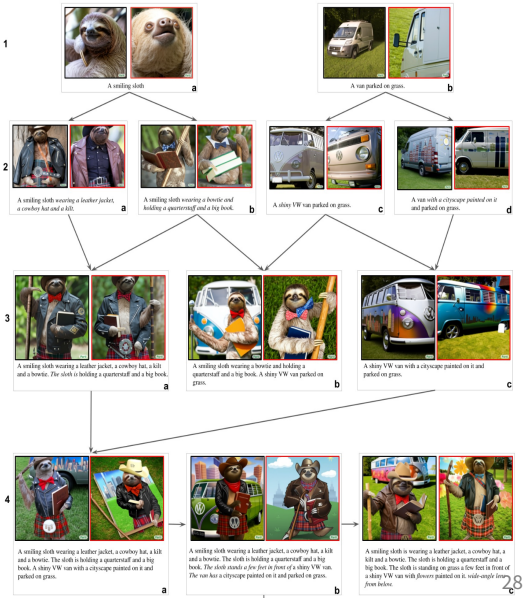


MM Benchmarks

Generative Models Benchmarks

How can we compare these models?

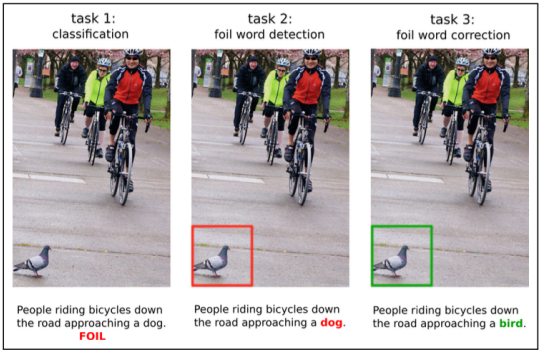
- Automatic image quality
  - Fréchet Inception Distance (FID)
- automatic image-text alignment
  - BLEU, CIDEr, METEOR and SPICE
- human evaluation
  - PartiPrompts (P2), DrawBench, Localized Narratives
- results are "cherry picked"
  - process of "growing a cherry tree"



MM Benchmarks

FOIL it! (Find One mismatch between Image and Language caption)

- builds on MS-COCO dataset
- dataset with minimal language bias
- introducing one single error per caption
- three different tasks
  1. Correct vs. foil classification
  2. Foil word detection
  3. Foil word correction



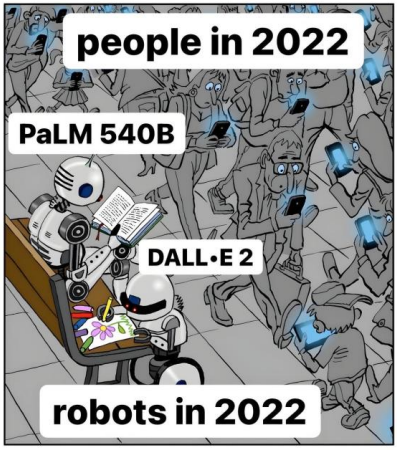
MM Benchmarks

VALSE

- builds on the idea of FOIL it!
- zero-shot benchmark
- basic linguistic phenomena affecting the linguistic and visual modalities

pieces	existence	plurality	counting	relations	actions	coreference
caption (blue) / foil (orange)	*There are no animals / animals shown.*	*A small copper vase with some flowers / exactly one flower in it.*	*There are four / six zebras.*	*A cat plays with a pocket knife on / underneath a table.*	*A man / woman shouts at a woman / man.*	*Buffalos walk along grass. Are they in a zoo? No / Yes.*
image						

Questions?



References

Bowman, S. R., & Dahl, G. E. (2021). What will it take to fix benchmarking in natural language understanding? *arXiv preprint arXiv:2104.02145*.

Rosset, C. (2020). Turing-nlg: A 17-billion-parameter language model by microsoft. *Microsoft Blog*, *1*(2).

Yuan, S., Shuai, Z., Jiahong, L., Zhao, X., Hanyu, Z., & Jie, T. (2022). Wudaomm: A large-scale multi-modal dataset for pre-training models. *arXiv preprint arXiv:2203.11480*.