EXPLORING MULTI-MODAL NEURAL MODELS AND THEIR APPLICATIONS

Lectures on Computational Linguistics

Pisa, May 30th 2023

LUCIA C. PASSARO
ALESSANDRO BONDIELLI

{lucia.passaro,alessandro.bondielli}@unipi.it

OUTLINE

MULTIMODALITY AND NLP(LAB 2)

Part I

Motivation and background

Problems and tasks

Data and pipelines

OUTLINE

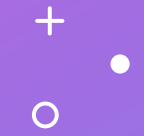
MULTIMODALITY AND NLP(LAB 2)

Part II

Learning strategies for VL models

Specializing pre-trained models

Generative models



Motivation and background

Motivation: humans and modalities

- Humans perceive the world through **many channels**, such as visual, haptic, and auditive
  - Individual channels are typically **incomplete** and **noisy**
  - Humans align and fuse information collected from multiple channels and grasp the **key concepts for interpreting the world**

Multi-modality

- Multi-modality involves the co-occurrence and co-operative use of different semiotic and perceptual modes
- It is a rising concept in current NLP research, since it is relevant from two perspectives:
  - The Cognitive perspective: **mental concepts are multi-modal constructions** (grounded cognition; Barsalou, 2008)
  - The Communication perspective: **communication is inherently multi-modal**, especially on the Web (multi-modal discourse analysis; Benson, 2016)

“One of the core aspirations in AI is to develop algorithms that endow computers with an ability to effectively learn from multi-modal (or, multi-channel) data.”

Gan et al., 2022

Background: intuitions and keywords

- Word / Visual Semantic embeddings
- Semantics and Similarity

Embeddings

"In the neural network literature, **any information "embedded" in a low-dimensional vector space**

Broadly speaking, all implicit distributional vectors are embeddings; in a narrow sense, word embeddings are distributional vectors built with neural networks"

Lenci, 2018

(word) embedding

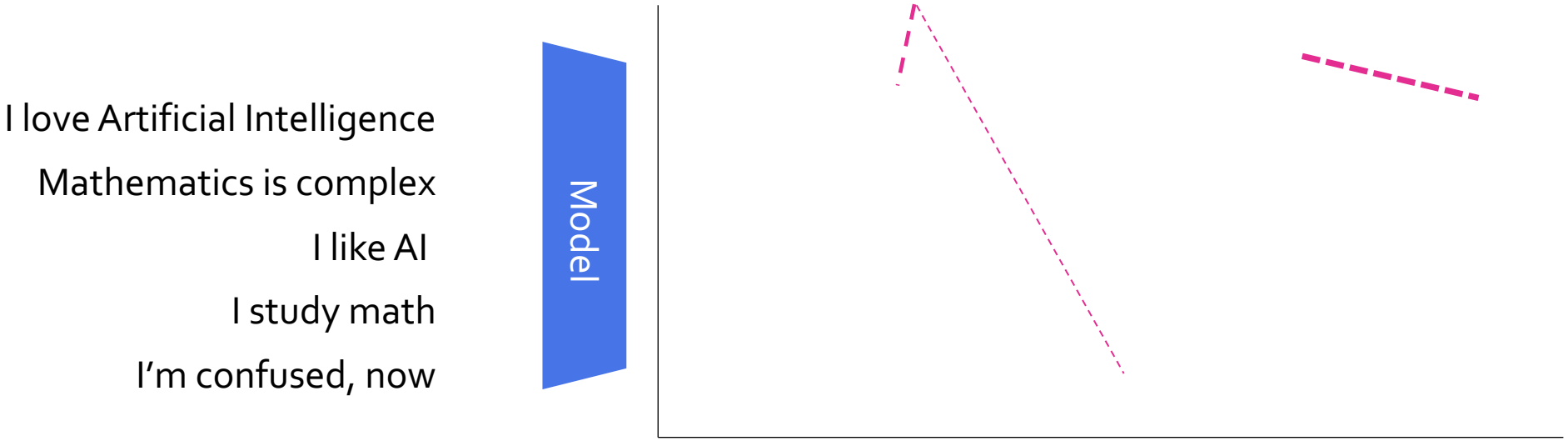
- **Word embeddings** are the most common text encoding technique when working with artificial neural networks
- They are **dense representations** of words (vectors), and are learnt by employing textual information
- **Similar** words (i.e., having a similar **meaning**) have similar embeddings
  - Similarity is measured with **cosine**

(image) embedding

- An image embedding is a low-dimensional representation of an image
- Similar to a word embedding, it is a **dense vector representation of the image** which can be used for many tasks
- Similar images have similar embeddings
  - Similarity is measured with cosine

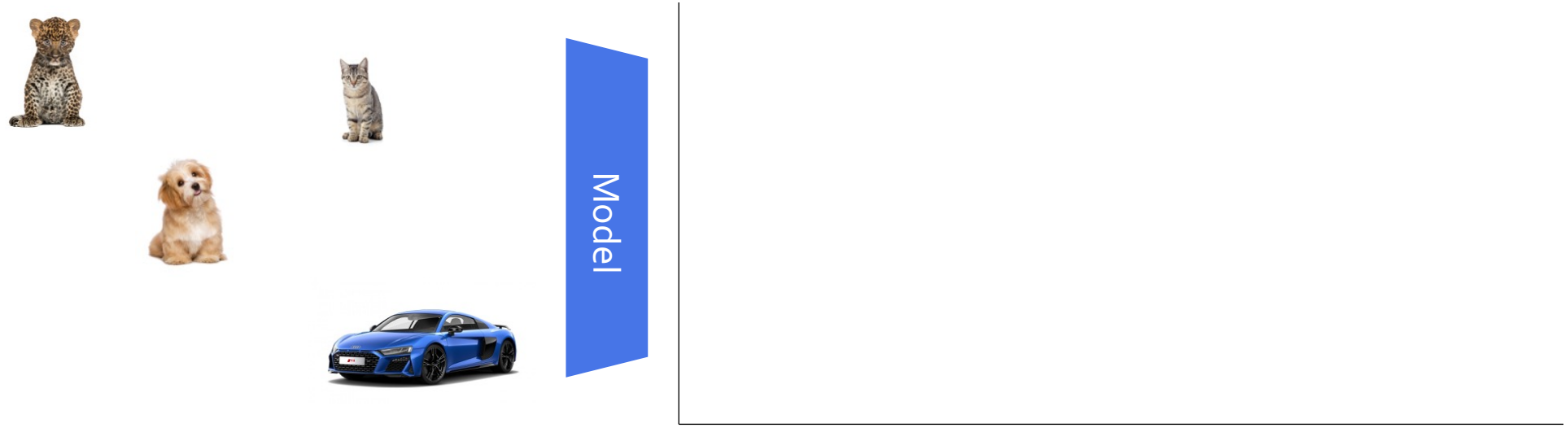
(word) embeddings

- Once the model is built, we can use embeddings to represent linguistic events (tokens, phrases, texts…) within the same distributional space
- This allows for measuring the distance (and the similarity) among the events



(image) embeddings

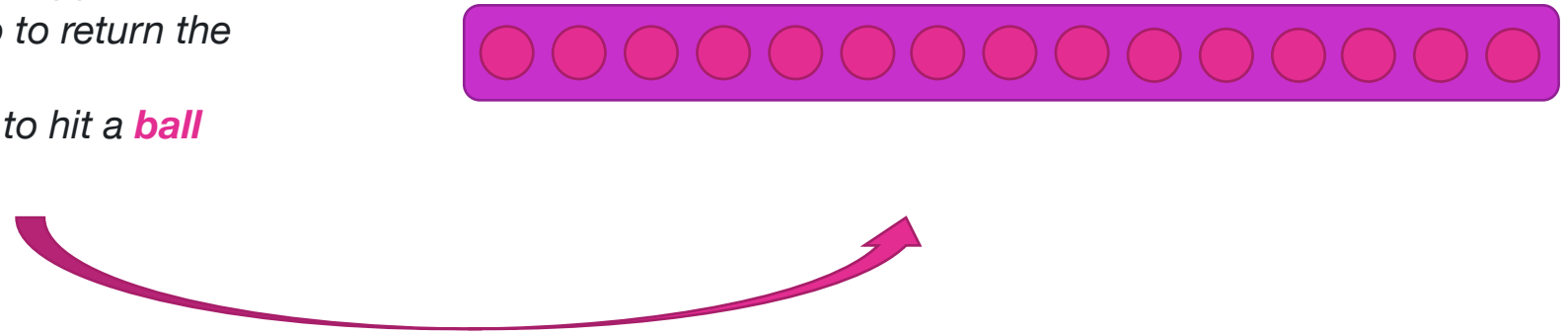
- Intuitively, the same idea has been applied over the years to images
  - Using a model trained on (lots of) visual data
  - We construct a «mathematical» representation of images based on probabilities



(visual-semantic) embeddings

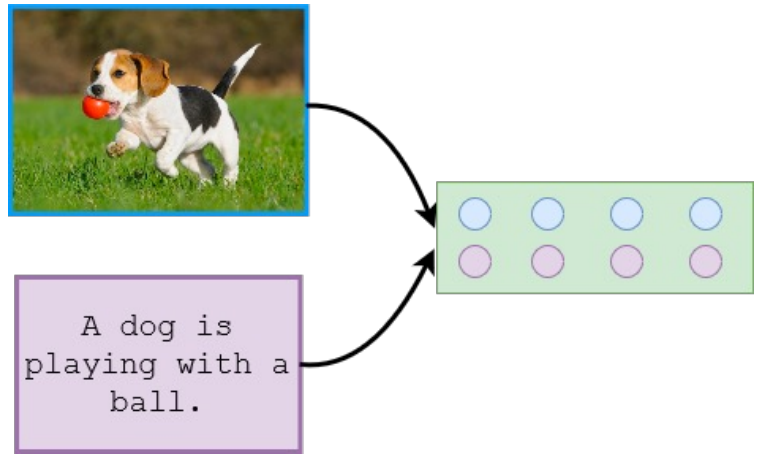
How can word embeddings (and textual representations in general) be enriched with **extra-linguistic (e.g., visual) information**?

…
*a young tennis player preparing to serve a tennis **ball** to her opponent*
*a tennis player leaps up to return the **ball***
*a tennis player is trying to hit a **ball** with a tennis racket*
…



(visual-semantic) embeddings

- **Visual-semantic embeddings** (aka multi-modal embeddings) encode both **visual** and **textual information**
- Textual information can concern either the word or the sentence level



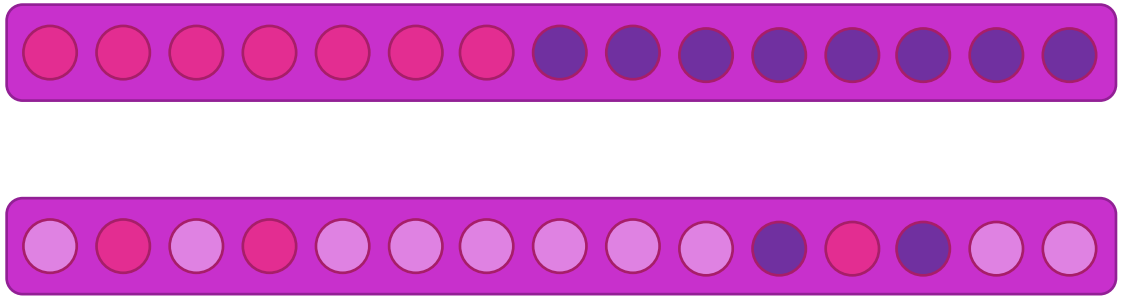
(visual-semantic) embeddings

- Multi-modal embeddings are created by
  - Fusion (**two-stage approach**)
    - Textual and visual features are extracted independently and then “combined”
  - Projection or joint learning (**one-stage approach**)
    - the model "learns" jointly by observing pairs (image, text)
    - representations related to different modalities are learned in a common multi-modal space
    - provides the model with multilayered information that can be transferred to images or texts it has never seen before

(visual-semantic) embeddings

- Multi-modal embeddings encode both visual and textual information
- Textual information can be at both the word (e.g., object) and sentence (e.g., caption) levels

…
*a young tennis player preparing to serve a tennis **ball** to her opponent*
*a tennis player leaps up to return the **ball***
*a tennis player is trying to hit a **ball** with a tennis racket*
…



Problems and Tasks

Vision-Language problems

- A first – broad – classification of multi-modal problems is between:
- Vision-Language **understanding** problems
  - a VL model to select the output from a given list of candidates
- Vision-Language **generation** problems
  - require a VL model to generate the output

Multi-modal problems

- A multi-modal AI system should process such multi-modal data in an effective and efficient way
- Following Gan et al., 2022, Vision-Language tasks can be grouped into three main categories:
  - Image-Text Tasks
  - Computer Vision Tasks as Vision-Language Problems
  - Video-Text Tasks

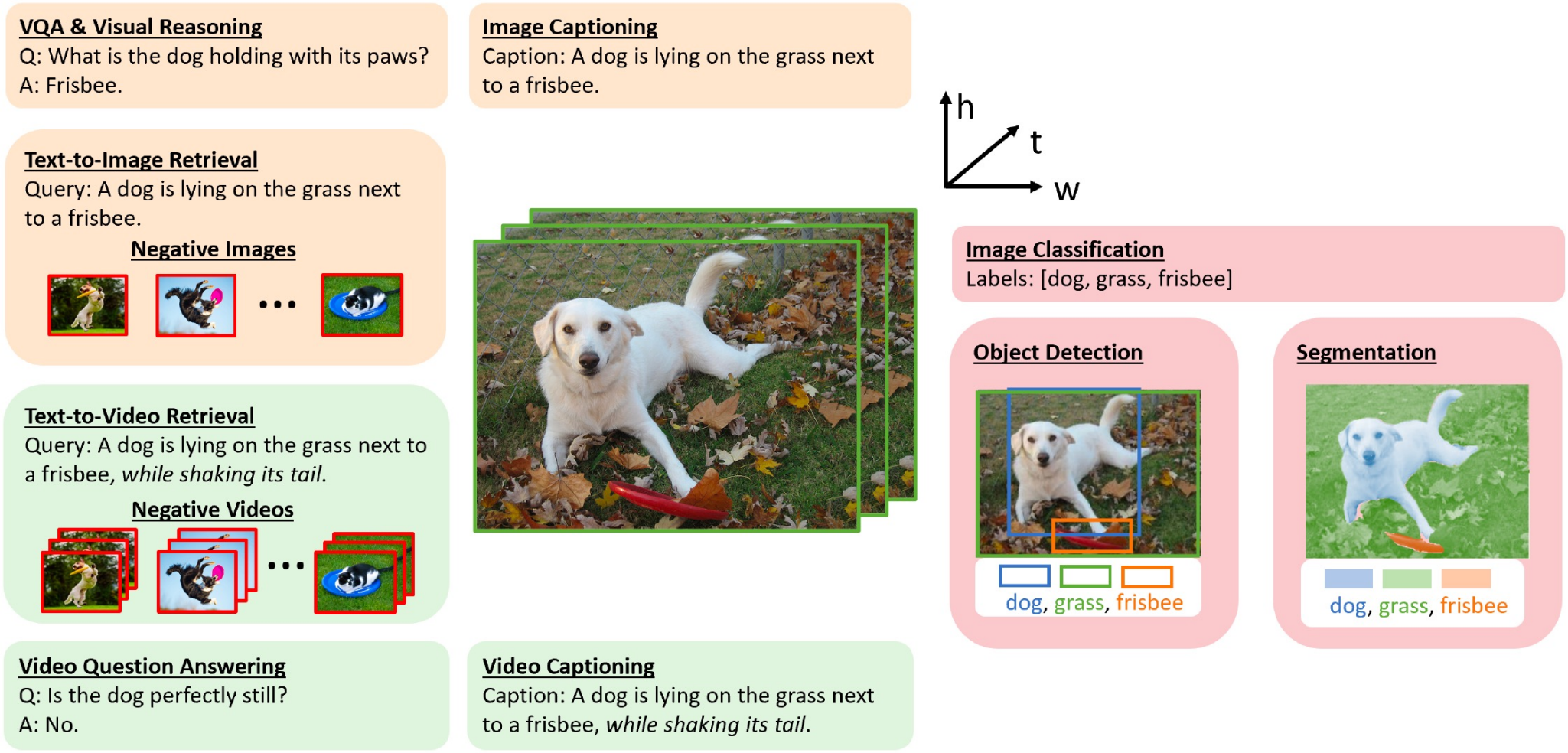


Figure 1.2: Illustration of representative tasks from three categories of VL problems covered in this paper: image-text tasks , vision tasks as VL problems , and video-text tasks .

Image credit: Gan et al., 2022

Image-Text Tasks

The most important and well-studied tasks in VL research are

- image-text retrieval
- image captioning (Vinyals et al., 2015)
- visual question answering (Antol et al., 2015)

Image Captioning
Caption: A dog is lying on the grass next to a frisbee.

VQA & Visual Reasoning
Q: What is the dog holding with its paws?
A: Frisbee.

Text-to-Image Retrieval
Query: A dog is lying on the grass next to a frisbee.

Negative Images

Image-Text Tasks: retrieval-related applications

- **Image-text retrieval**. The models are asked to retrieve the most relevant text (or image) from a large corpus, given the image (or text) query
- **Visual grounding**. Instead of text outputs, requires bounding box outputs, where the model needs to predict the bounding box corresponding to the input text query

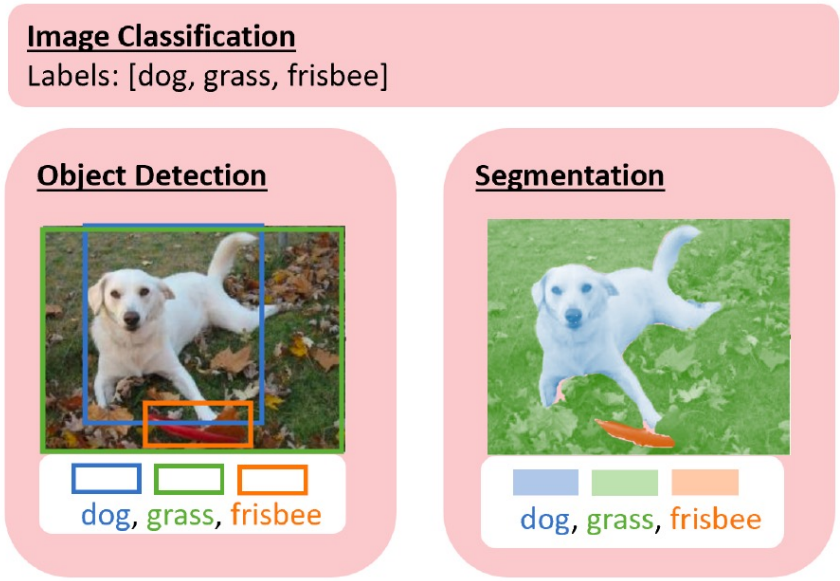
Image-Text Tasks: VQA-related applications

- VQA and visual reasoning models are asked to provide **open-ended free-form texts**, or **selecting an option** from multiple choices:
  - visual reasoning (Hudson and Manning, 2019b; Suhr et al., 2019)
  - visual commonsense reasoning (Zellers et al., 2019)
  - visual dialog (Das et al., 2017)
  - knowledge-based VQA (Marino et al., 2019)
  - scene-text-based VQA (Singh et al., 2019)

Image-Text Tasks: Captioning-related applications

- The typical setting in captioning is generating a short sentence (Lin et al., 2014)
- Related tasks include:
  - image **paragraph** captioning (Krause et al., 2017)
  - **scene-text-based** image captioning (Sidorov et al., 2020)
  - visual **storytelling** (Huang et al., 2016)
- **Text-to-image generation** is a dual task of image captioning: the system is required to **create an image based on the text input**

"Aided" Computer Vision Tasks



- **Core vision problems** such as image classification, object detection, and segmentation **benefit of language supervision**
- Instead of treating the supervision signals (e.g., class labels) as one-hot vectors, the **semantic meaning of the labels** is exploited to generalize the traditional close-set classification or detection models to recognizing **unseen concepts** (i.e., open-vocabulary object detection)

Video-Text Tasks

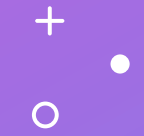
Text-to-Video Retrieval
Query: A dog is lying on the grass next to a frisbee, *while shaking its tail.*

Negative Videos

Video Question Answering
Q: Is the dog perfectly still?
A: No.

Video Captioning
Caption: A dog is lying on the grass next to a frisbee, *while shaking its tail.*

- All aforementioned image-text tasks have their video-text counterparts, such as video captioning, retrieval, and question answering
- AI systems are not only asked to capture **spatial information** within a single video frame, but also the **temporal dependencies** among **frames**



Multi-modal datasets

Multi-modal datasets

- Vision-language models are pre-trained on large multi-modal datasets
- Text data typically include **human-generated captions**, object labels, image metadata and other **information relevant for specific tasks** (e.g., questions and answers)

Trends on multi-modal data

We focus on general-purpose datasets for Vision-Language Models!

- **Pioneering** multi-modal datasets were designed for training a model on **specific tasks** such as captioning, VQA, object detection etc.
- Nowadays, following the trend observed for Large Language Models VL models are typically:
  - **pre-trained** on a vast number of multi-modal data
  - and then **fine-tuned** (or used under a **zero** / **few shot setting**)

Pioneering multi-modal datasets

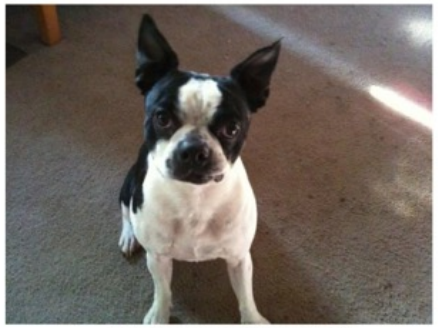
- SBU Captions (Ordonez et al., 2011), Conceptual Captions (CC3M, Sharma et al., 2018) and Flickr30K (Young et al., 2014): **captioning datasets**
  - datasets scraped from the web with accompanying captions
- COCO (Chen et al., 2015) : **object detection** and **image captioning**
  - consists of image instances with object labels and natural sentence descriptions.
- Visual Genome, VG (Krishna et al., 2017): **object detection**, **region descriptions** and **relationships**, **questions & answers**
  - contains objects, attributes, and relationships within each image

SBU Captions, Conceptual Captions and Flickr30K

Amazing colours in the sky at sunset with the orange of the cloud and the blue of the sky behind.

A female mallard duck in the lake at Luukki Espoo

Fresh fruit and vegetables at the market in Port Louis Mauritius.



Street dog in Lijiang

Tree with red leaves in the field in autumn.

One monkey on the tree in the Ourika Valley Morocco

Clock tower against the sky.

The river running through town I cross over this to get to the train

Strange cloud formation literally flowing through the sky like a river in relation to the other clouds out there.

The sun was coming through the trees while I was sitting in my chair by the river

Image credit: Ordonez et al., 2011

SBU Captions, **Conceptual Captions** and Flickr30K

Alt-text: A Pakistani worker helps to clear the debris from the Taj Mahal Hotel November 7, 2005 in Balakot, Pakistan.

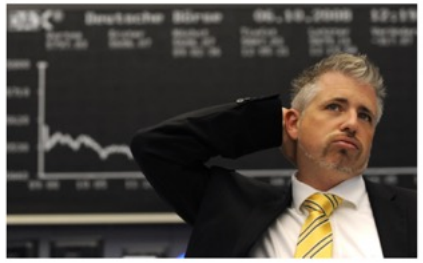
Conceptual Captions: a worker helps to clear the debris.

Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

Conceptual Captions: pop artist performs at the festival in a city.

Image credit: Sharma et al., 2018

SBU Captions, Conceptual Captions and **Flickr30K**



Gray haired man in black suit and yellow tie working in a financial environment.
A graying man in a suit is perplexed at a business meeting.
A businessman in a yellow tie gives a frustrated look.
A man in a yellow tie is rubbing the back of his neck.
A man with a yellow tie looks concerned.

A butcher cutting an animal to sell.
A green-shirted man with a butcher's apron uses a knife to carve out the hanging carcass of a cow.
A man at work, butchering a cow.
A man in a green t-shirt and long tan apron hacks apart the carcass of a cow while another man hoses away the blood.
Two men work in a butcher shop; one cuts the meat from a butchered cow, while the other hoses the floor.

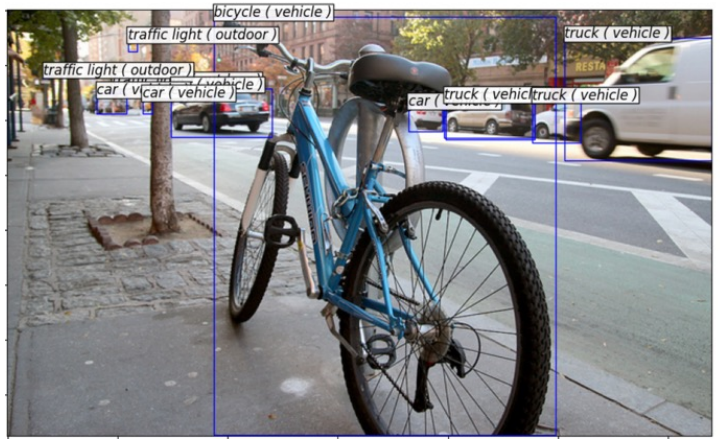
Image credit: Young et al., 2014

COCO

Captions (5)
a blue bike parked on a side walk
a bicycle is chained to a fixture on a city street
a blue bicycle sits on a sidewalk near a street.
a bicycle is locked up to a post
a bike sits parked next to a street

Objects

19765 traffic light
19766 traffic light
19767 traffic light
20403 bicycle
20896 car
20897 car
20898 car
20899 car
20900 car
22828 truck
22829 truck
22830 truck



Segmentation

Key points

Visual Genome (VG)

- **scene graphs contain:**
  - images
  - objects (provided with synsets)
  - relations (provided with synsets)
  - region descriptions

    (in natural language)
  - questions & answers

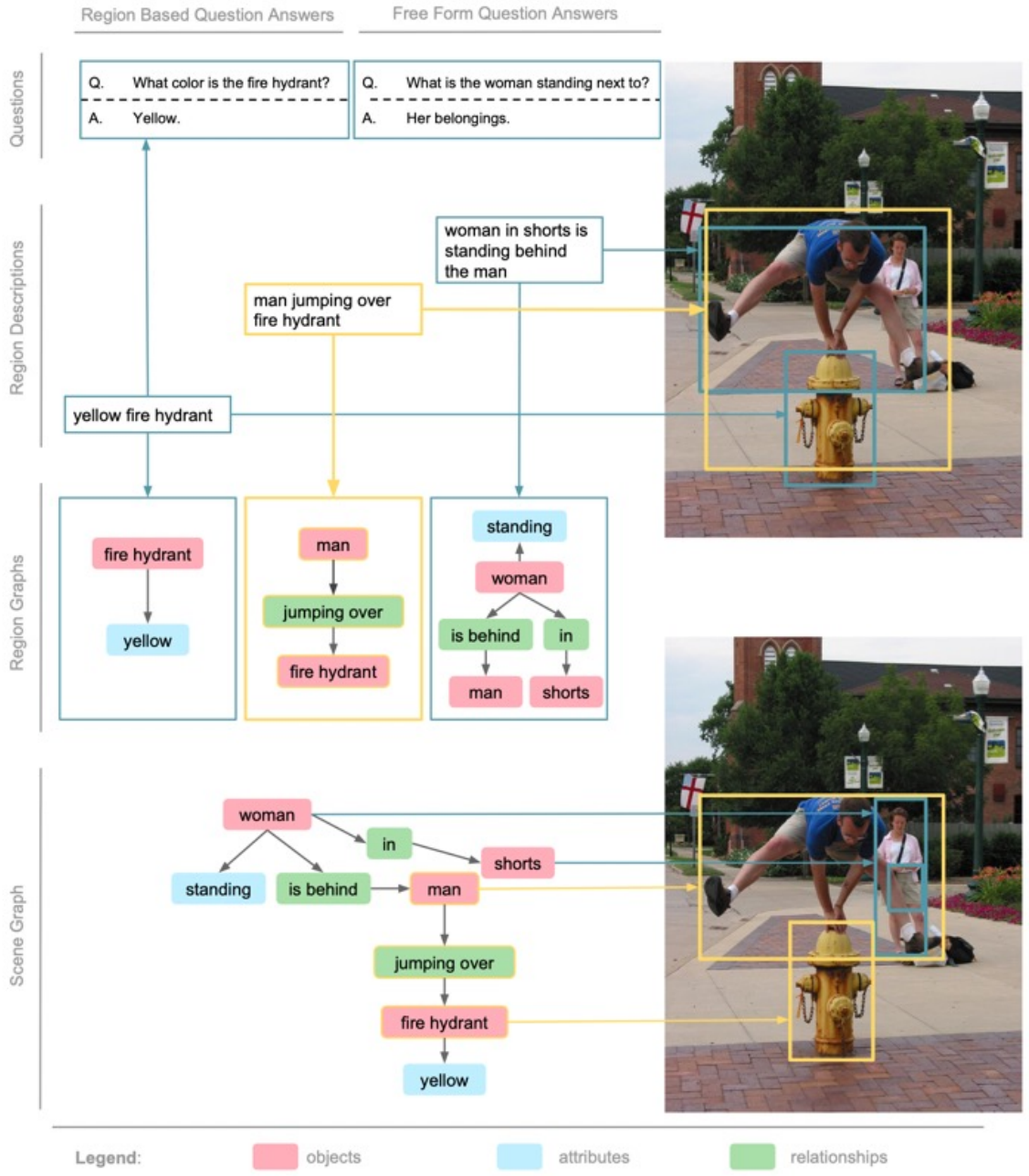


Image credit: Krishna et al., 2017

Visual Genome (VG)

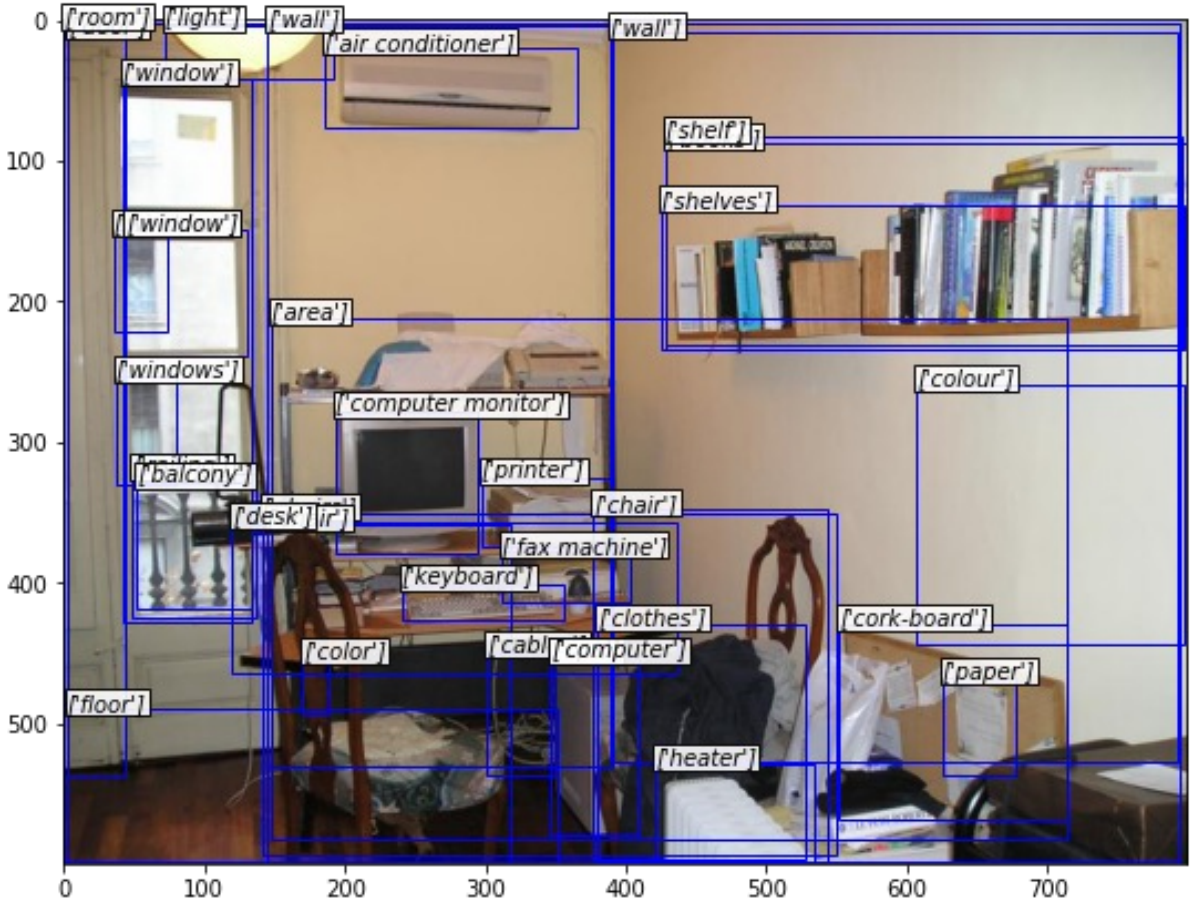
- **scene graphs contain:**
  - **images**
  - objects (provided with synsets)
  - relations (provided with synsets)
  - region descriptions

    (in natural language)
  - questions & answers

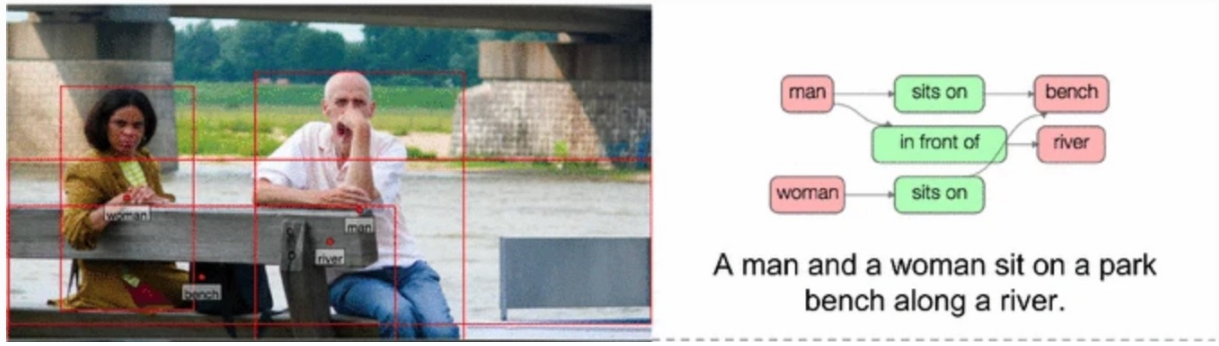
Visual Genome (VG)

- **scene graphs contain:**
  - images
  - **objects** (provided with synsets)
  - relations (provided with synsets)
  - region descriptions

    (in natural language)
  - questions & answers



Visual Genome (VG)



- **scene graphs contain:**
  - images
  - objects (provided with synsets)
  - **relations** (provided with synsets)
  - region descriptions

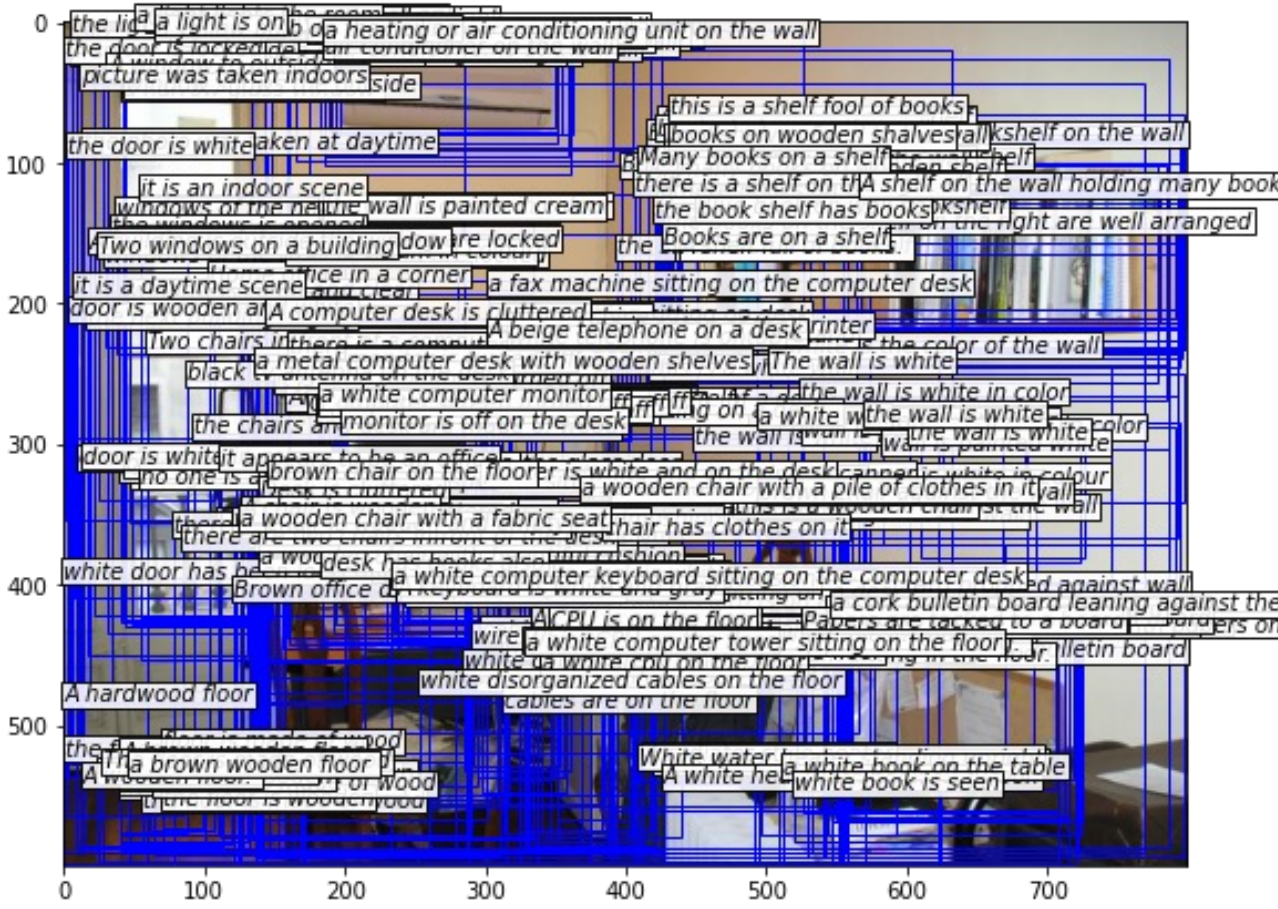
    (in natural language)
  - questions & answers

... {'synsets': ['along.r.01'], 'predicate': 'ON', 'relationship_id': 3187870, 'object_id': 1060360, 'subject_id': 1060364}, {'synsets': ['along.r.01'], 'predicate': 'ON', 'relationship_id': 3187871, 'object_id': 1060364, 'subject_id': 1060345}, {'synsets': [], 'predicate': 'with', 'relationship_id': 3187872, 'object_id': 1060351, 'subject_id': 1060375}, {'synsets': ['earlier.r.01'], 'predicate': 'before', 'relationship_id': 3187873, 'object_id': 1060368, 'subject_id': 1060342}, {'synsets': ['along.r.01'], 'predicate': 'ON', 'relationship_id': 3187874, 'object_id': 1060360, 'subject_id': 1060364}, {'synsets': ['along.r.01'], 'predicate': 'ON' 'relationship_id': 3187875, 'object_id': 1060376, 'subject_id': 1060365}, {'synsets': ['along.r.01'], 'predicate': 'ON', 'relationship_id': 3187876, 'object_id': 1060371, 'subject_id': 1060377}, {'synsets': ['along.r.01'], 'predicate': 'ON', 'relationship_id': 3187877, 'object_id': 1060359, 'subject_id': 1060378}, {'synsets': [], 'predicate': 'against', 'relationship_id': 3187878, 'object_id': 1060360, 'subject_id': 1060348}, {'synsets': ['along.r.01'], 'predicate': 'ON', 'relationship_id': 3187879, 'object_id': 1060374, 'subject_id': 1060362}, {'synsets': [], 'predicate': 'against', 'relationship_id': 3187880, 'object_id': 1060360, 'subject_id': 1060371} ...

Visual Genome (VG)

- **scene graphs contain:**
  - images
  - objects (provided with synsets)
  - relations (provided with synsets)
  - **region descriptions**

    (in natural language)
  - questions & answers



Visual Genome (VG)

- **scene graphs contain:**
  - images
  - objects (provided with synsets)
  - relations (provided with synsets)
  - region descriptions

    (in natural language)
  - **questions & answers**

Visual Genome (VG)

Image credit: Krishna et al., 2017

The era of Large Vision-Language Models

- Vision-language models are trained on image and text datasets with different structures on the pre-training objective.
- Pre-trained models are fine-tuned on downstream tasks with task-specific datasets.

Pre-Training Datasets

- In a **typical academic setting**, VL models are pre-trained on a collection of four commonly used image-caption datasets:
  - COCO (Chen et al., 2015)
  - Visual Genome (VG) (Krishna et al., 2017)
  - Conceptual Captions (CC3M) (Sharma et al., 2018)
  - SBU Captions (Ordonez et al., 2011)

Pre-Training Datasets

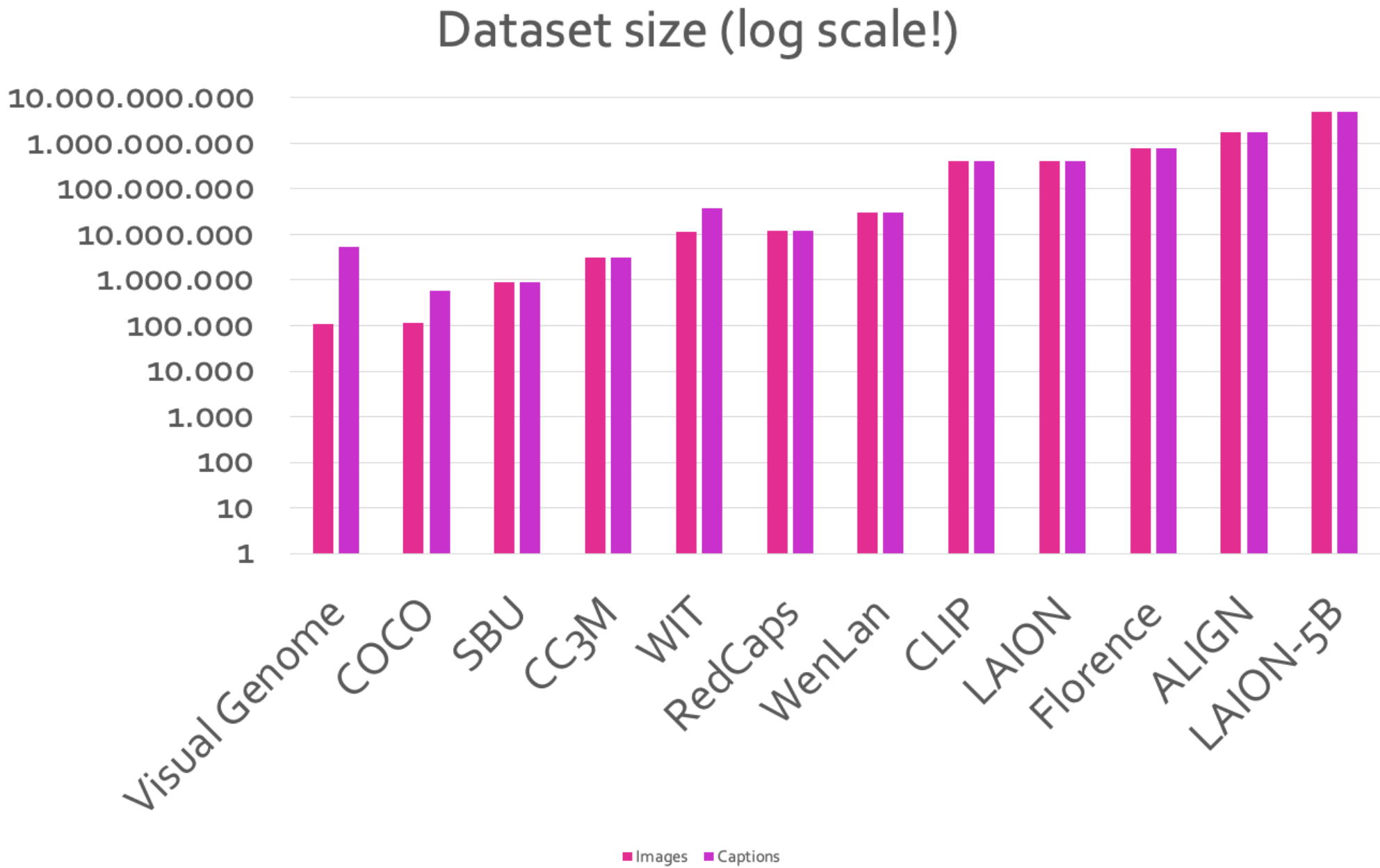
- In a **typical industrial setting**, datasets are **large scale** and **web-crawled**. Most of them are proprietary. Some of them are:
  - The dataset used in CLIP (Radford et al., 2021) consists of 400 M image-text pairs, PROPRIETARY
  - The dataset used in ALIGN (Jia et al., 2021) has 1.8 B image-text pairs, PROPRIETARY
  - The dataset used in Florence (Yuan et al., 2021) and GIT (Wang et al., 2022d) contains 800 M image-text pairs, PROPRIETARY
  - WenLan (Huo et al., 2021) consists of 30 million image-text pairs (web-collected exploiting topic models, elaborate cleaning process), PROPRIETARY

Pre-Training Datasets

- In a **typical industrial setting**, datasets are **large scale** and **web-crawled**.

  Most of them are proprietary. Some of them are:

  - LAION-400M/5B (Schuhmann et al., 2021) has 400 M or 5 B image-text pairs, PUBLIC
  - RedCaps (Desai et al., 2021) includes 12 M image-text pairs (subreddits), PUBLIC
  - The Wikipedia-based Image-Text Dataset (WIT) (Srinivasan et al., 2021) is composed of 11.5 M unique images and 37.6 M texts (multilingual, across 108 languages), PUBLIC


Dataset size (log scale!)
10.000.000.000
1.000.000.000
100.000.000
10.000.000
1.000.000
100.000
10.000
1.000
100
10
1
Visual Genome
COCO
SBU
CC3M
WIT
RedCaps
WenLan
CLIP
LAION
Florence
ALIGN
LAION-5B
Images
Captions

Case study:

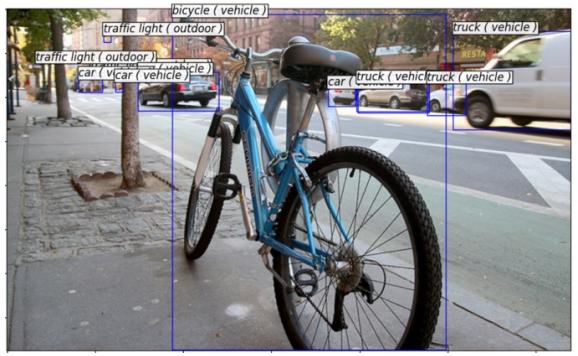
The COCO dataset

Exploring COCO

Captions (5)
a blue bike parked on a side walk
a bicycle is chained to a fixture on a city street
a blue bicycle sits on a sidewalk near a street.
a bicycle is locked up to a post
a bike sits parked next to a street

Objects

19765 traffic light
19766 traffic light
19767 traffic light
20403 bicycle
20896 car
20897 car
20898 car
20899 car
20900 car
22828 truck
22829 truck
22830 truck



Notebook:
Exploring COCO data from scratch!

Segmentation

Datasets and models

Hugging Face

Hugging Face

A **community** and data science **platform** offering:

- Advanced Tools: Empowering users to create, train, and deploy machine learning models using open-source code and technologies.
- **Collaborative Space**: A platform that brings together a diverse community of data scientists, researchers, and ML engineers to exchange ideas, receive support, and actively contribute to open-source projects.

… for our purposes

- Hugging Face is very useful for accessing available
  - Pre-Trained (multi-modal) **models**
    - Both **foundational** and **specialized** on tasks
  - **Datasets**
    - With a common data format, ready to be used with the pre-trained models

Hugging Face Datasets

- A library for accessing and sharing **datasets** (Text, Audio, Vision, Multi-modal)
  - **Integrated** with the Hugging Face Hub
- Datasets can be easily loaded
  - single line of code
- Apache Arrow format
  - enables large amounts of data to be processed and moved quickly
  - language-agnostic, column-oriented, supported by ML tools
  - stores data in a columnar memory layout (many, possibly nested, column types)

Notebook:
The COCO Dataset in HuggingFace

Hugging Face Pipelines

- Using the pipelines is the simplest way to use **pretrained models** for inference.
- Both existing models (available in huggingface) and to your own can be used.
- They deal with both **uni-modal** and **multi-modal tasks**.
- A pipeline is a wrapper built for working around all the other available pipelines.

  It can be executed on:
  - a single item
  - a list of items (list)
  - a Dataset

Notebook:
Using Hugging Face Pipelines for inference

MAY 30TH 2023

LCL 2023

THANK YOU FOR YOUR ATTENTION

Notebooks and Slides here +

References

References

- Barsalou, Lawrence W. (2008). "Grounded Cognition". In: Annual Review of Psychology 59.1, pp. 617–645.
- Benson, Phil (2016). The Discourse of YouTube: Multimodal Text in a Global Context. London; New York: Routledge.
- Gan, Z., Li, L., Li, C., Wang, L., Liu, Z., & Gao, J. (2022). Vision-language pre-training: Basics, recent advances, and future trends. Foundations and Trends® in Computer Graphics and Vision, 14(3–4), 163-352.
- Lenci, A. (2018). Distributional models of word meaning. Annual review of Linguistics, 4, 151-171.
- Huo, Y., Zhang, M., Liu, G., Lu, H., Gao, Y., Yang, G., Wen, J., Zhang, H., Xu, B., Zheng, W., et al. (2021). Wenlan: Bridging vision and language by large-scale multi-modal pre-training. arXiv preprint arXiv:2103.06561.

References

- Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In CVPR.
- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). Vqa: Visual question answering. In ICCV.
- Hudson, D. A. and Manning, C. D. (2019b). Gqa: A new dataset for real-world visual reasoning and compositional question answering. In CVPR.
- Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., and Artzi, Y. (2019). A corpus for reasoning about natural language grounded in photographs. In ACL.

References

- Zellers, R., Bisk, Y., Farhadi, A., and Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. In CVPR.
- Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., Parikh, D., and Batra, D. (2017). Visual dialog. In CVPR.
- Marino, K., Rastegari, M., Farhadi, A., and Mottaghi, R. (2019). Ok-vqa: A visual question answering benchmark requiring external knowledge. In CVPR.
- Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. (2019). Towards vqa models that can read. In CVPR.

References

- Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2, 67-78.
- Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Doll´ar, P., and Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV.
- Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In ACL.

References

- Ordonez, V., Kulkarni, G., and Berg, T. (2011). Im2text: Describing images using 1 million captioned photographs. In NeurIPS.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In ICML.
- Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In ICML.
- Srinivasan, K., Raman, K., Chen, J., Bendersky, M., and Najork, M. (2021). Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. arXiv preprint arXiv:2103.01913.

References

- Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. (2021). Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114.
- Desai, K., Kaul, G., Aysola, Z., and Johnson, J. (2021). Redcaps: Web-curated image-text data created by the people, for the people. In NeurIPS, Track on Datasets and Benchmarks.
- Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., Liu, C., Liu, M., Liu, Z., Lu, Y., Shi, Y., Wang, L., Wang, J., Xiao, B., Xiao, Z., Yang, J., Zeng, M., Zhou, L., and Zhang, P. (2021). Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432.
- Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., and Wang, L. (2022d). Git: A generative image-to-text transformer for vision and language. arXiv preprint arXiv:2205.14100.