Deep Representation: Building a Semantic Image Search Engine

Emmanuel Ameisen

INSIGHT

PINTEREST SEARCH

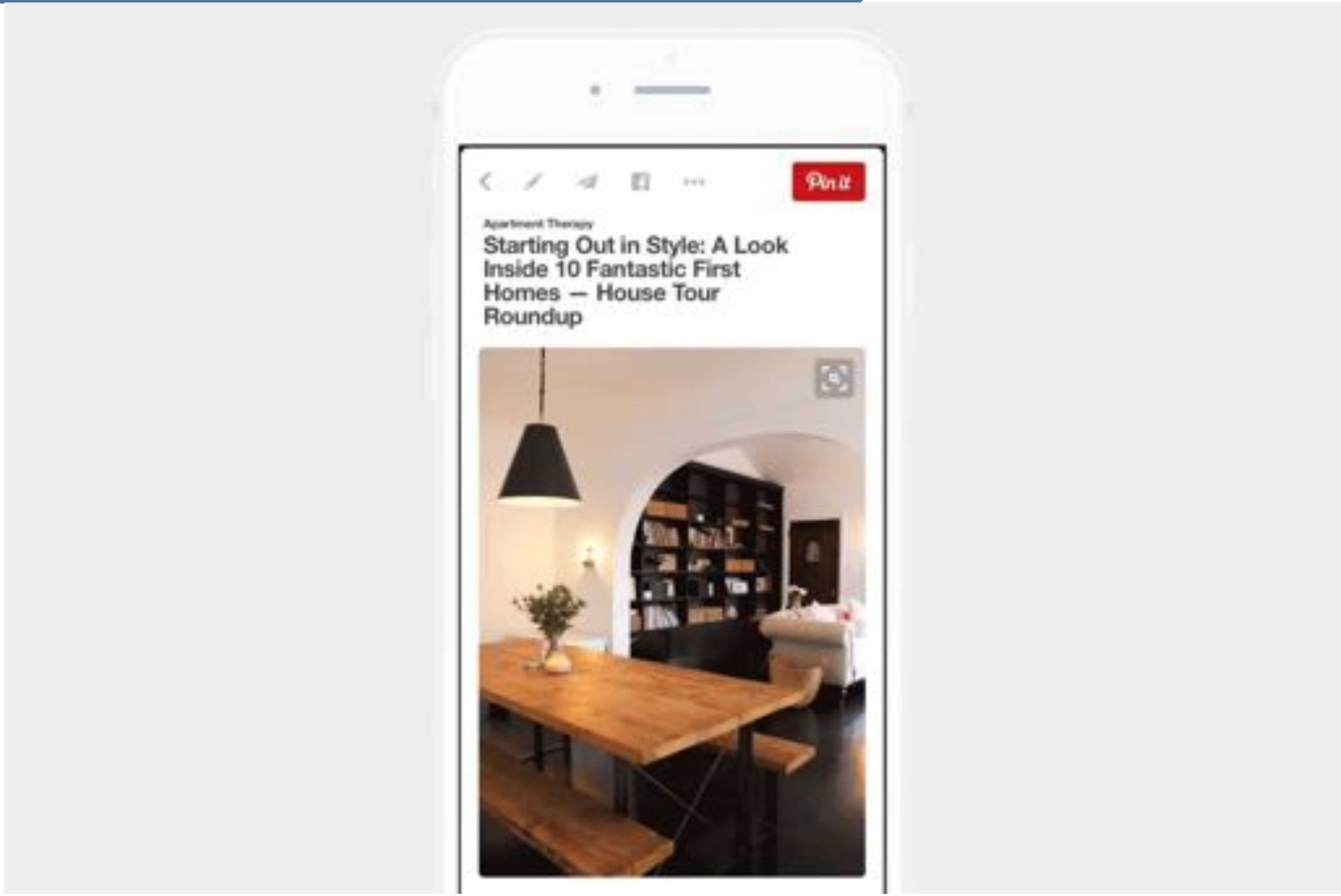


IMAGE SEARCH ENGINE

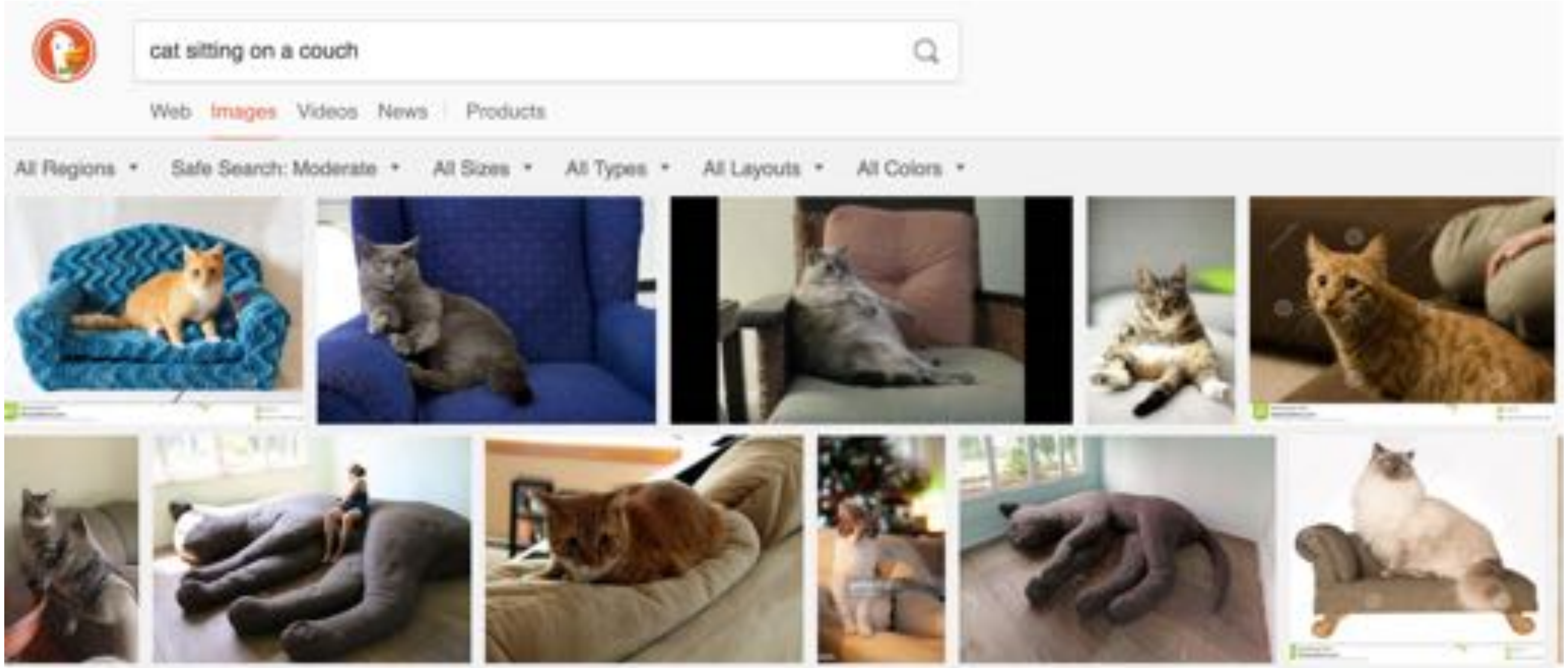


IMAGE TAGGING

thenextweb.com

BACKGROUND

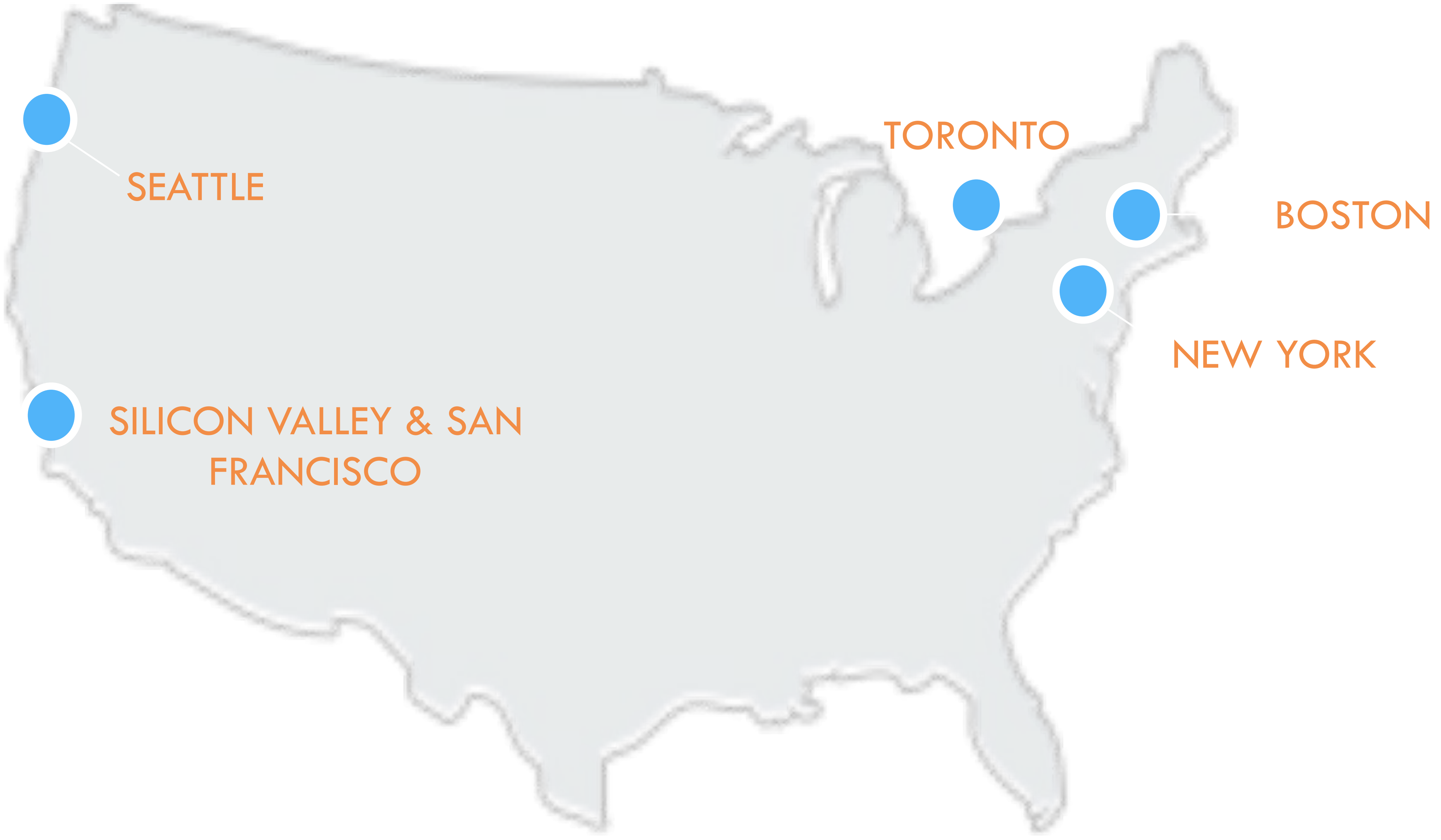
- Why am I speaking about this?

INSIGHT

ABOUT INSIGHT

7-Week Fellowship in

- DATA SCIENCE
- DATA ENGINEERING
- HEALTH DATA
- ARTIFICIAL INTELLIGENCE
- PRODUCT MANAGEMENT
- DEVOPS



+ REMOTE

www.insightdata.ai

INSIGHT

INSIGHT DATA – FELLOW PROJECTS

FASHION CLASSIFIER

AUTOMATIC REVIEW GENERATION

9/24/2017

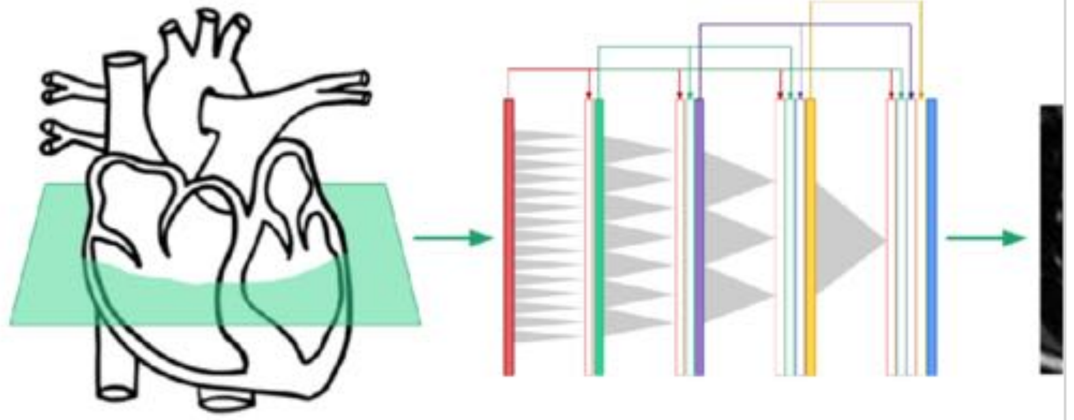
Great service! The place is very relaxed. The curry is outstanding. I am always satisfied with the food and the ambiance.

READING TEXT IN VIDEOS

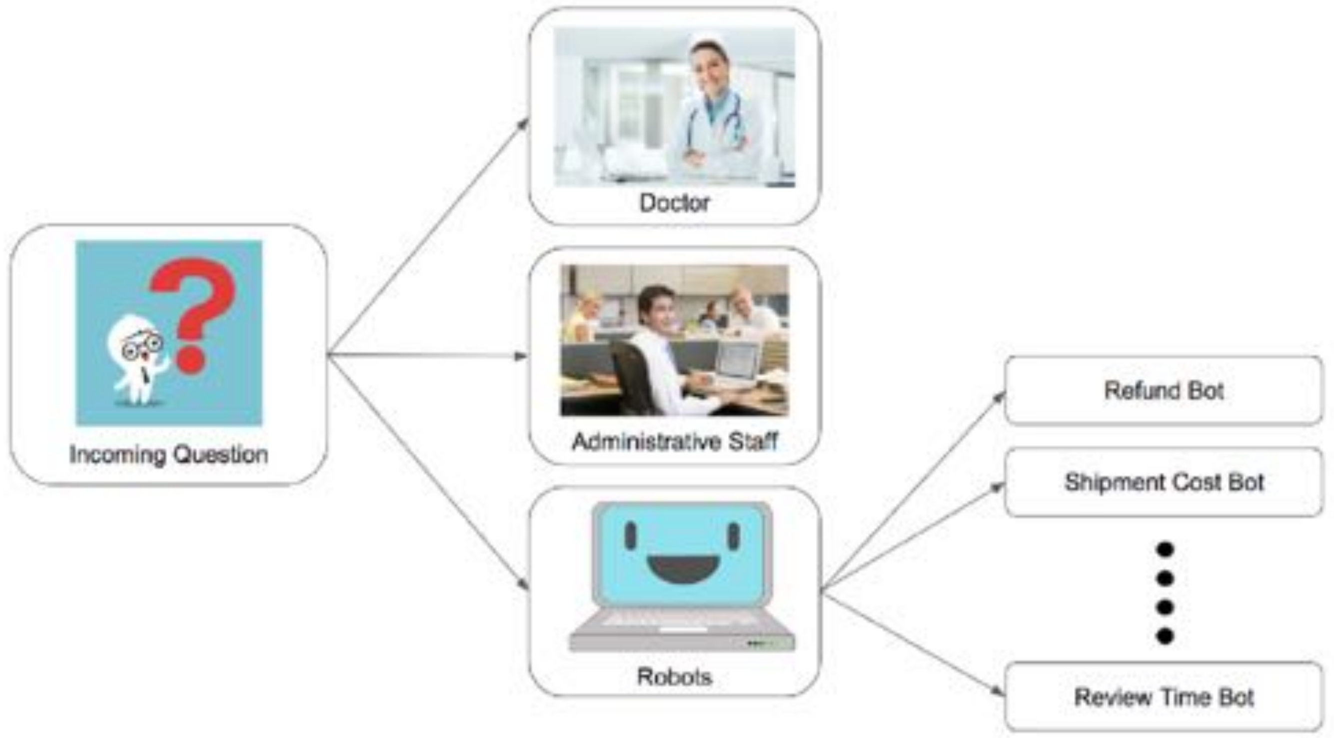
HEART SEGMENTATION

Heart Disease Diagnosis with Deep Learning

State-of-the-art results with 60x fewer parameters

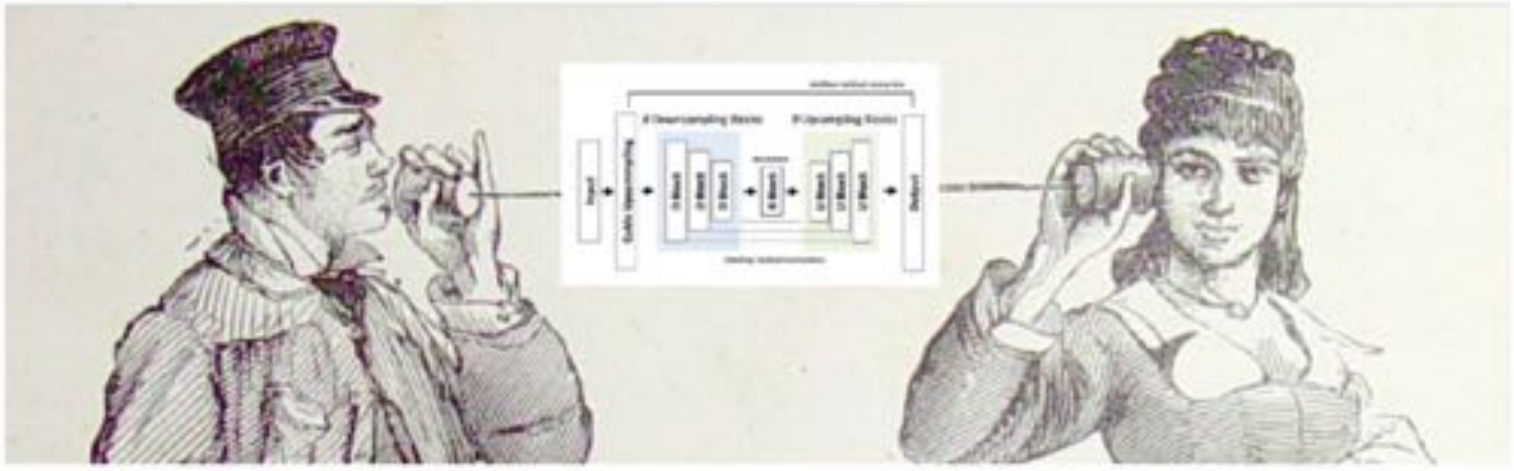


SUPPORT REQUEST CLASSIFICATION



SPEECH UNSAMPLING

Using Deep Learning to Reconstruct High-Resolution Audio



INSIGHT

INSIGHT FELLOWS ARE DATA SCIENTISTS AND DATA ENGINEERS EVERYWHERE

400+

COMPANIES

ON THE MENU

- A quick overview of Computer Vision (CV) tasks and challenges
- Natural Language Processing (NLP) tasks and challenges
- Challenges in combining both
- Representations learning in CV
- Representation learning in NLP
- Combining both

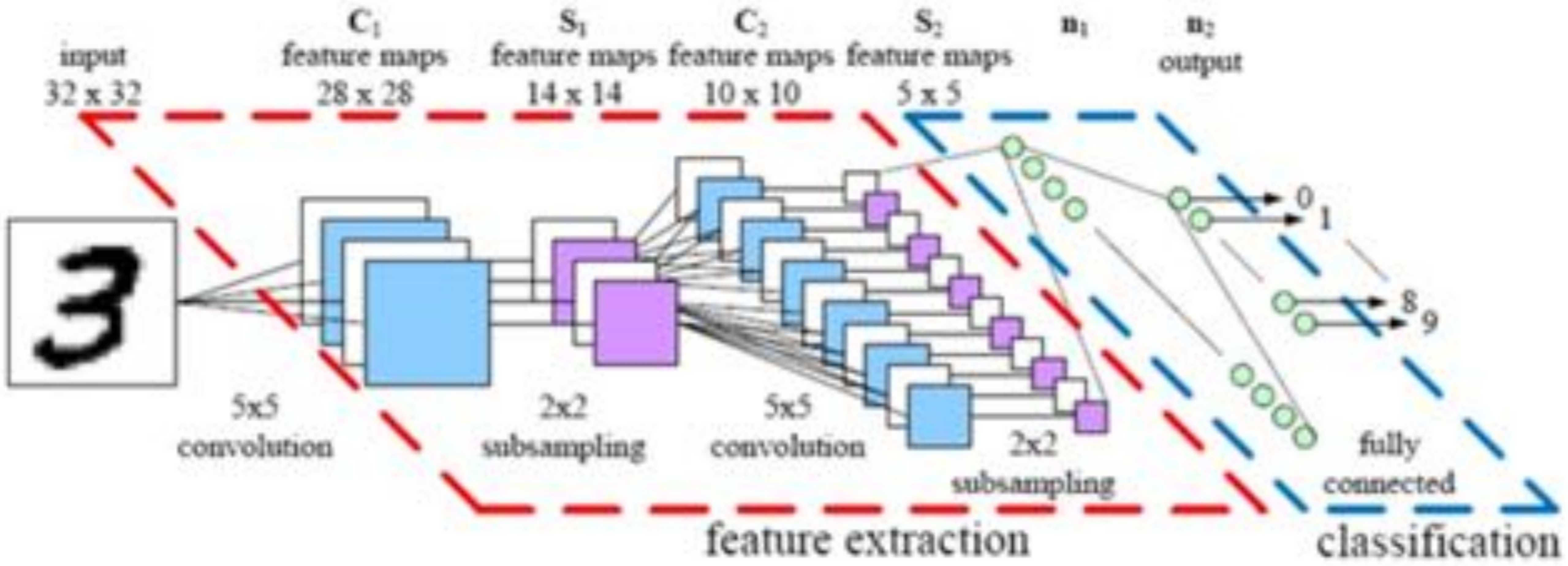
ON THE MENU

- A quick overview of Computer Vision (CV) tasks and challenges
- Natural Language Processing (NLP) tasks and challenges
- Challenges in combining both
- Representations learning in CV
- Representation learning in NLP
- Combining both

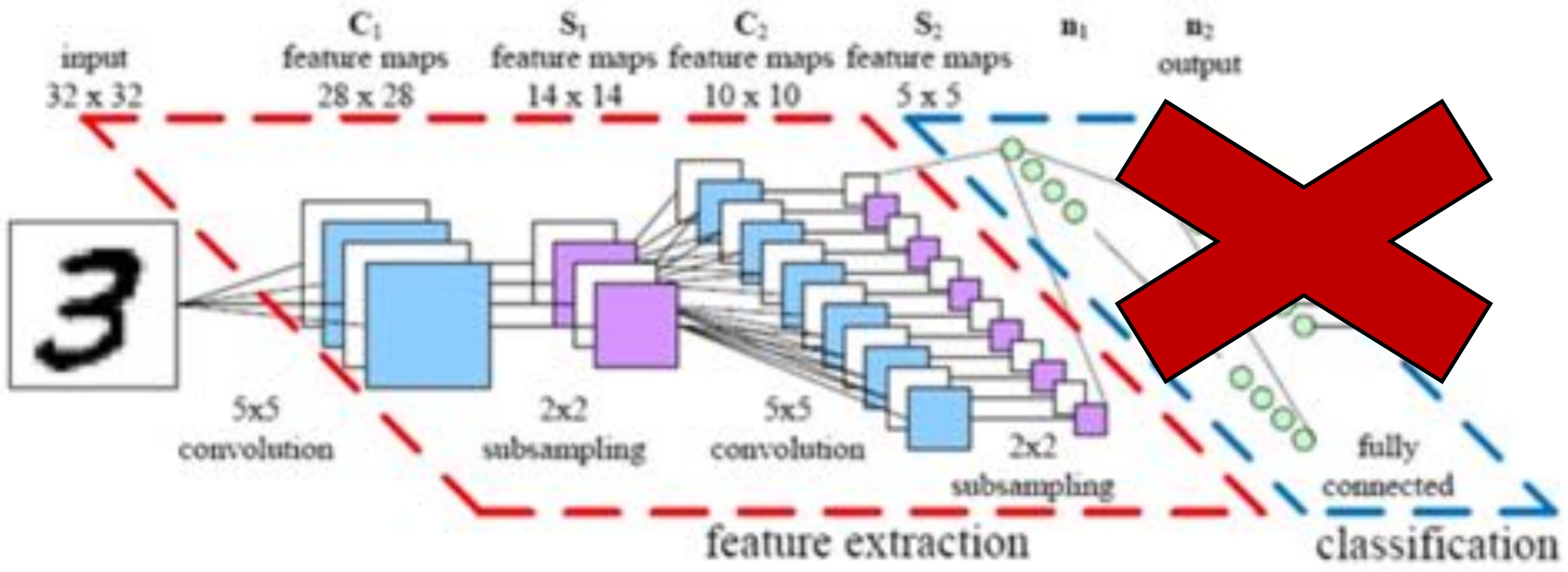
CONVOLUTIONAL NEURAL NETWORKS (CNN)

- Massive models
  - Dataset of 1M+images
  - For multiple days
- Automates feature engineering

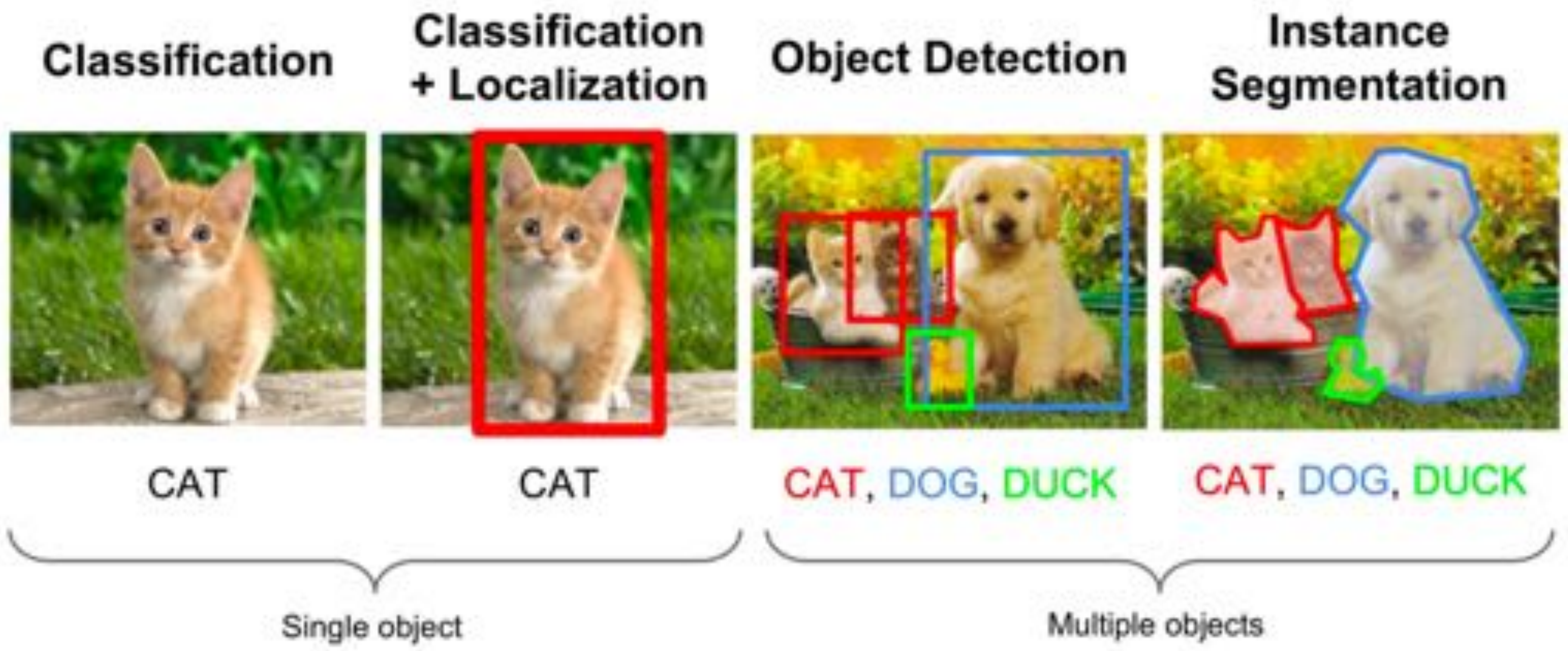
- Use cases
  - Fashion
  - Security
  - Medicine
  - …



EXTRACTING INFORMATION



- Incorporates local and global information
- Use cases
  - Medical
  - Security
  - Autonomous Vehicles

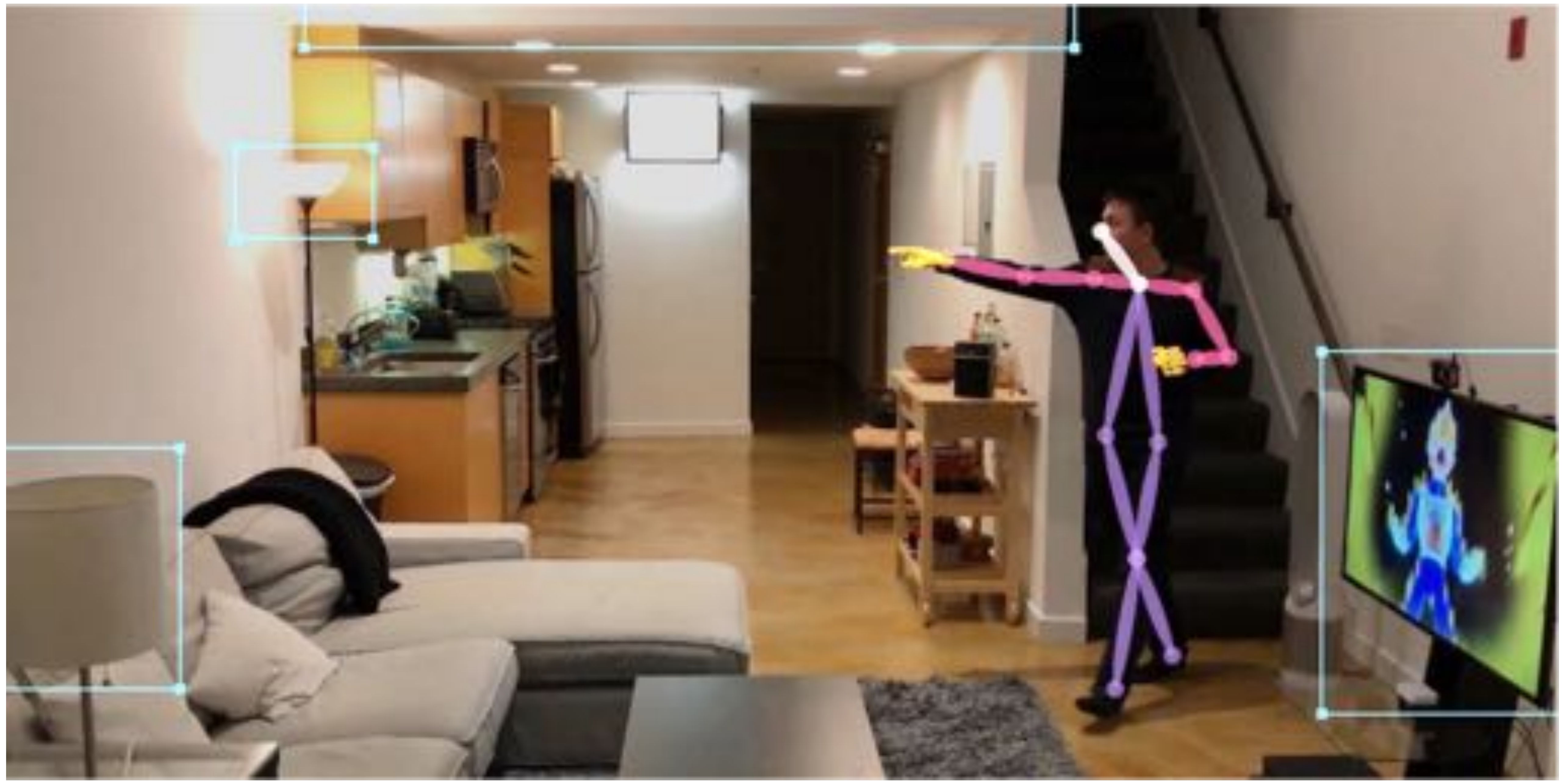


@arthur_ouaknine

ADVANCED APPLICATIONS

- Pose Estimation
- Scene Parsing
- 3D Point cloud estimation

Insight Fellow Project with Piccolo



Felipe Mejia

ON THE MENU

- A quick overview of Computer Vision (CV) tasks and challenges
- Natural Language Processing (NLP) tasks and challenges
- Challenges in combining both
- Representations learning in CV
- Representation learning in NLP
- Combining both

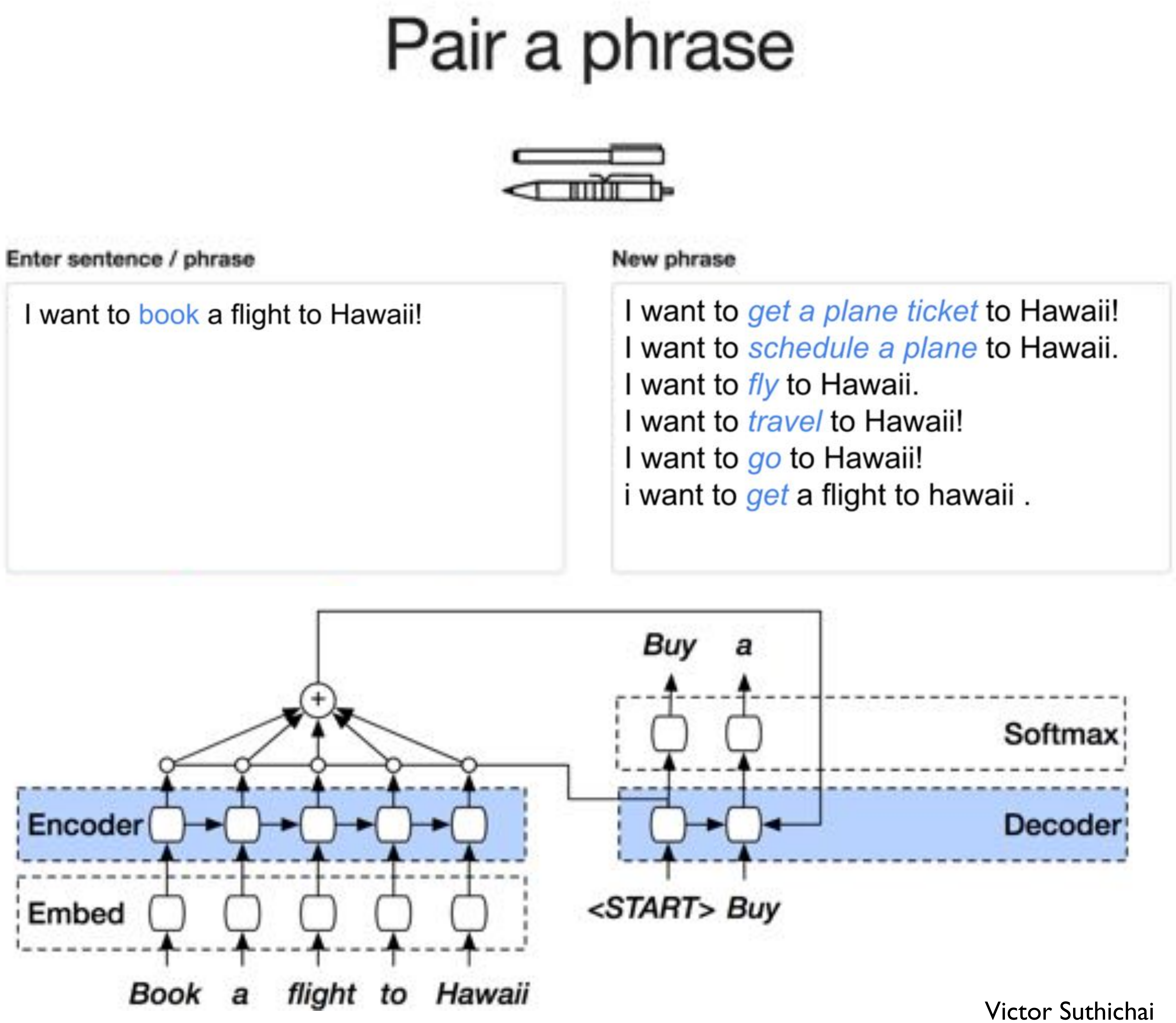
NLP

- Traditional NLP tasks
  - Classification (sentiment analysis, spam detection, code classification)
- Extracting Information
  - Named Entity Recognition, Information extraction
- Advanced applications
  - Translation, sequence to sequence learning

SENTENCE PARAPHRASING

- Sequence to sequence models are still often too rough to be deployed, even with sizable datasets
  - Recognized Tosh as a swear word

- They can be used efficiently for data augmentation
  - Paired with other latent approaches



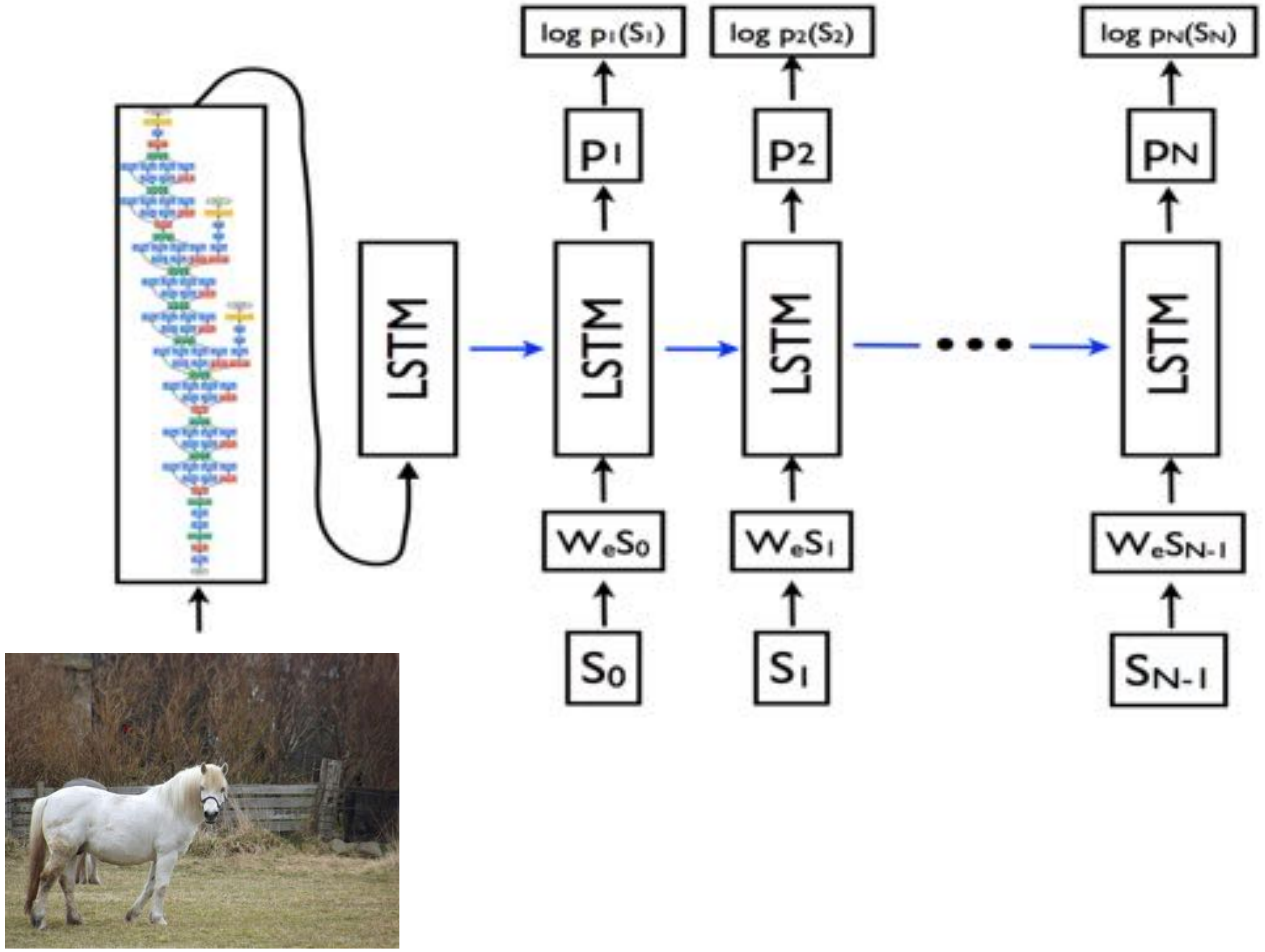
ON THE MENU

- A quick overview of Computer Vision (CV) tasks and challenges
- Natural Language Processing (NLP) tasks and challenges
- Challenges in combining both
- Representations learning in CV
- Representation learning in NLP
- Combining both

IMAGE CAPTIONING

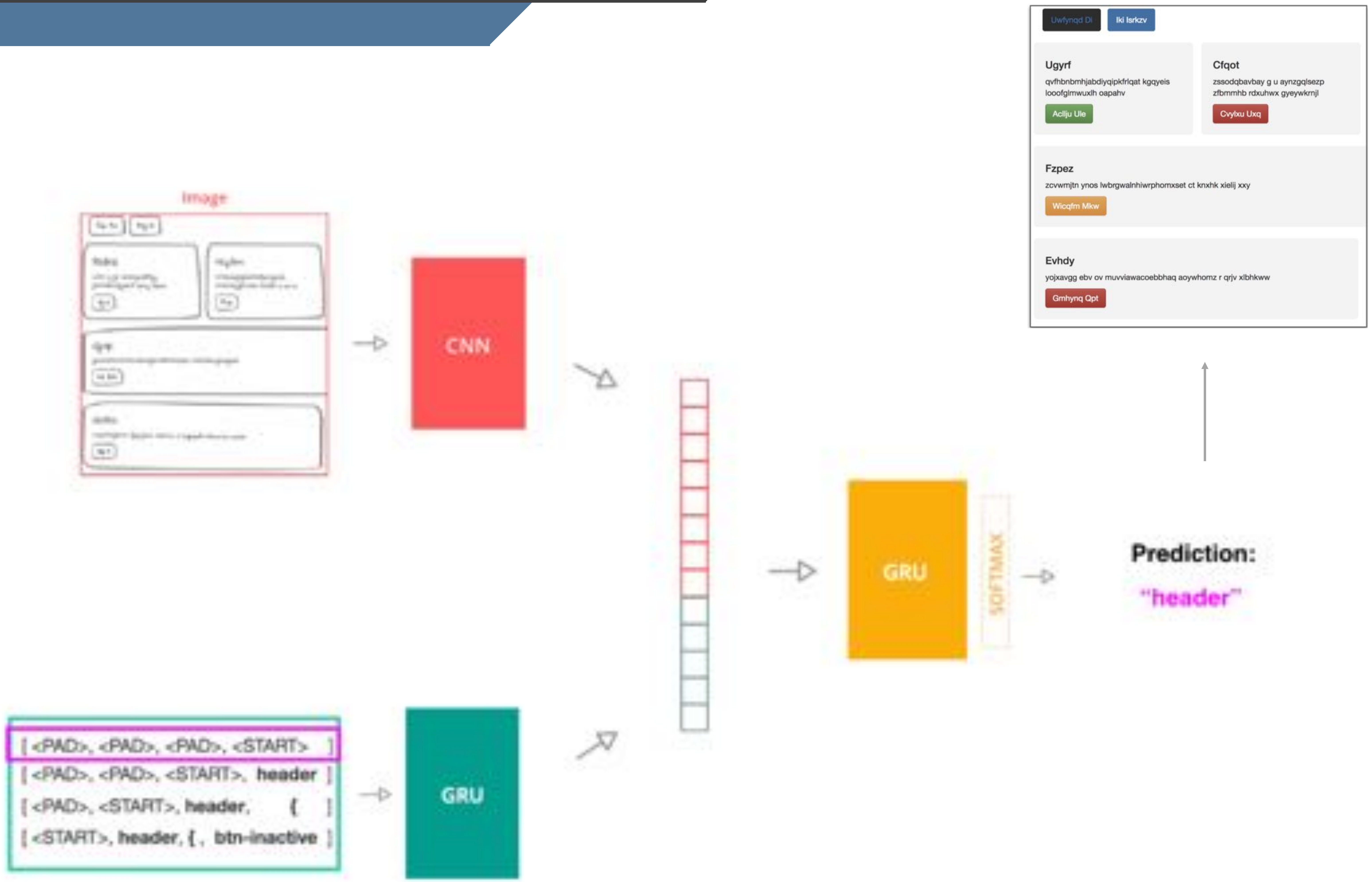
- Prime language model with features extracted from CNN
- Feed to an NLP language model
- End-to-end
  - Elegant
  - Hard to debug and validate
  - Hard to productionize

A horse is standing in a field with a fence in the background.



CODE GENERATION

- Harder problem for humans
  - Anyone can describe an image
  - Coding takes specific training
- We can solve it using a similar model
- The trick is in getting the data!



BUT DOES IT SCALE?

- These methods mix and match different architectures
- The combined representation is often learned implicitly
  - Hard to cache and optimize to re-use across services
  - Hard to validate and do QA on
- The models are entangled
  - What if we want to learn a simple joint representation?

Image Search

Goals

- Searching for **similar images to an input image**
  - Computer Vision: (Image → Image)
- Searching for **images using text &** generating **tags for images**
  - Computer Vision + Natural Language Processing: (Image ↔ Text)
- Bonus: finding **similar words to an input word**
  - Natural Language Processing: (Text → Text)

ON THE MENU

- A quick overview of Computer Vision (CV) tasks and challenges
- Natural Language Processing (NLP) tasks and challenges
- Challenges in combining both
- Representations learning in CV
- Representation learning in NLP
- Combining both

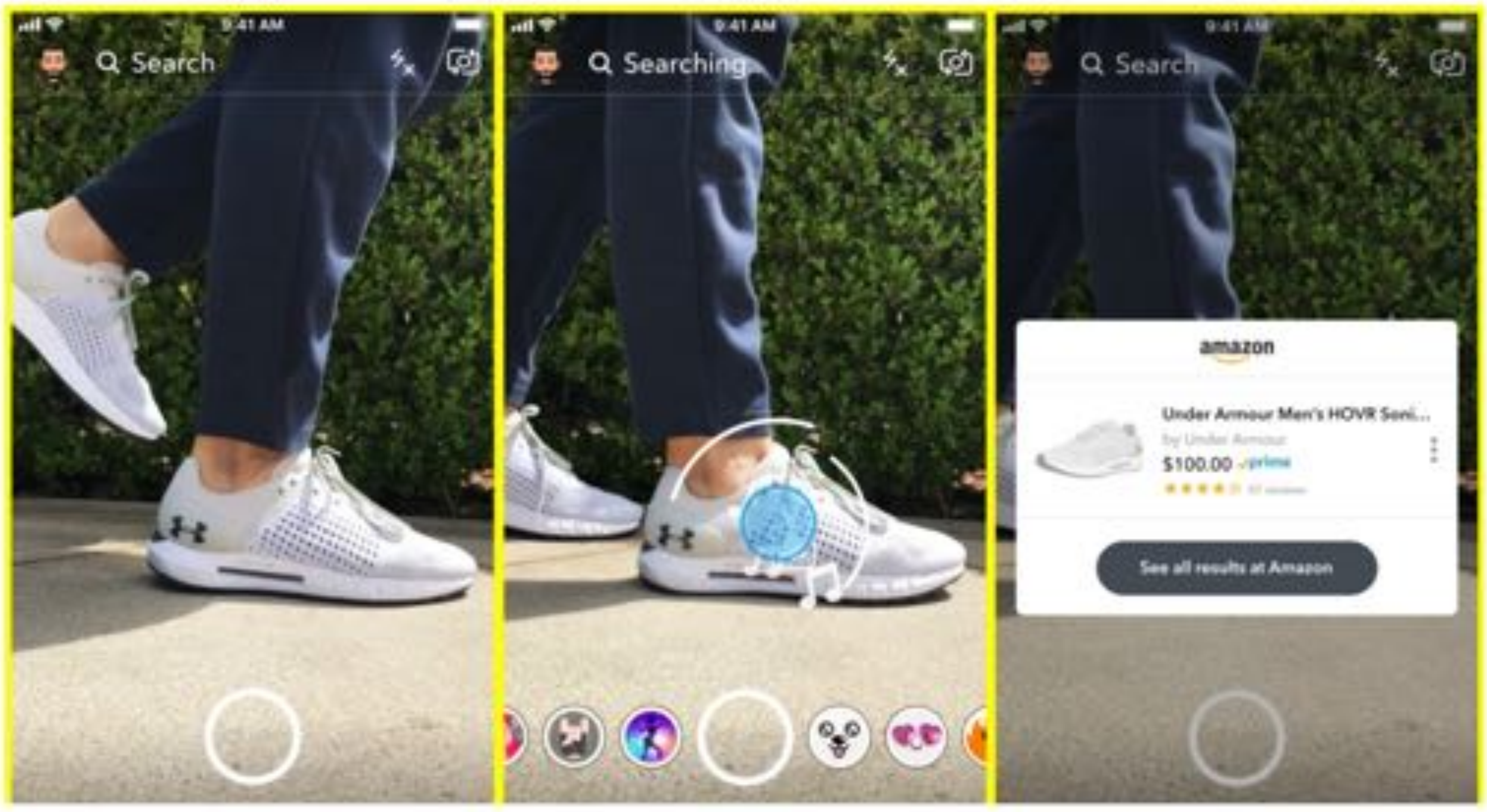
Image Based Search

Let’s build this!

Snapchat lets you take a photo of an object to buy it on Amazon

Josh Constine @joshconstine / 4 weeks ago

Comment



Dataset

- 1000 images
  - 20 classes, 50 images per class
- 3 orders of magnitude smaller than usual deep learning datasets
- Noisy

aeroplane bicycle bird boat bottle bus car cat chair cow
dining_table dog horse motorbike person potted_plant sheep sofa
train tv_monitor

Credit to Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier for the dataset.

WHICH CLASS?

`aeroplane` `bicycle` `bird` `boat` `bottle` `bus` `car` `cat` `chair` `cow` `dining_table` `dog` `horse` `motorbike` `person` `potted_plant` `sheep` `sofa` `train` `tv_monitor`

DATA PROBLEMS

Bottle ☹

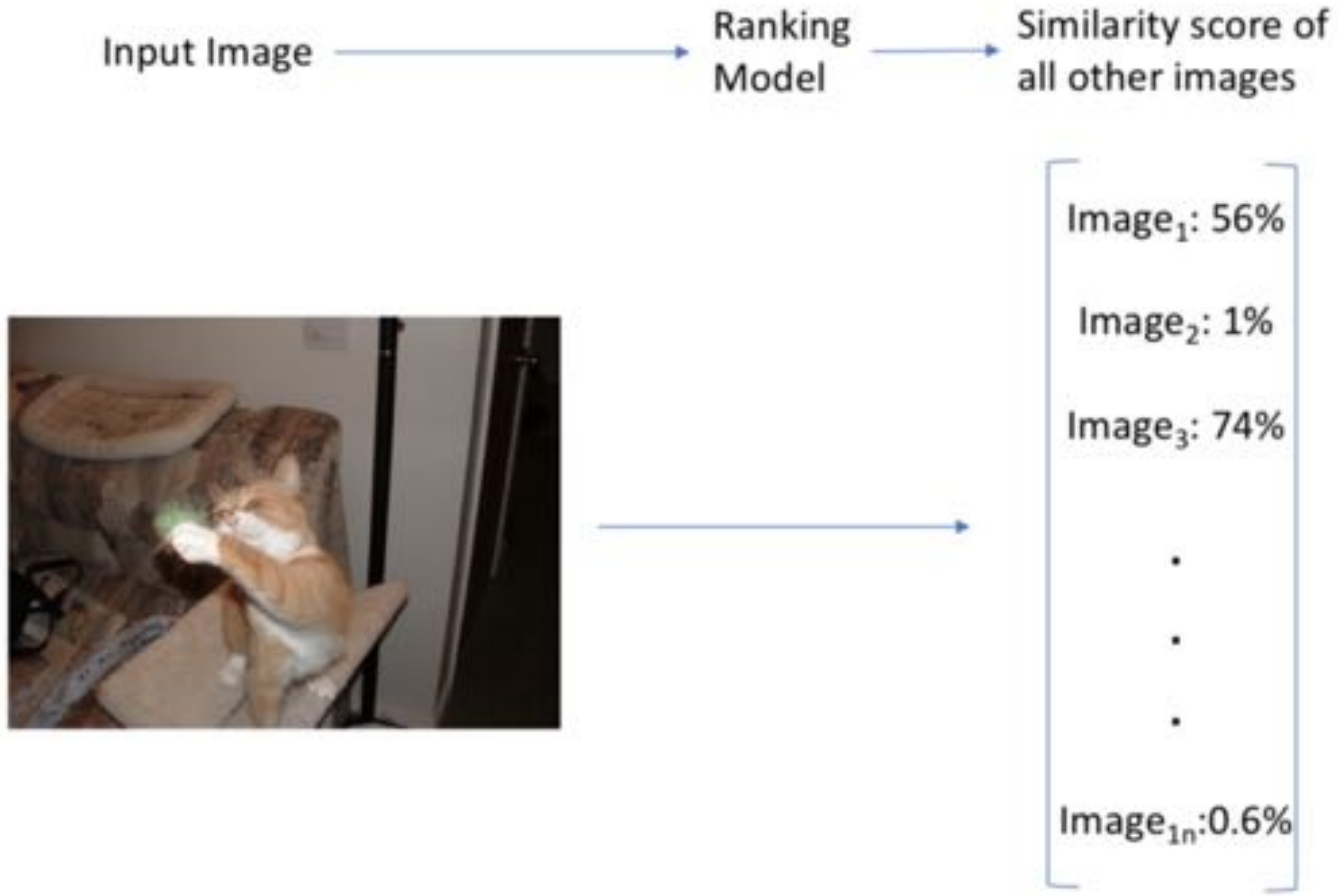
INSIGHT

A FEW APPROACHES

- Ways to think about searching for similar images

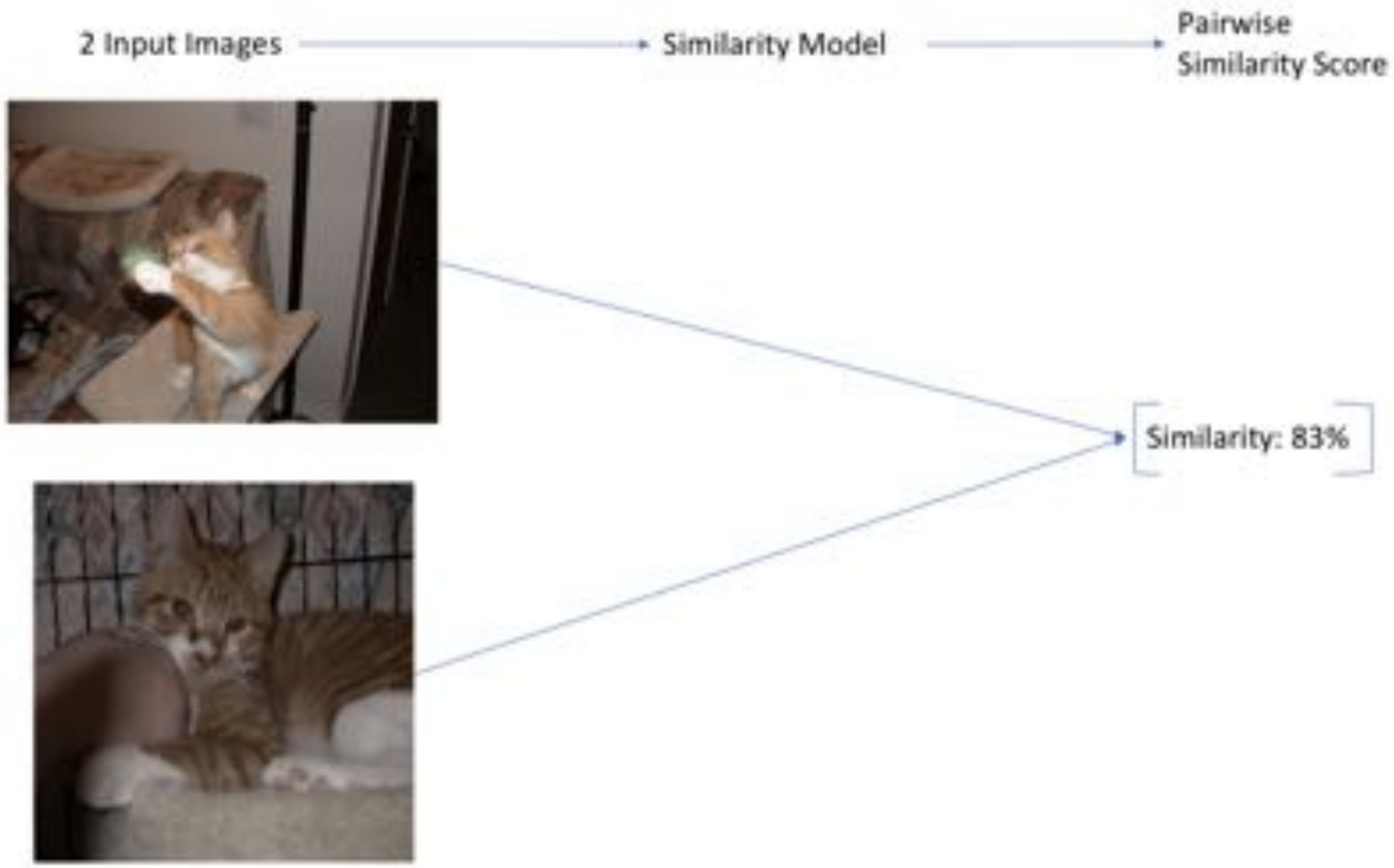
IF WE HAD INFINITE DATA

- Train on all images
- Pros:
  - One Forward Pass (fast inference)
- Cons:
  - Hard too optimize
  - Poor scaling
  - Frequent Retraining



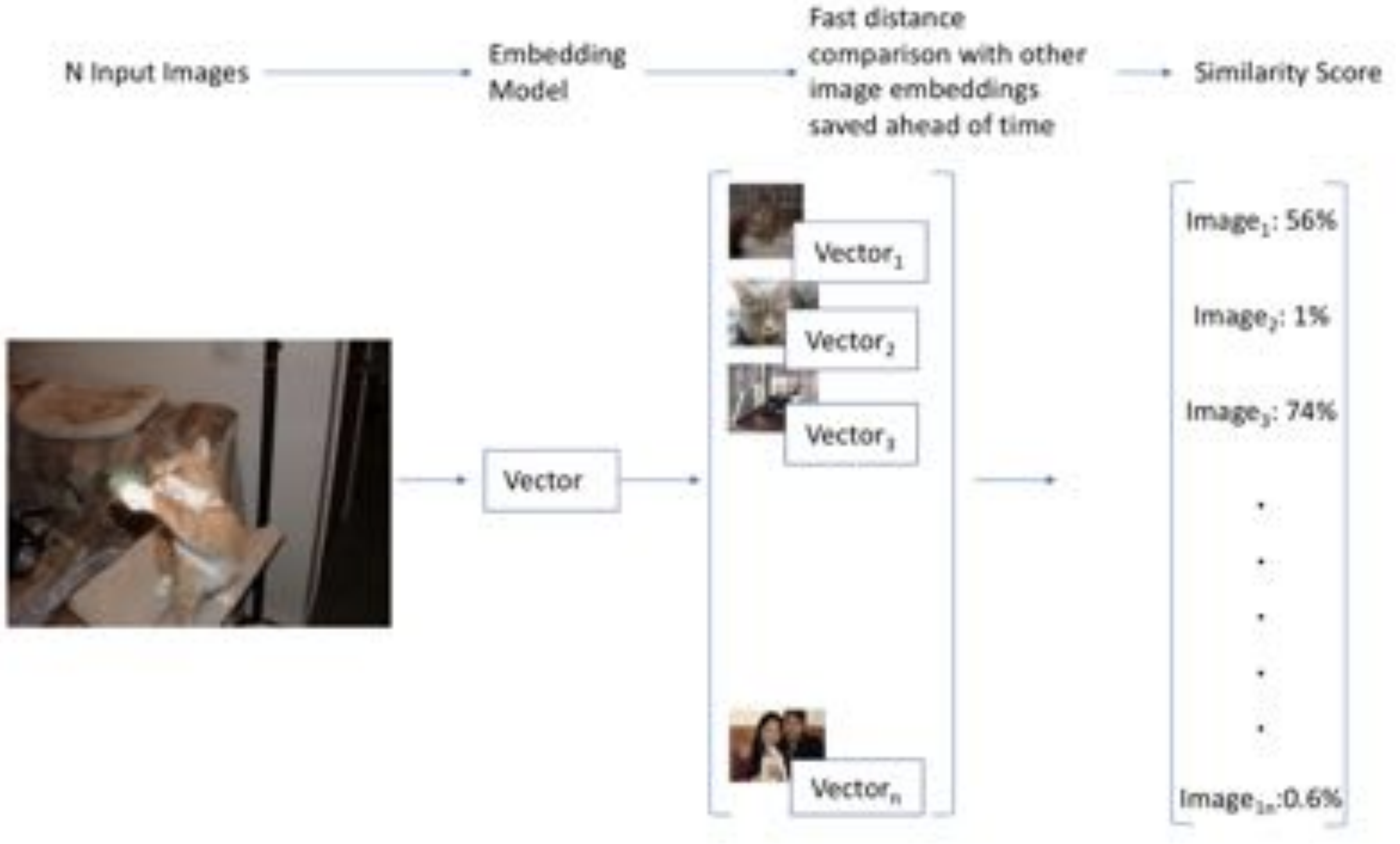
SIMILARITY MODEL

- Train on each image pair
- Pros:
  - Scales to large datasets
- Cons:
  - Slow
  - Does not work for text
  - Needs good examples

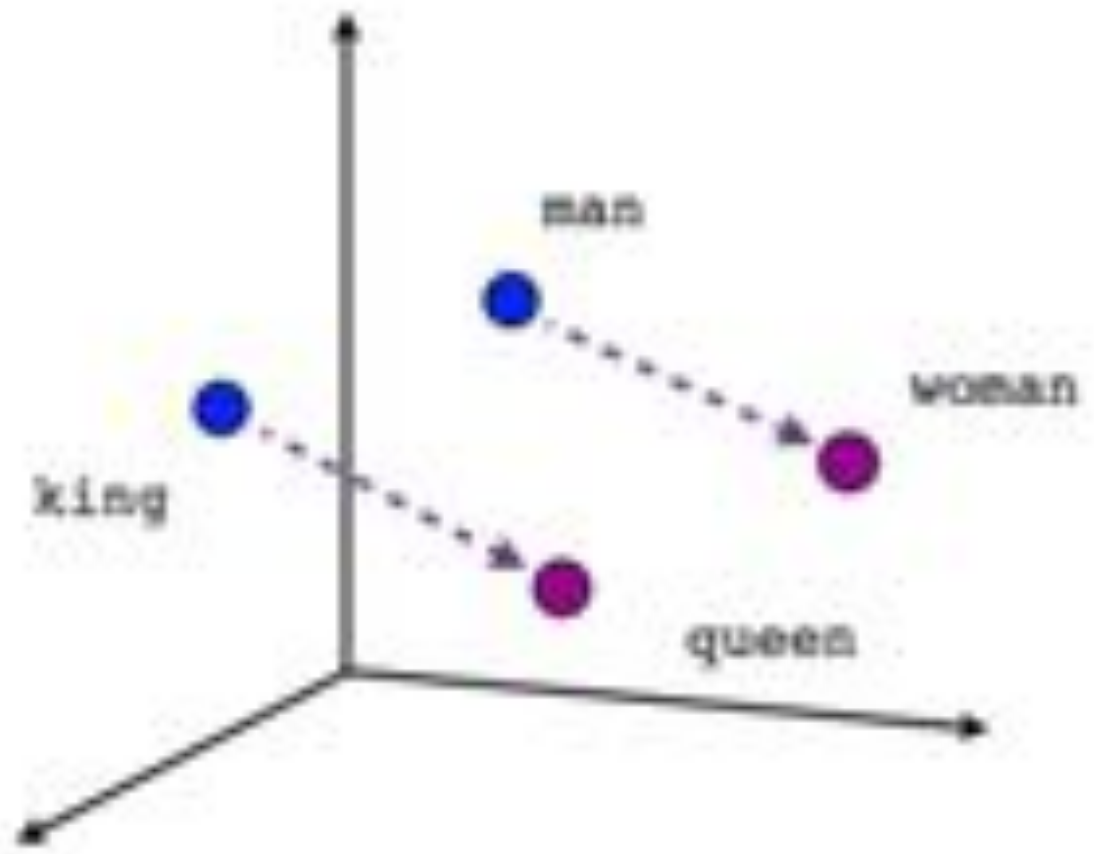


EMBEDDING MODEL

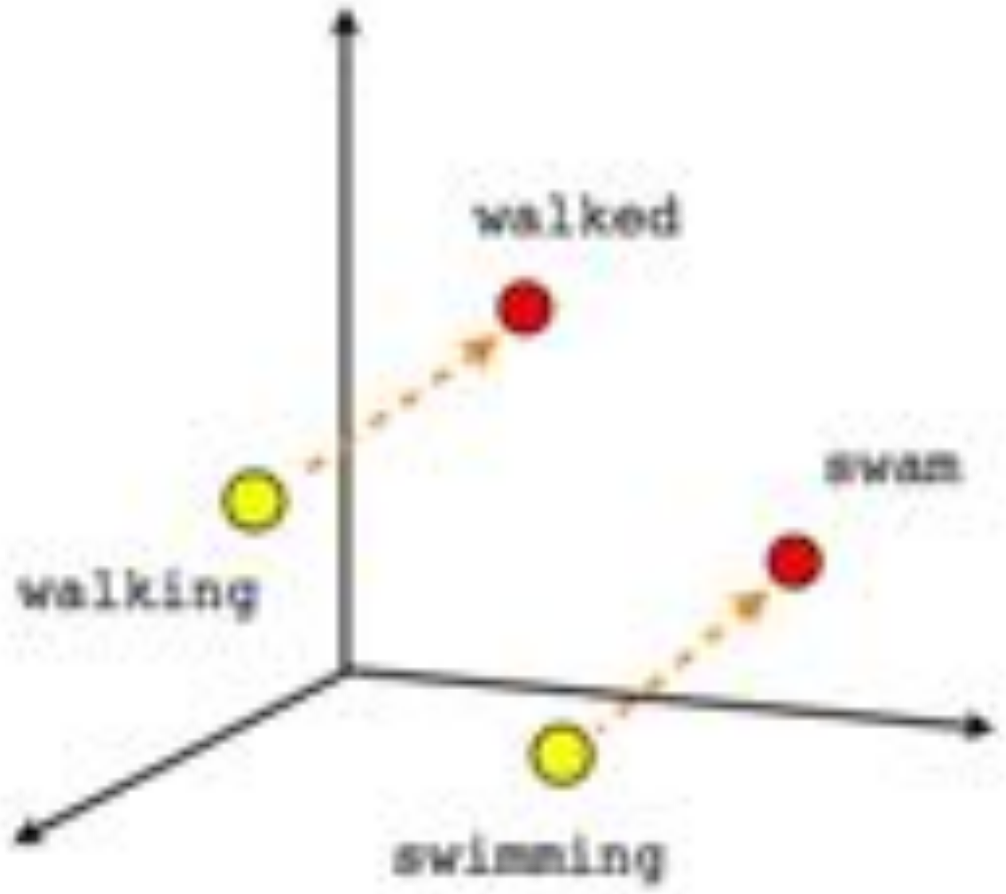
- Find embedding for each image
- Calculate ahead of time
- Pros:
  - Scalable
  - Fast
- Cons:
  - Simple representations



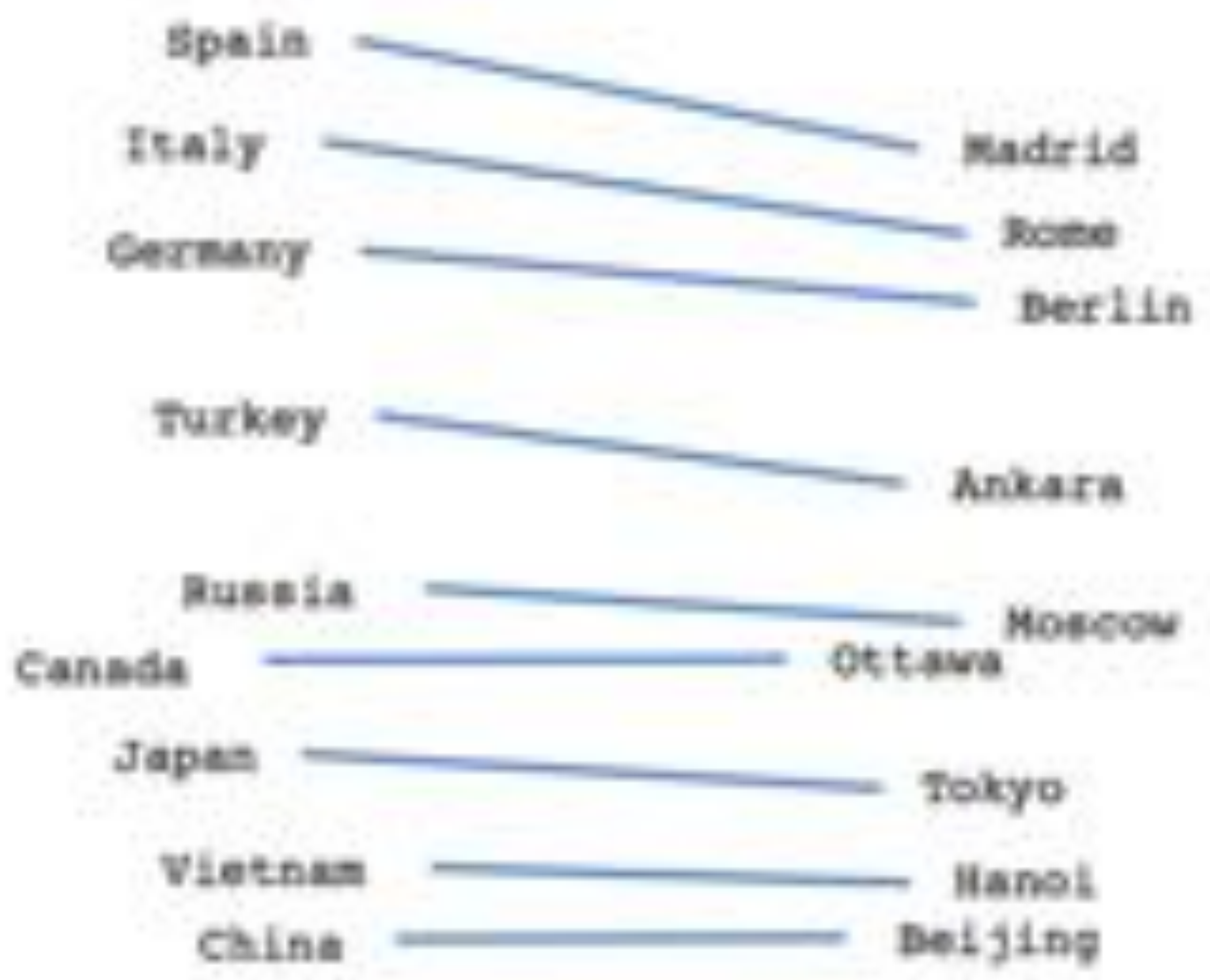
WORD EMBEDDINGS



Male-Female



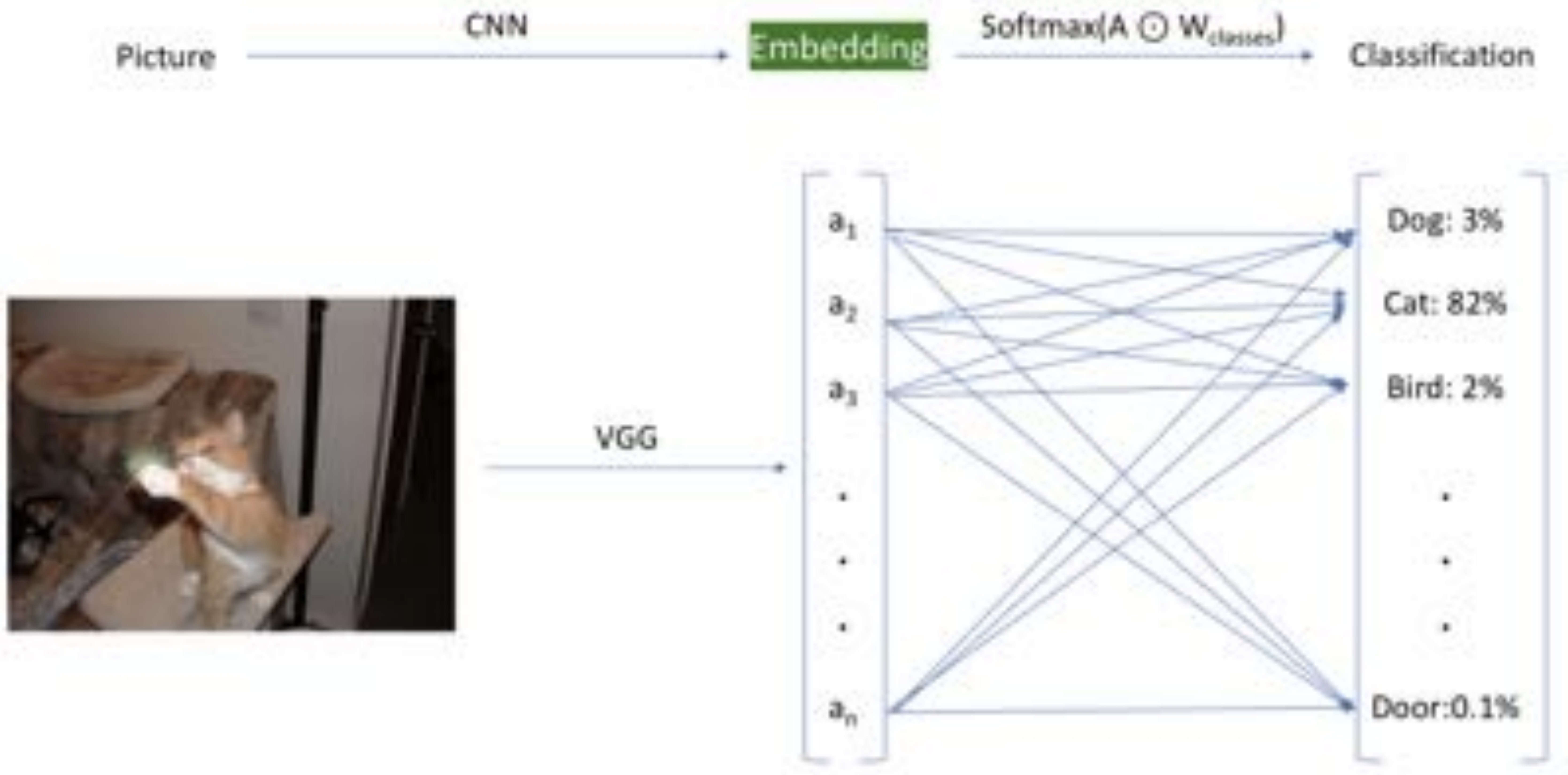
Verb tense



Country-Capital

Mikolov et Al. 2013

LEVERAGING A PRE-TRAINED MODEL



HOW AN EMBEDDING LOOKS

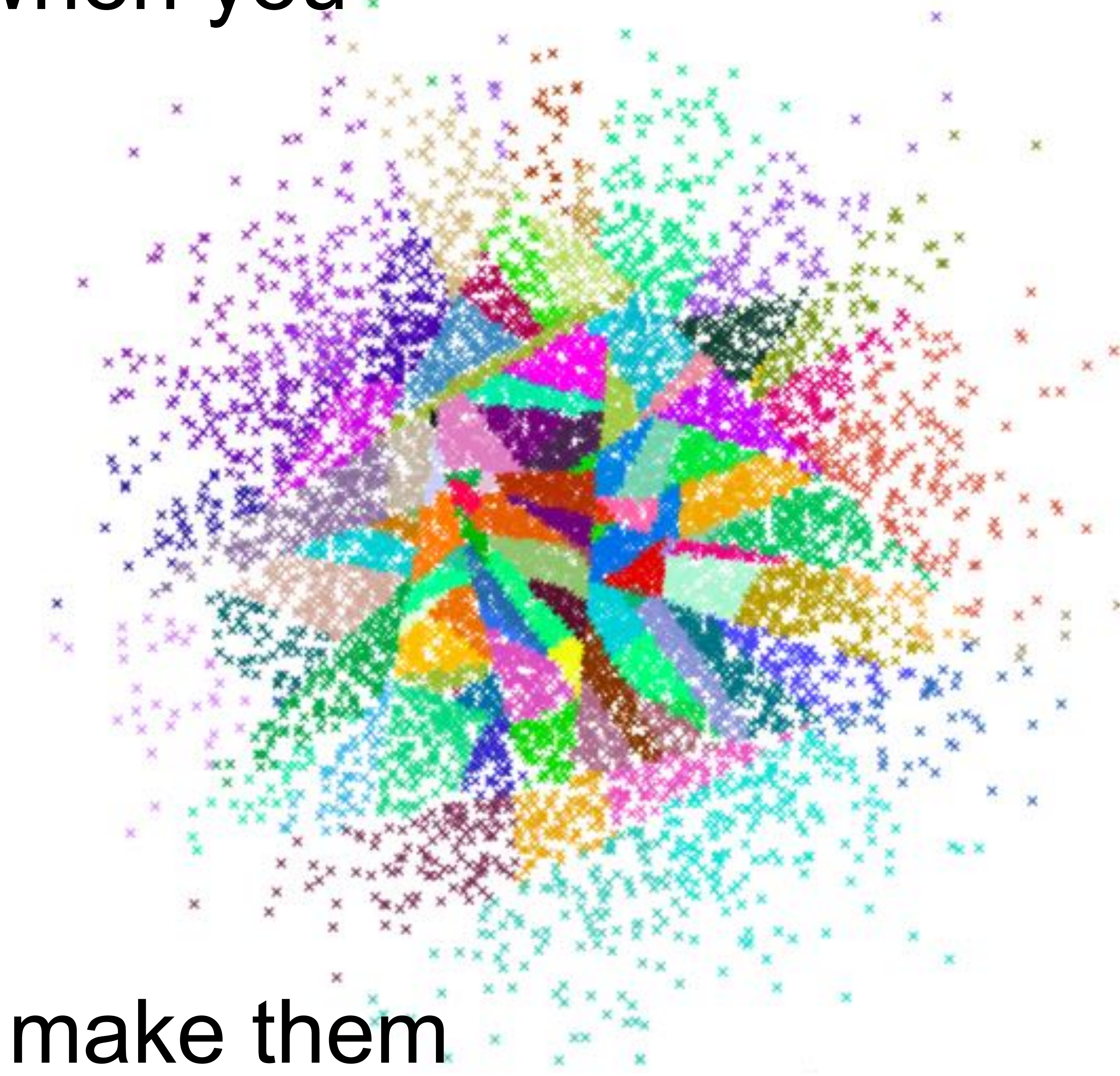
	4	4885	4886	4887	4888	4889	4890	4891	4892	4893	4894	4895
0	0	0.4546	0	6.1516	0	1.3763	0	0	1.1756	0.4681	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	1.8355	0	0	0	0	0.6923	0	0	0
3	5	0	0	0	0.4420	5.2478	0	0	1.6778	0	0	0
4	6	0.1924	0	0	0	0	0	0	0.8986	0	0	0
5	9	0	0	2.0729	0	0	0	0	0.0996	0.2377	0	0
6	0	0	0	6.8556	0	1.3900	0	0	0.6859	1.1272	0	0
7	0	0	0	2.2498	0	0	0	1.8188	0.1840	0	0	0
8	0	0	0	6.0193	0	0	0	0	0	0	0	0
9	5	0	0	0	0.9562	0	0.3197	0	1.8738	3.5308	0	1.3911
10	6	0	0	0.2099	0	0	0	1.4480	1.3150	1.1056	2.1684	0

INSIGHT

PROXIMITY SEARCH IS FAST

How do you find the 5 most similar images to a given one when you have over a million users?

- Fast index search
  - Spotify uses annoy (we will as well)
  - Flickr uses LOPQ
  - Nmslib is also very fast
- Some rely on making the queries approximate in order to make them fast

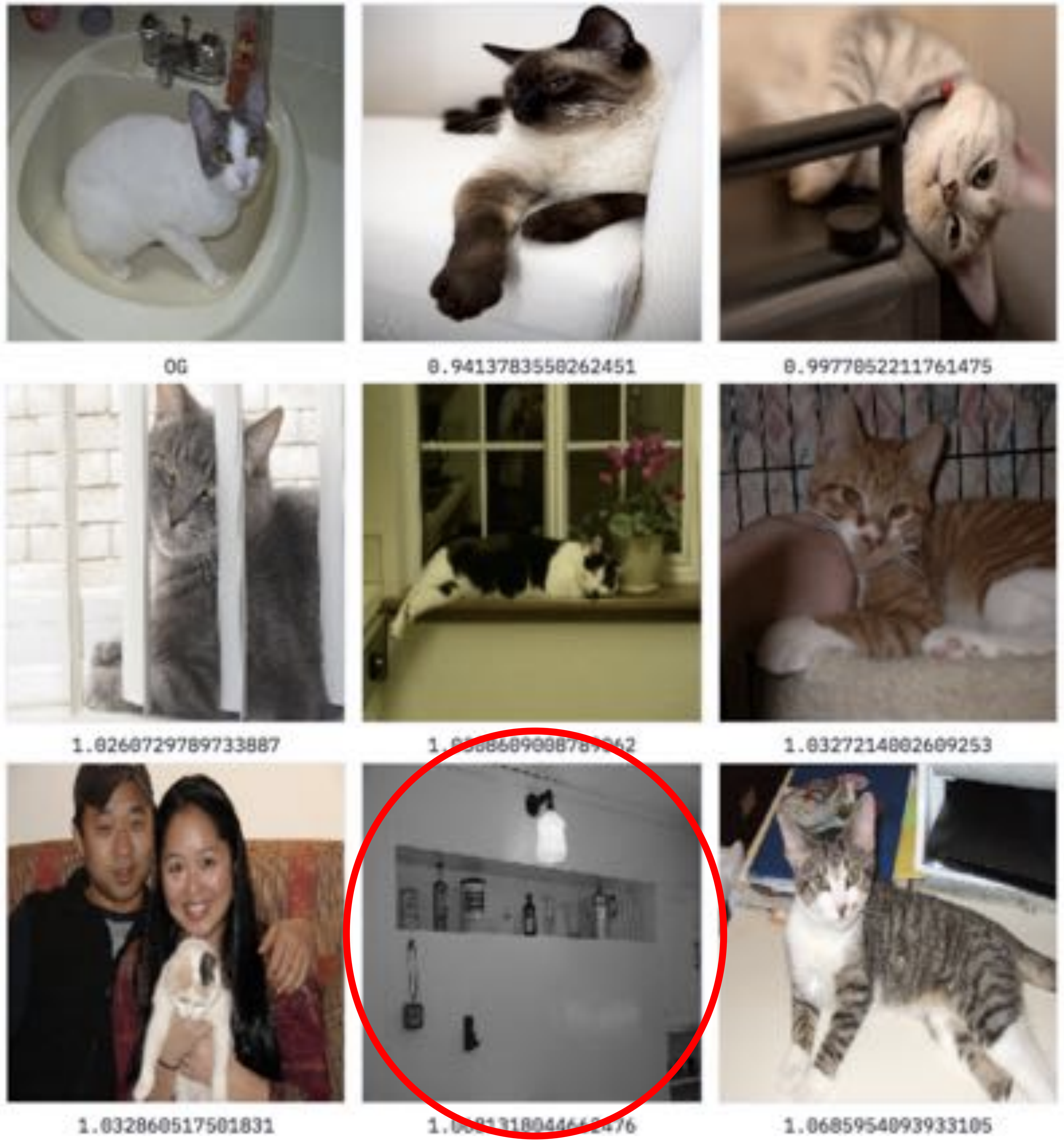


INSIGHT

PRETTY IMPRESSIVE!

IN

OUT



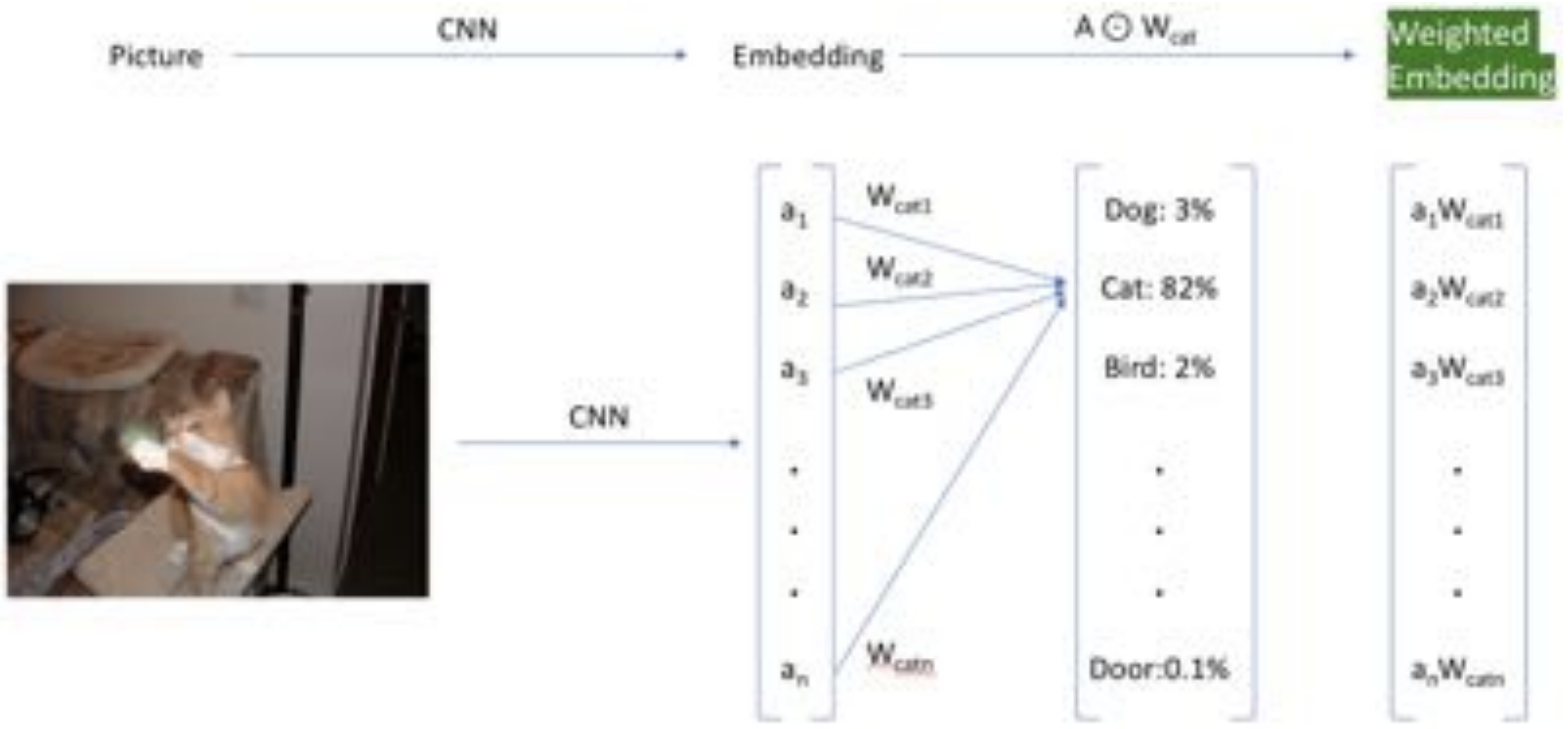
FOCUSING OUR SEARCH

- Sometimes we are only interested in **part of the image**.
- For example, given an image of a cat and a bottle, we might be only interested in similar cats, not similar bottles.
- How do we incorporate this information

IMPROVING RESULTS: STILL NO TRAINING

- **Computationally expensive approach:**
  - Object detection model first
    - *(We don't do this)*
  - Image search on a cropped image
    - *(We don't do this)*

- **Semi-Supervised approach:**
  - Hacky, but efficient!
  - re-weighing the activations
  - Only use the class of interest to re-weigh embeddings



EVEN BETTER

IN

OUT

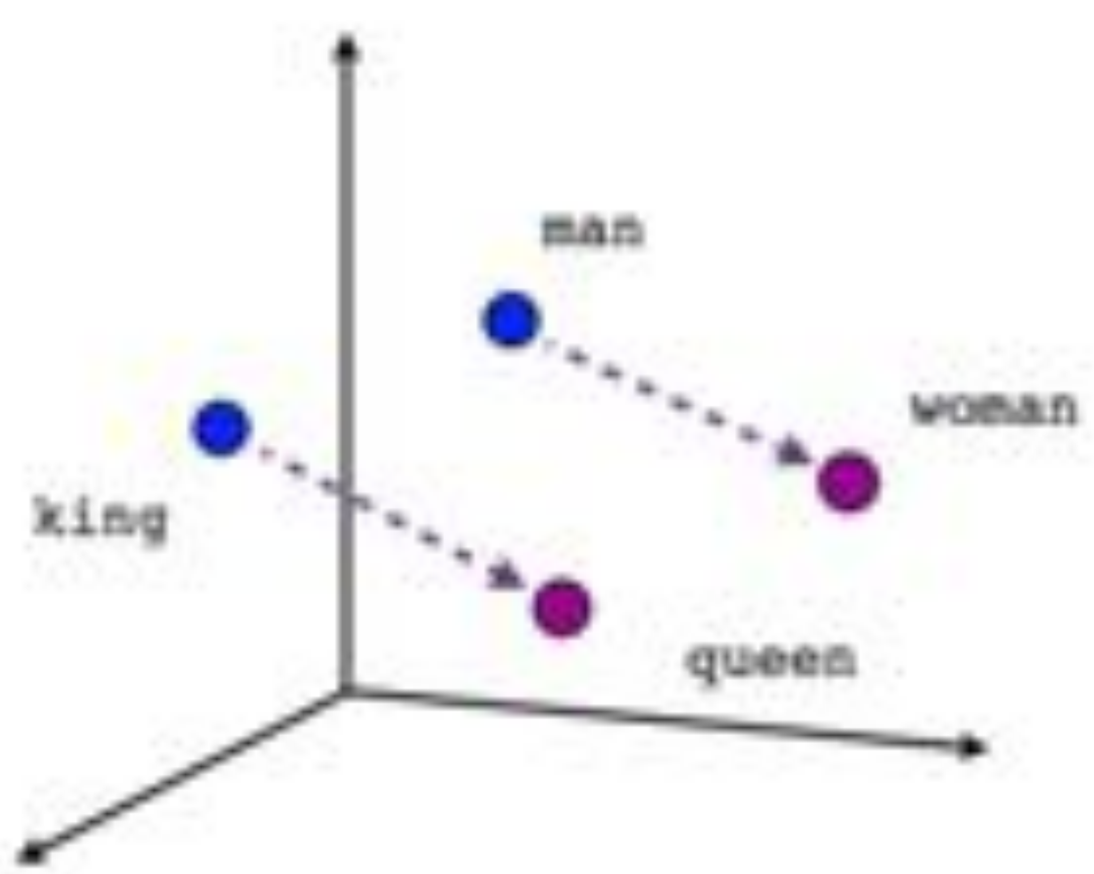
ON THE MENU

- A quick overview of Computer Vision (CV) tasks and challenges
- Natural Language Processing (NLP) tasks and challenges
- Challenges in combining both
- Representations learning in CV
- Representation learning in NLP
- Combining both

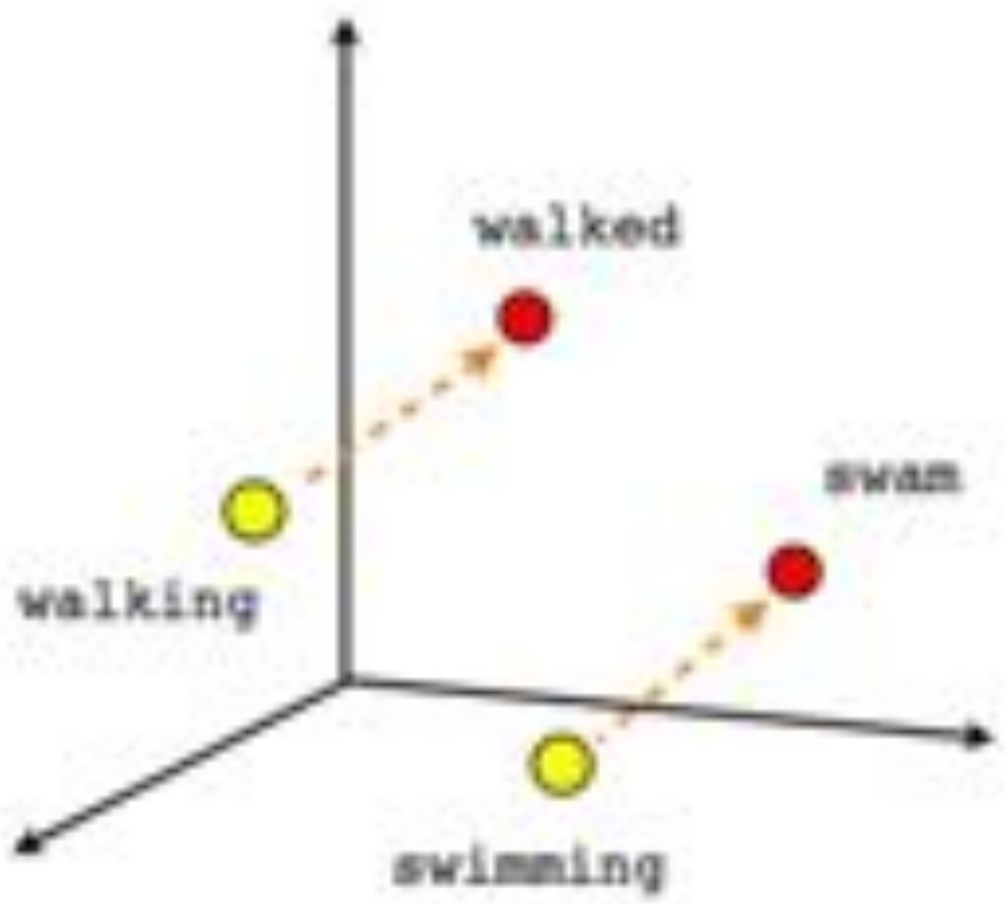
GENERALIZING

- We have added some ability to guide the search, but it is limited to classes our model was initially trained on
- We would like to be able to **use any word**
- How do we combine words and images?

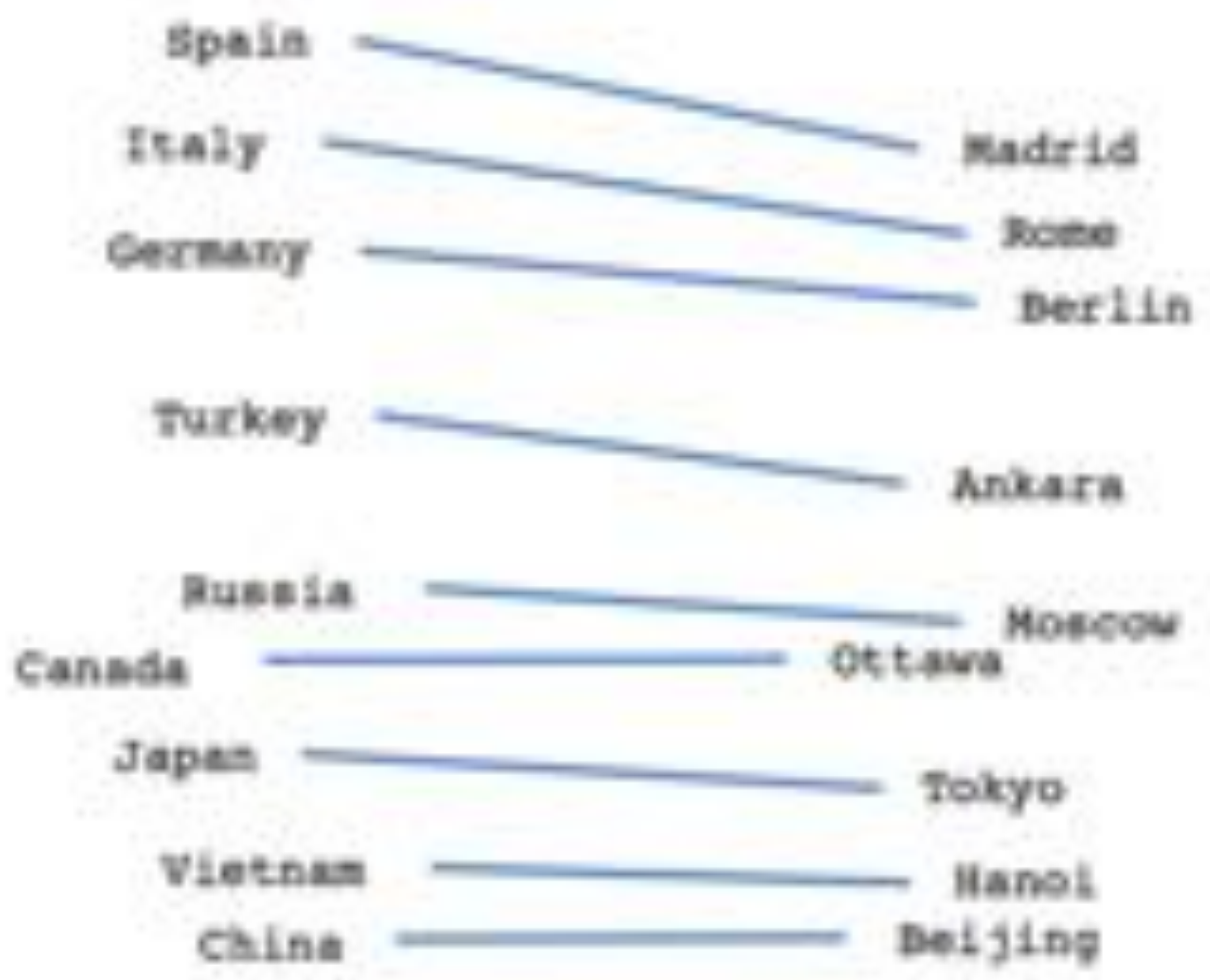
WORD EMBEDDINGS



Male-Female



Verb tense



Country-Capital

Mikolov et Al. 2013

SEMANTIC TEXT!

- **Load a set of pre-trained vectors (GloVe)**
  - Wikipedia data
  - Semantic relationships
- **One big issue:**
  - The embeddings for images are of size 4096
  - While those for words are of size 300
  - And both models trained in a different fashion
- **What we need: *Joint model!***

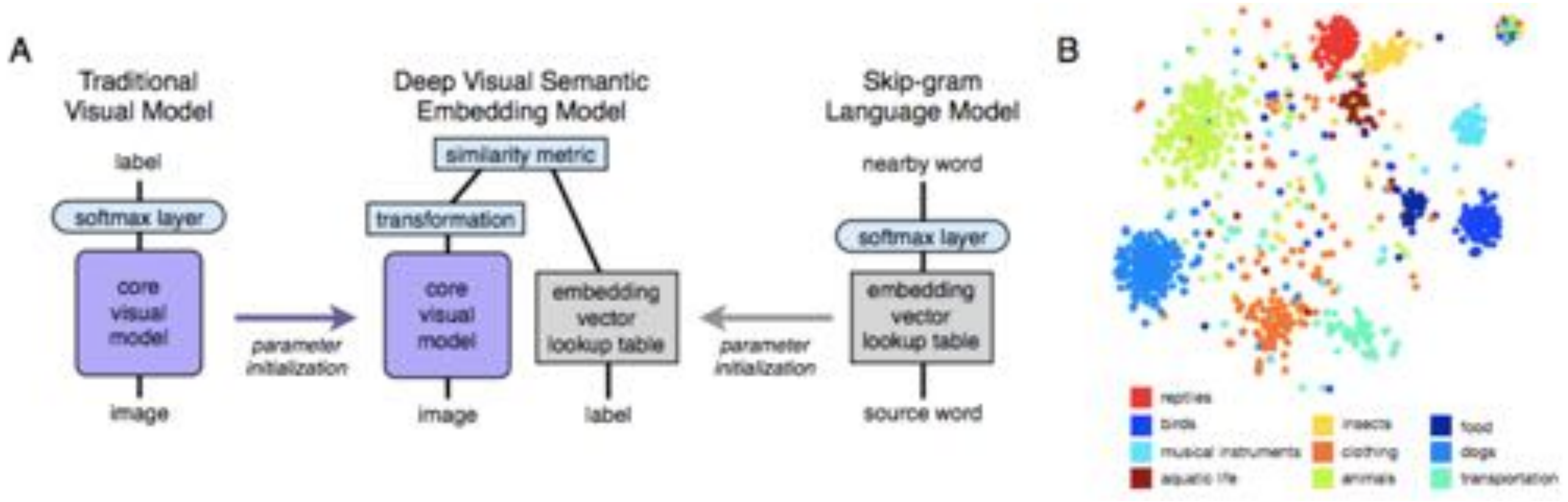
Searching for `said`, for example, returns this list of [`word`, `distance`]:

- `['said', 0.0]`
- `['told', 0.688713550567627]`
- `['spokesman', 0.7859575152397156]`
- `['asked', 0.872875452041626]`
- `['noting', 0.9151610732078552]`
- `['warned', 0.915908694267273]`
- `['referring', 0.9276227951049805]`
- `['reporters', 0.9325974583625793]`
- `['stressed', 0.9445104002952576]`
- `['tuesday', 0.9446316957473755]`

ON THE MENU

- A quick overview of Computer Vision (CV) tasks and challenges
- Natural Language Processing (NLP) tasks and challenges
- Challenges in combining both
- Representations learning in CV
- Representation learning in NLP
- Combining both

Inspiration

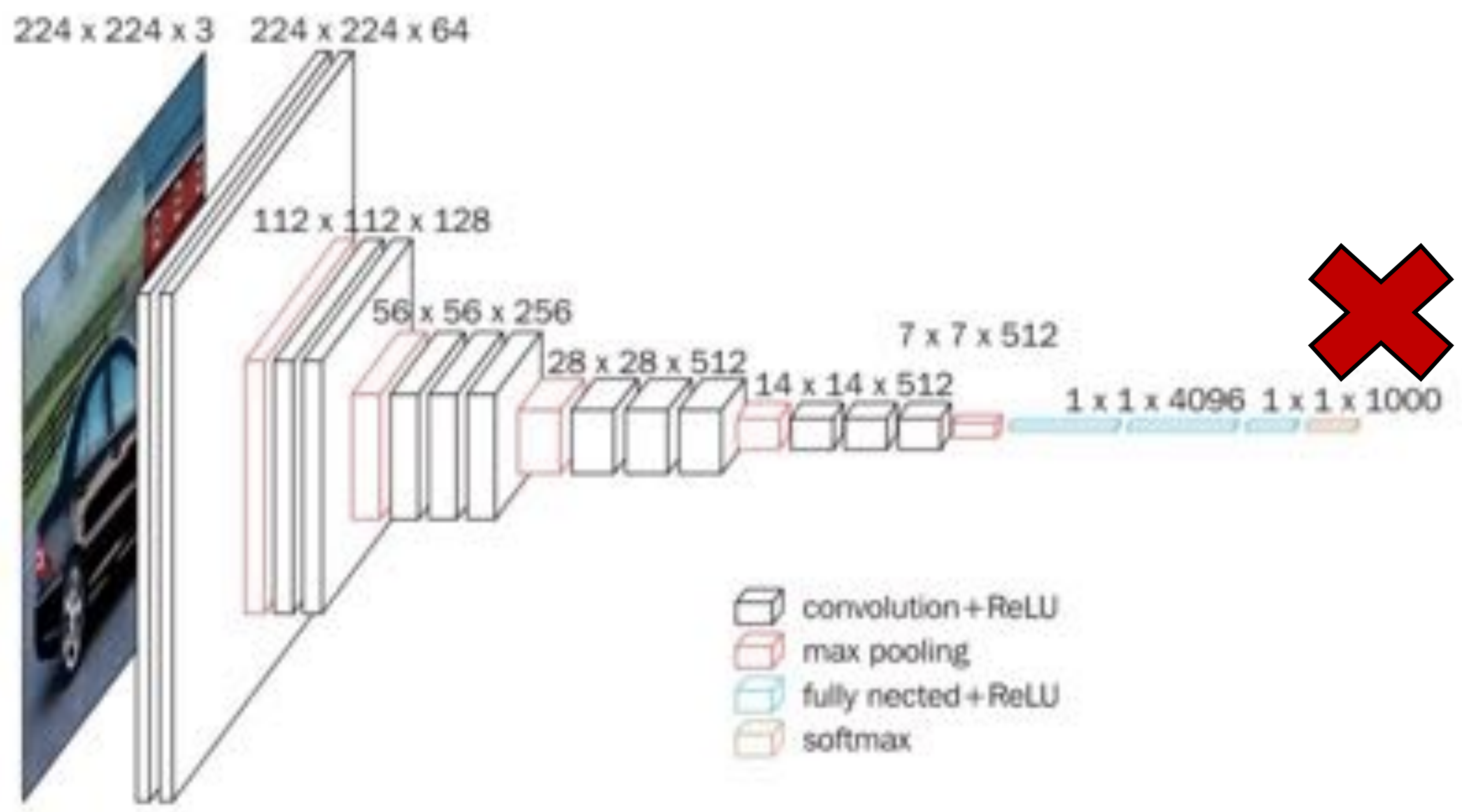


DeViSE: A Deep Visual-Semantic Embedding Model

Andrea Frome*, Greg S. Corrado*, Jonathon Shlens*, Samy Bengio
Jeffrey Dean, Marc'Aurelio Ranzato, Tomas Mikolov
* These authors contributed equally.

{afrome, gcorrado, shlens, bengio, jeff, ranzato, tmikolov}@google.com
Google, Inc.
Mountain View, CA, USA

TIME TO TRAIN



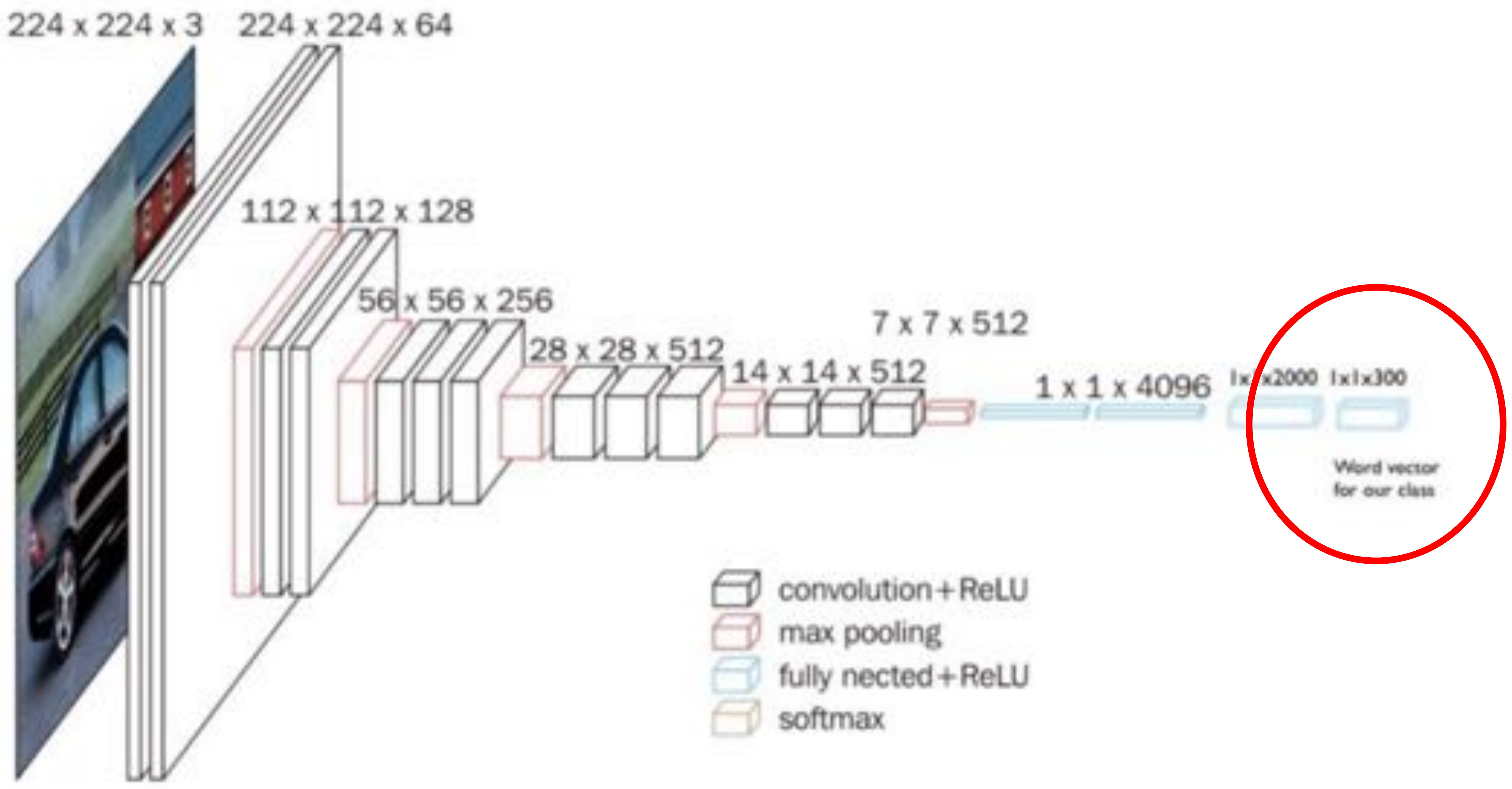


Image → Image

Image → Text

IMAGE → TEXT

- **Re-train model to predict the word vector**
  - i.e. 300-length vector associated with cat
- **Training**
  - Takes more time per example than image → class
  - But *much* faster than on Imagenet (7 hours, no GPU)
- **Important to note**
  - Training data can be *very small:* ~1000 images
  - Miniscule compared to Imagenet (1+ Million images)
- **Once model is trained**
  - Build a new fast index of images
  - Save to disk

How do you think this model will perform?

INSIGHT

IMAGE → TEXT

Here are the generated tags:

- `[6676, 'bottle', 0.3879561722278595]`
- `[7494, 'bottles', 0.7513495683670044]`
- `[12780, 'cans', 0.9817070364952087]`
- `[16883, 'vodka', 0.9828150272369385]`
- `[16720, 'jar', 1.0084964036941528]`
- `[12714, 'soda', 1.0182772874832153]`
- `[23279, 'jars', 1.0454961061477661]`
- `[3754, 'plastic', 1.0530102252960205]`
- `[19045, 'whiskey', 1.061428427696228]`
- `[4769, 'bag', 1.0815287828445435]`

GENERALIZED IMAGE SEARCH WITH MINIMAL DATA

IN: "DOG"

OUT

SEARCH FOR WORD NOT IN DATASET

IN: "OCEAN"

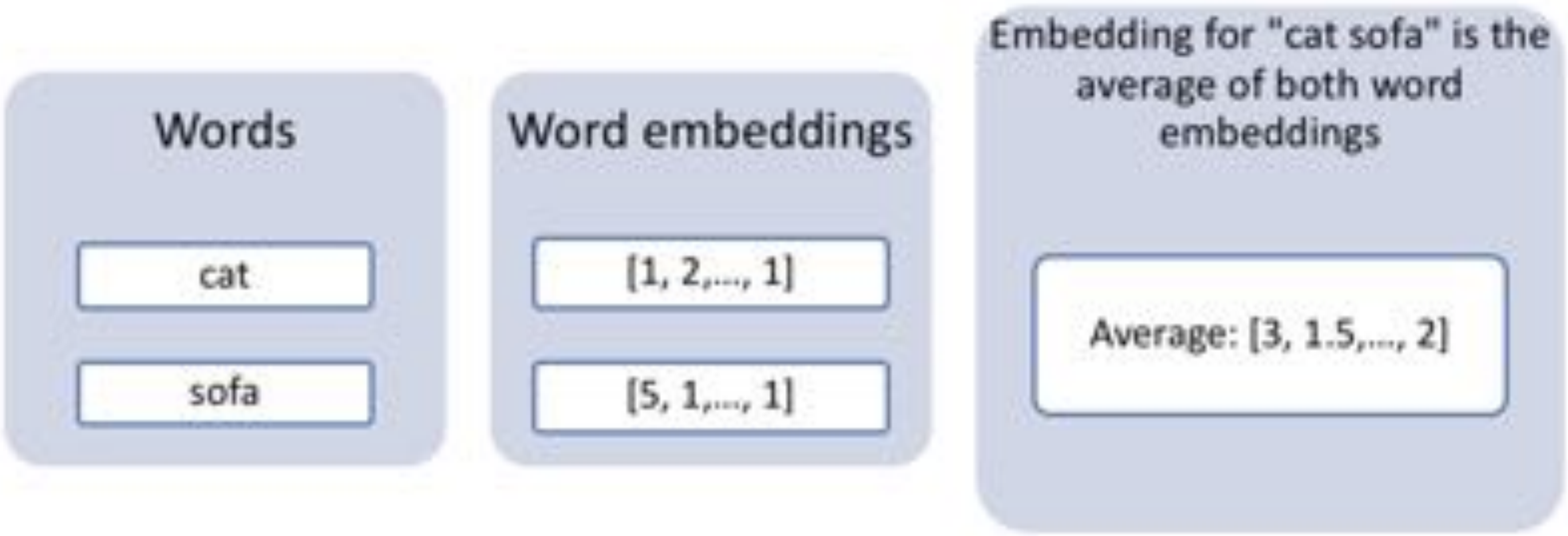
OUT

SEARCH FOR WORD NOT IN DATASET

IN: “STREET”

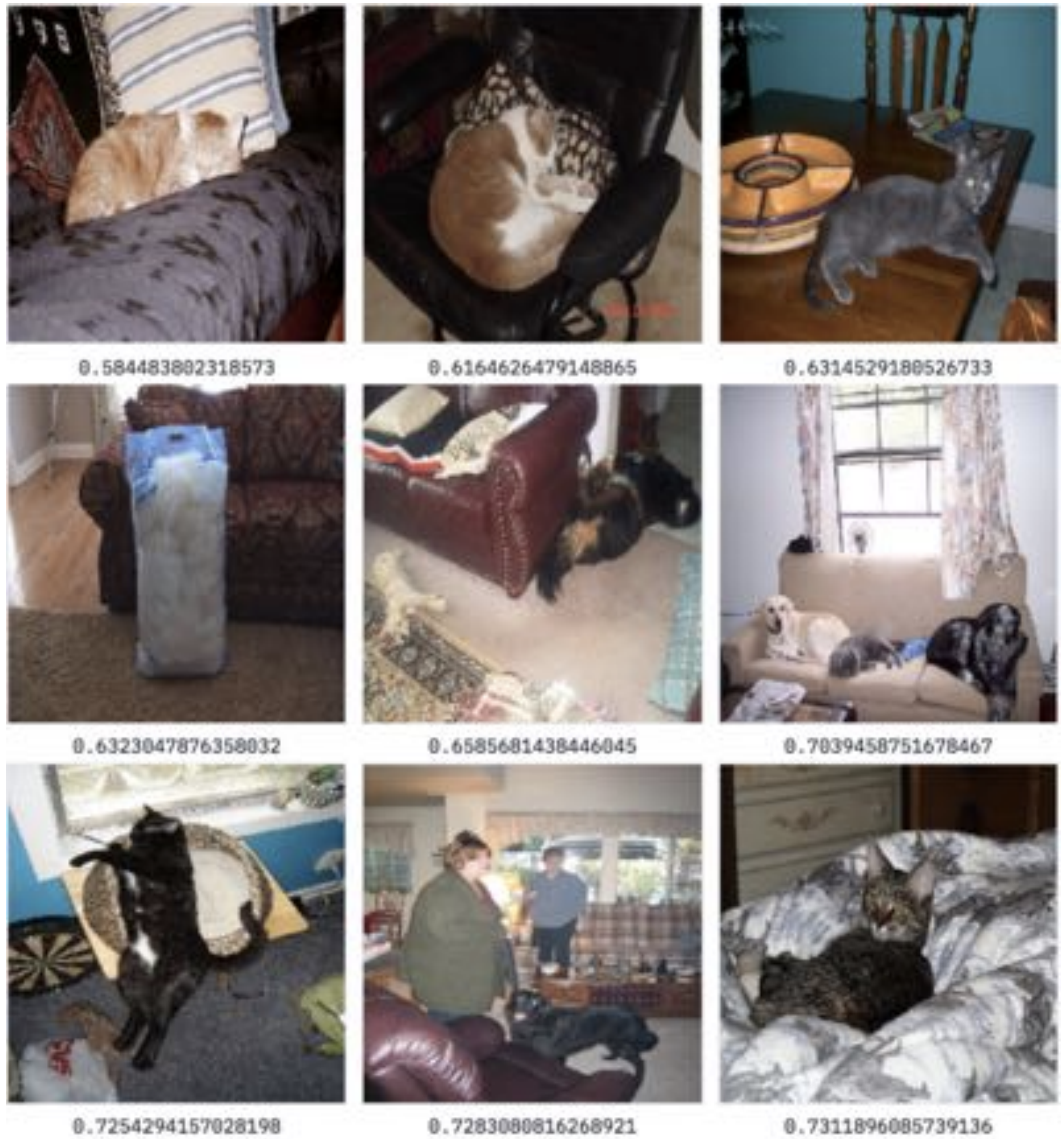
OUT

MULTIPLE WORDS!



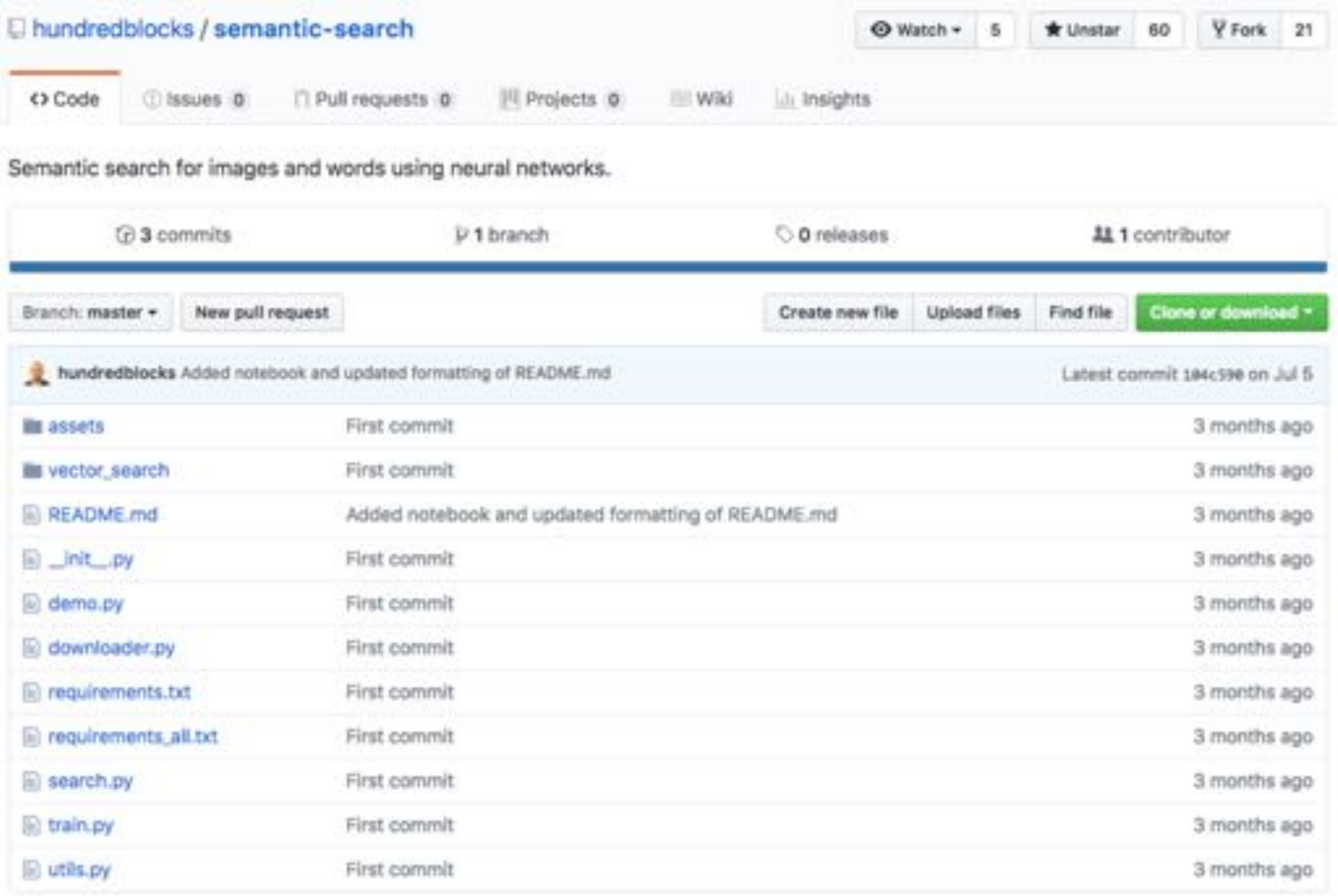
MULTIPLE WORDS!

IN: "CAT SOFA"   OUT



INSIGHT

Learn More:
Find the repo on Github!



Next steps

- **Incorporating user feedback**
  - Most real world image search systems use user clicks as a signal
- **Capturing domain specific aspects**
  - Often times, users have different meanings for similarity
- **Keep the conversation going**
  - Reach me on Twitter @EmmanuelAmeisen

INSIGHT

EMMANUEL AMEISEN

Head of AI, ML Engineer

emmanuel@insightdata.ai

@emmanuelameisen

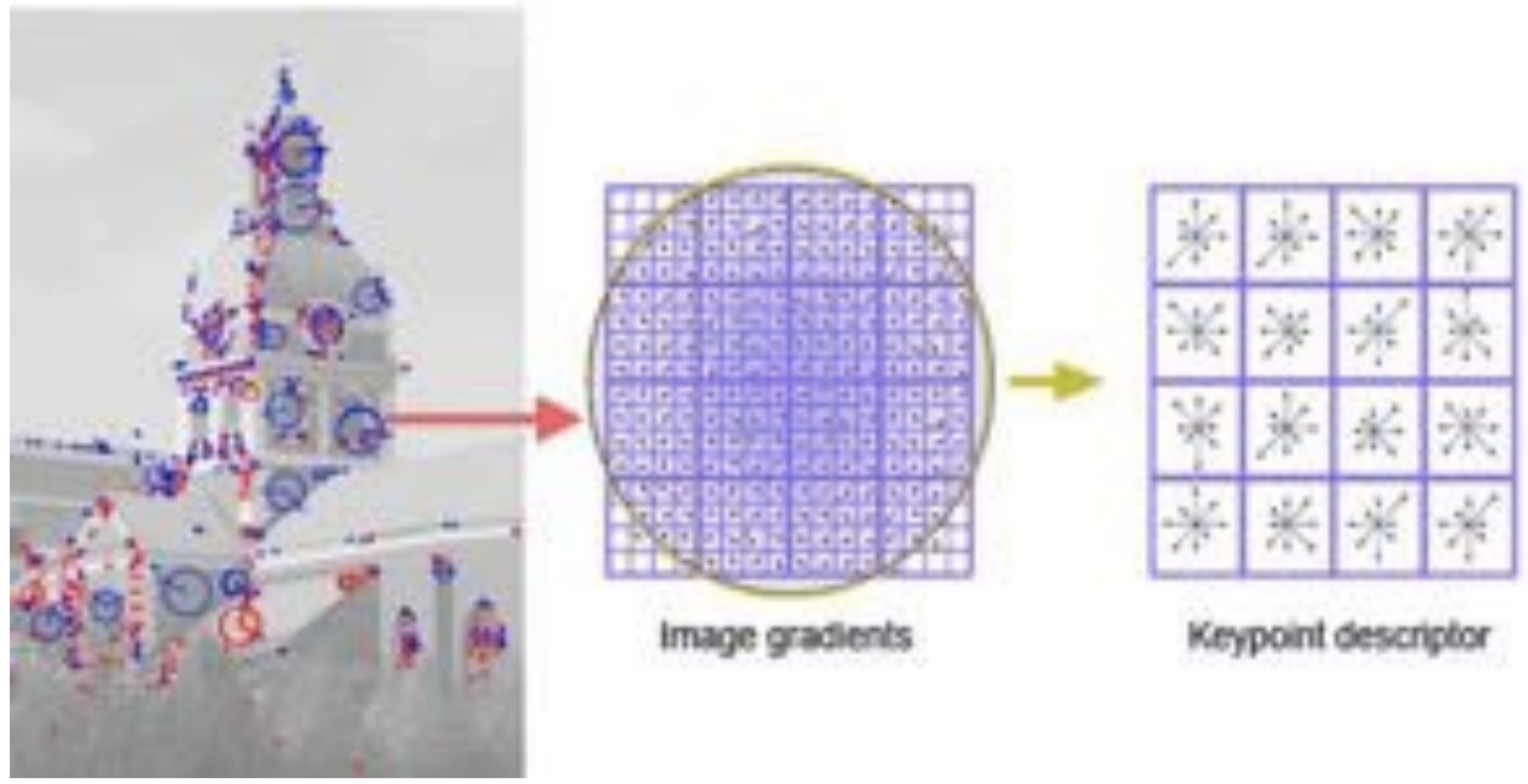
bit.ly/imagefromscratch

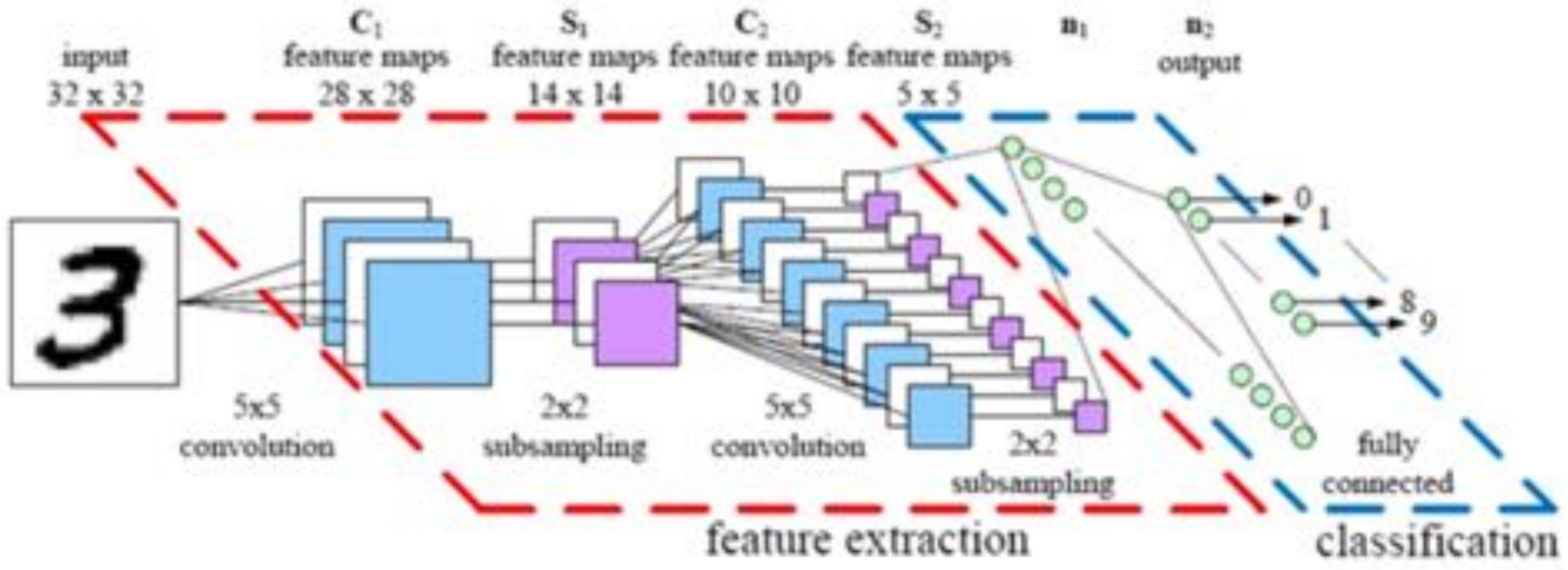
www.insightdata.ai/apply

CV Approaches

White-box Algorithms

Black-Box Algorithms





@Andrey Nikishaev

CLASSIFICATION

- NLP Classification is generally more shallow
  - Logistic Regression/Naïve Bayes
  - Two layer CNN
- This is starting to change
  - The triumph of pre-training and transfer learning