Neural Networks and Deep Learning

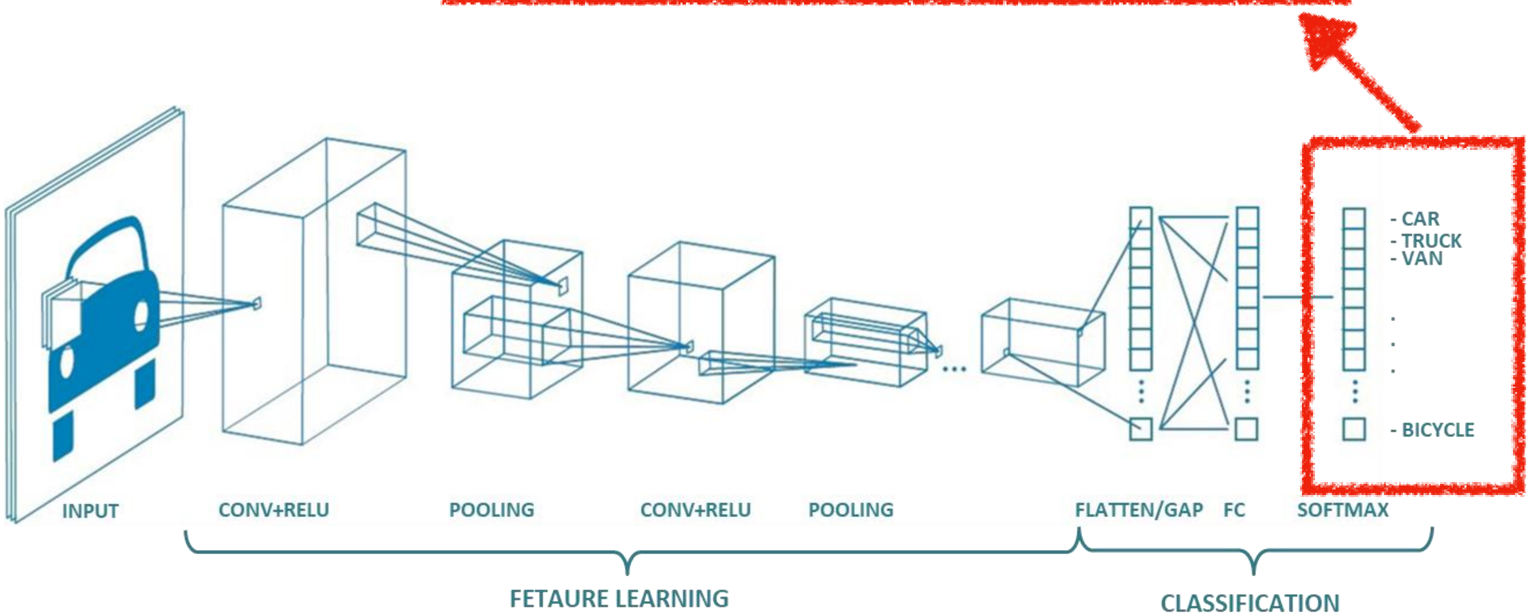
Metric Learning

Credit to: Moitreya Chatterjee, Yunan Luo

Supervised Learning so far

$$s \longrightarrow \boxed{\text{Softmax}} \longrightarrow \boxed{\text{Cross-Entropy Loss}}$$

$$f(s)_i = \frac{e^{s_i}}{\sum_j^C e^{s_j}} \qquad CE = -\sum_i^C t_i log(f(s)_i)$$



Few-shot Learning

Supervised ML focused on learning from limited number of examples, inspired by human/animal ability to rapidly generalize from few examples.

Typical scenarios:

- Learning for rare cases – when obtaining labelled data is hard or impossible
- Reducing data gathering effort and computational cost

Variations:

- (<50, e.g. 5,10)-shot learning: General case, where only few examples are given
- 1-shot learning: Extreme case, where only one data example is given (e.g. 1 image per class)
- 0-shot learning: Instead of image, we have description of new class

e.g.: person re-id   hundreds of pedestrians, 2-5 images per pedestrian

1 – Y. Wang et al, Generalizing from a Few Examples: A Survey on Few-Shot Learning, 2019 Arxiv: 1904.05046

Few-shot learning

Ned Stark

Robert Baratheon

Daenerys Targaryen

1-shot learning

Ned Stark

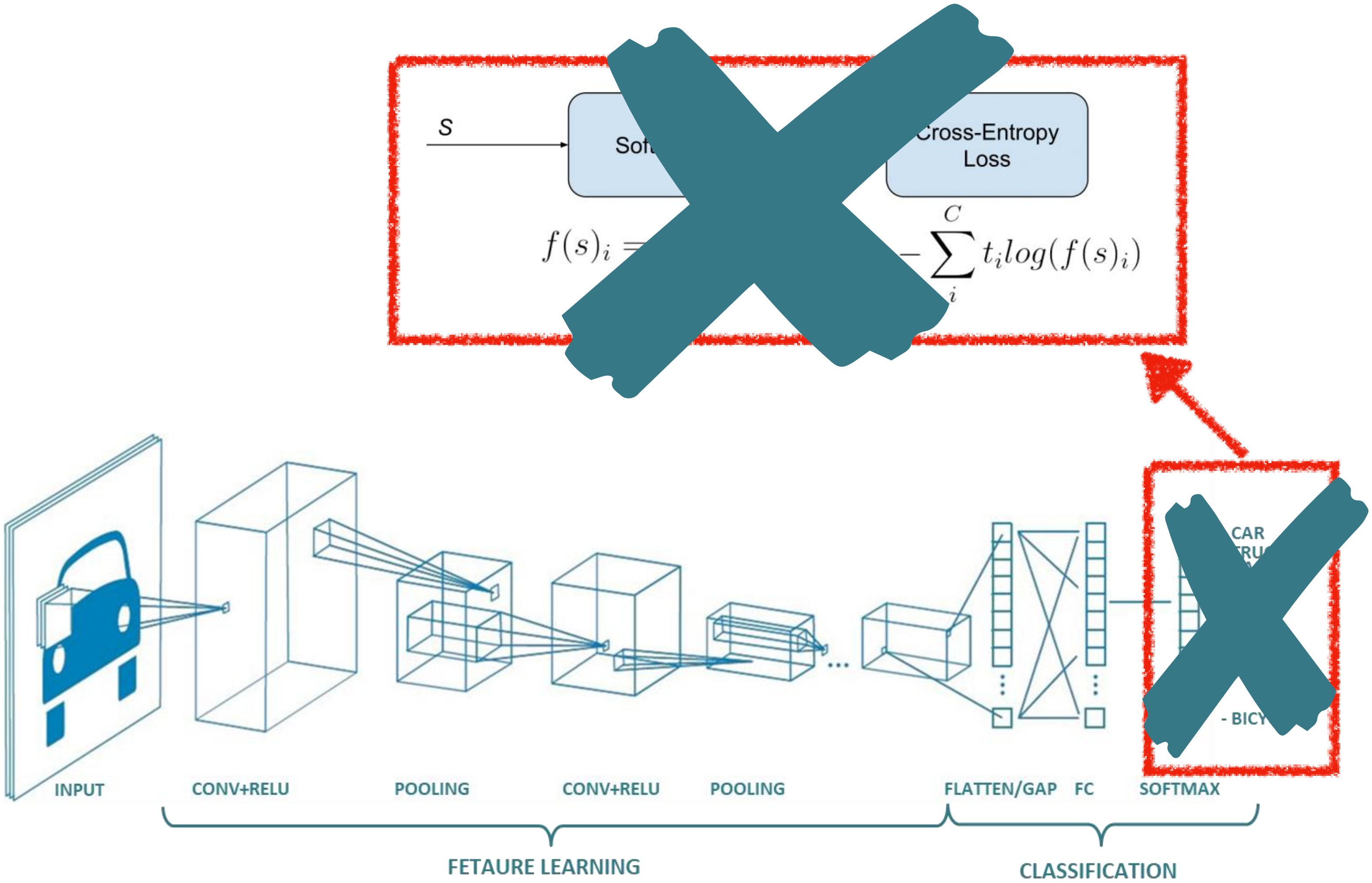
Robert Baratheon

Daenerys Targaryen

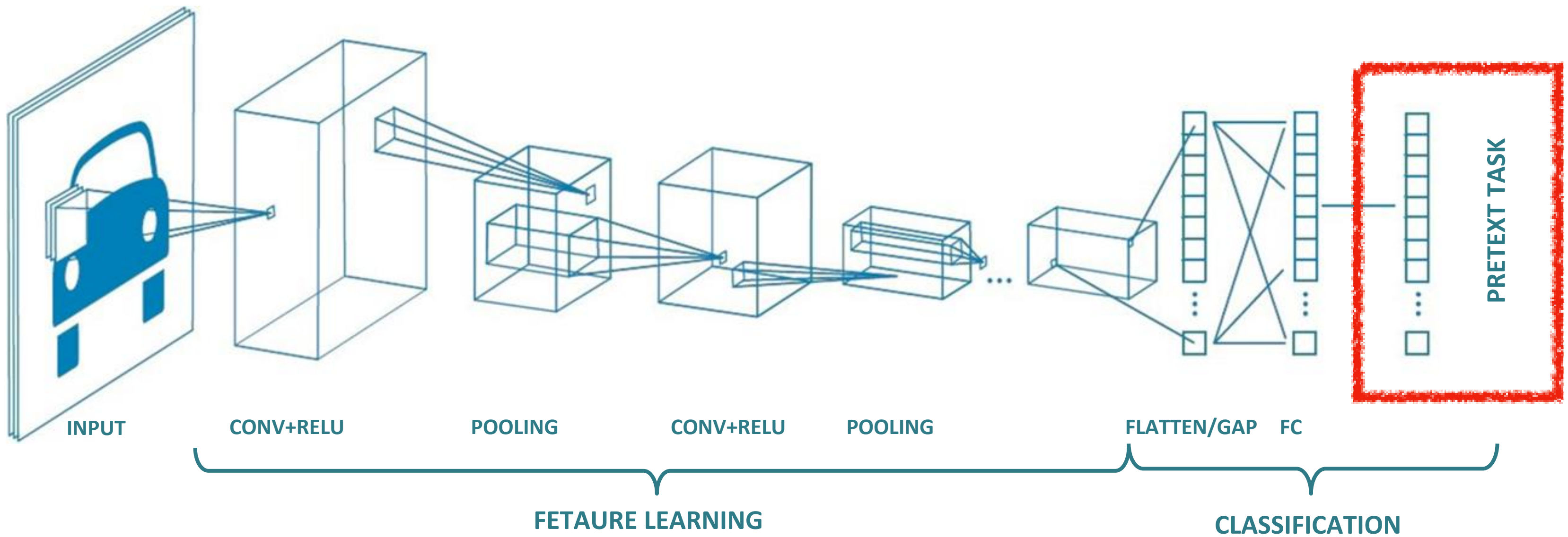
0-shot learning

"Blond, long hair, blue eyes"

The Self-Supervised Revolution



The Self-Supervised Revolution



Focus on Feature Learning for a 'random' given task (Pretext Task) e.g:

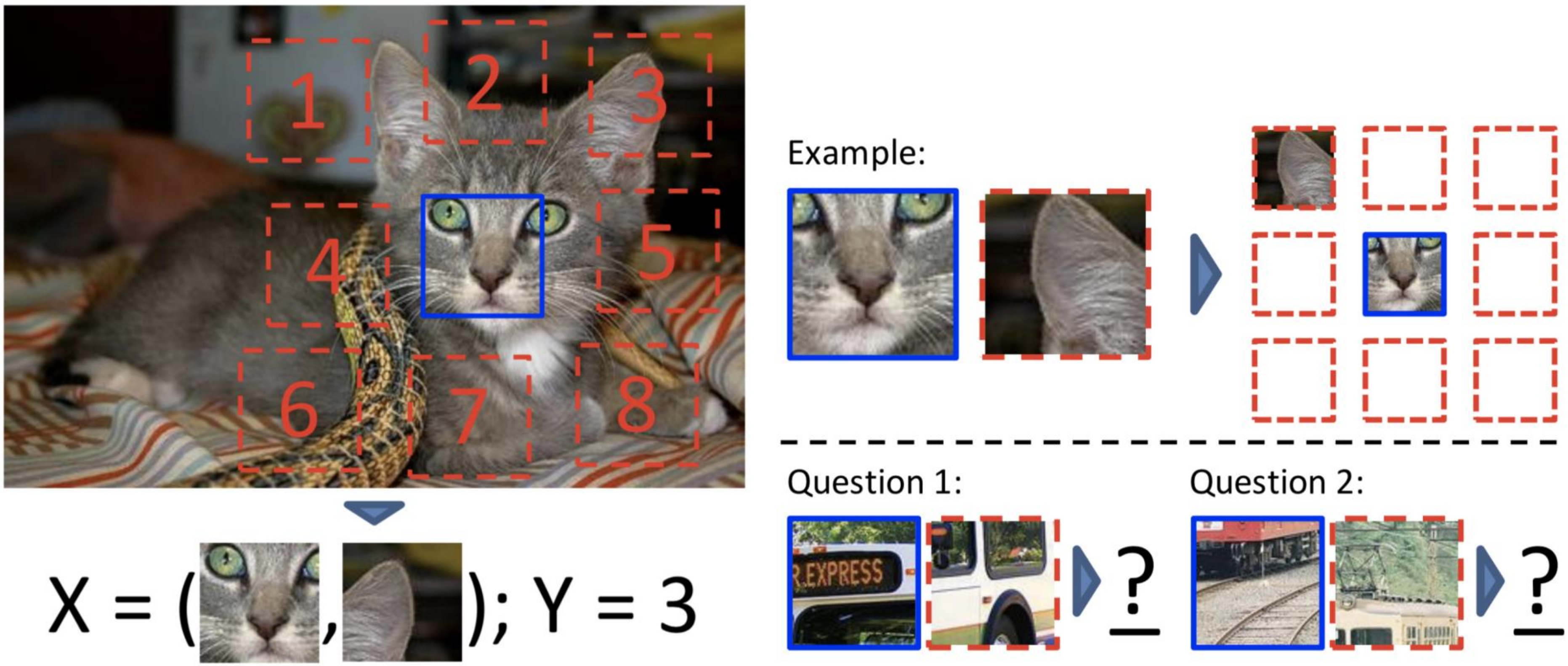
- Center word, next sentence, sentence swapping, ...

Pretext tasks: Tons of data, easy to extract goal

https://amitness.com/2020/05/self-supervised-learning-nlp/

Learning from pretext tasks

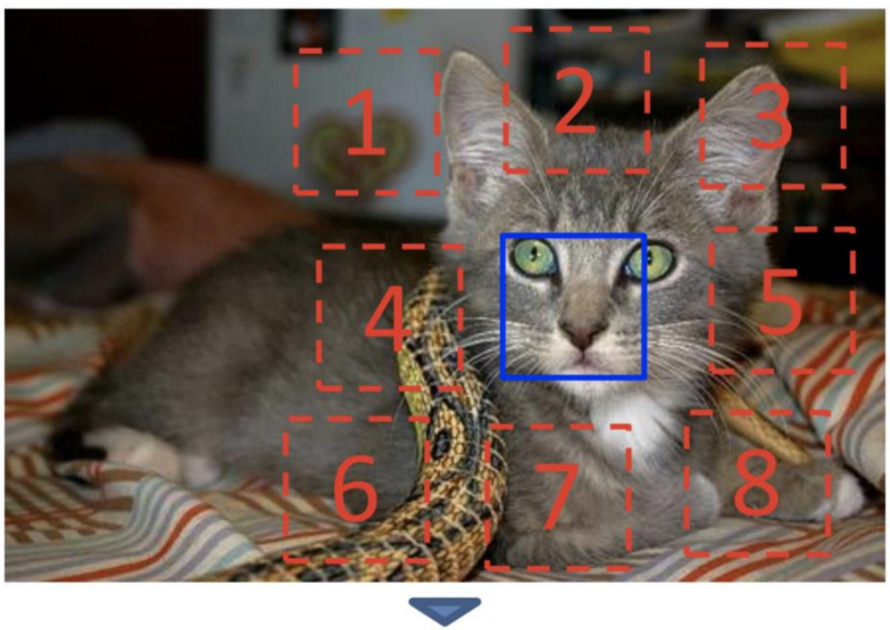
Jigsaw puzzle: what is the relative position

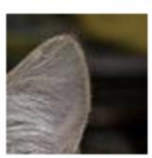


[Doersch et al., 2015]

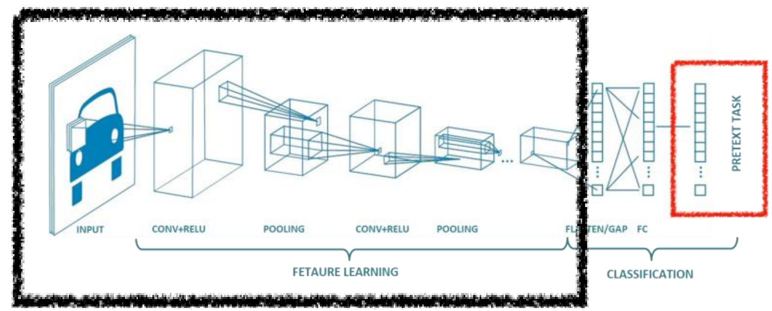
Learning from pretext tasks

Jigsaw puzzle: what is the relative position

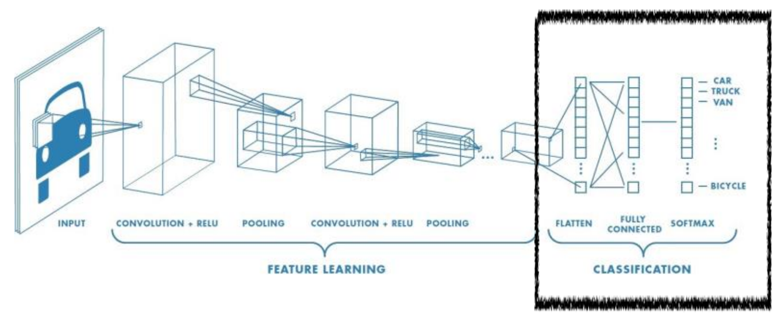


X = (, ); Y = 3

1. Train Feature Representation



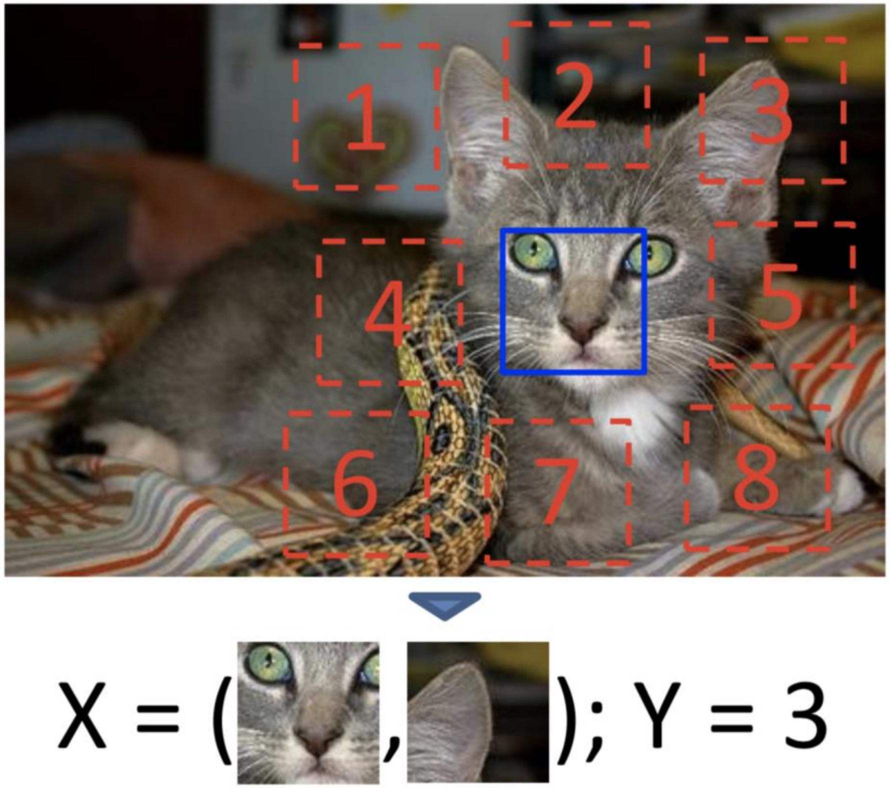
2. Finetune classifier w/ some labeled data



[Doersch et al., 2015]

Learning from pretext tasks

Jigsaw puzzle: what is the relative position

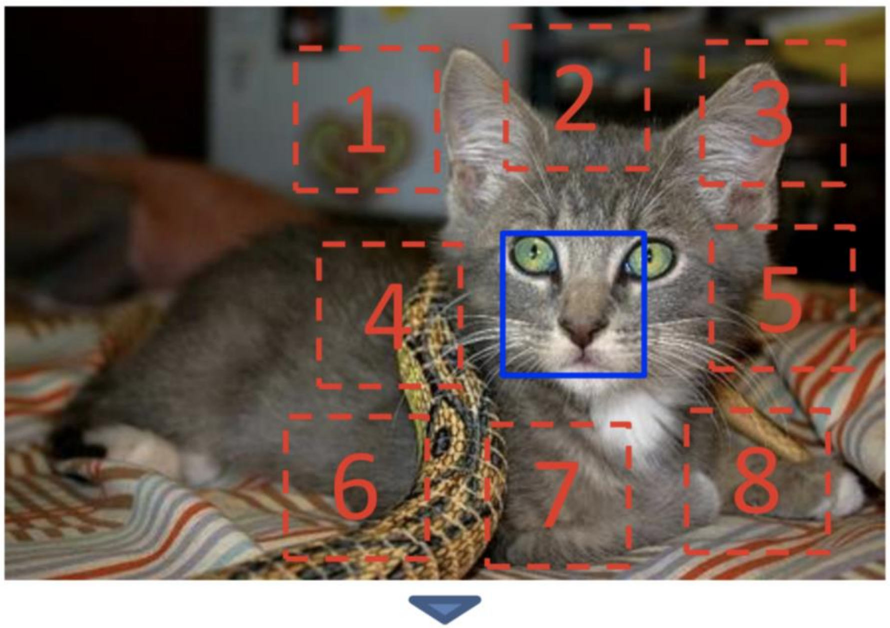


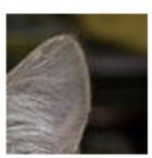
Results are good when a lot of unlabeled data is available and little labeled data available

[Doersch et al., 2015]

Learning from pretext tasks

Jigsaw puzzle: what is the relative position



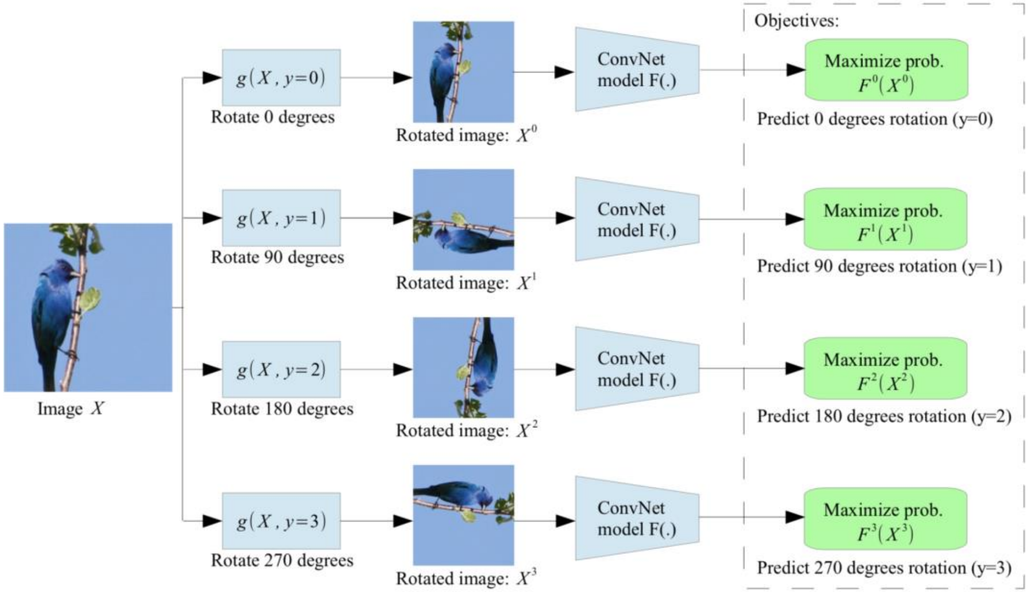
X = (, ); Y = 3

Shortcut problem: The Network always tries to “cheat”.
Ex: network can use the camera lens distortion

(Doersch et al. found a solution for this particular problem)

[Doersch et al., 2015]

Learning from pretext tasks



There are other pretext tasks

- Rotation
- In-painting
- Colorization
- Up to your imagination…

[Gidaris et al., 2018]

Pretext Tasks for NLP

- Center Word Prediction (CBOW)
- Neighbour Word Prediction (skip-gram)
- Neighbour Sentence Prediction (skip-gram, sentence level)
- Auto-regressive Language Modeling (predict next word)
- Masked Language Modeling (e.g. BERT)
- Next Sentence Prediction (classify if sentence comes from same or different document)
- Sentence Order Prediction (classify if correct order or not)
- Sentence Permutation (re-order messed up sentences)
- Emoji Prediction

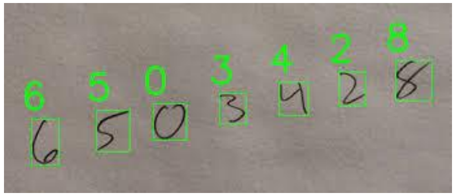
Does any embedded space suffice?

Most ML algorithms are based on comparing samples and to decide if they ‘are the same’, and to define borders between different concepts.

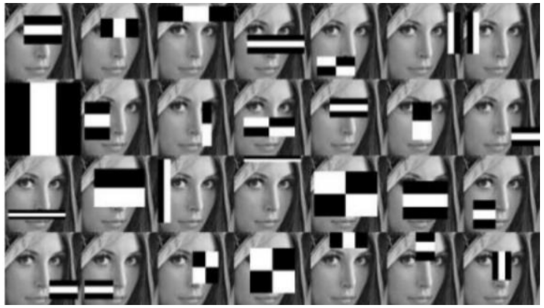
Goal: define the **similarity** between subjects.

Need for Similarity Measures

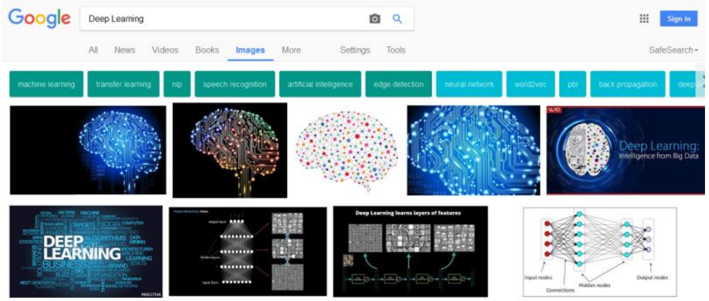
- Recognizing handwriting in checks.



- Automatic detection of faces in a camera image.



- Search Engines, such as Google, matching a **query** (could be text, image, etc.) with a set of **indexed documents** on the web.



Notion of a Metric

- A **Metric** is a function that quantifies a "distance" between every pair of elements in a set, thus inducing a measure of similarity.
- A metric f(x,y) must satisfy the following properties for all x, y, z belonging to the set:
  - *Non-negativity*: $f(x, y) \geq 0$
  - *Identity of Discernible*: $f(x, y) = 0 \Leftrightarrow x = y$
  - *Symmetry*: $f(x, y) = f(y, x)$
  - *Triangle Inequality*: $f(x, z) \leq f(x, y) + f(y, z)$

Types of Metrics

In broad strokes metrics are of two kinds:

- **Pre-defined Metrics**: Metrics which are fully specified without the knowledge of data.

  E.g. Euclidian Distance: $f(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T(\mathbf{x} - \mathbf{y})$

- **Learned Metrics**: Metrics which can only be defined with the **knowledge** of the **data**.

  E.g. Mahalanobis Distance: $f(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T\mathbf{M}(\mathbf{x} - \mathbf{y})$; where **M** is a matrix that is estimated from the data.

  Learned Metrics are of two types:

  - **Unsupervised** : Use unlabeled data
  - **Supervised** : Use labeled data

DEEP LEARNING TECHNIQUES

Deep Learning to the Rescue!

CNNs can **jointly optimize** the representation of the input data conditioned on the “similarity” measure being used, aka end-to-end learning.

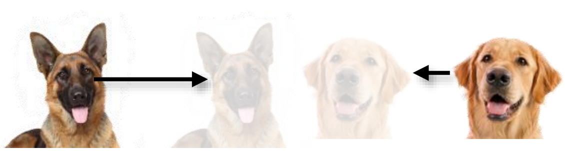
State the Problem

Input: Given a pair of input images, we want to know how “similar” they are to each other.

Output: The output can take a variety of forms:

- Either a binary label, i.e. 0 (same) or 1 (different).
- A **Real** number indicating how similar a pair of images are.

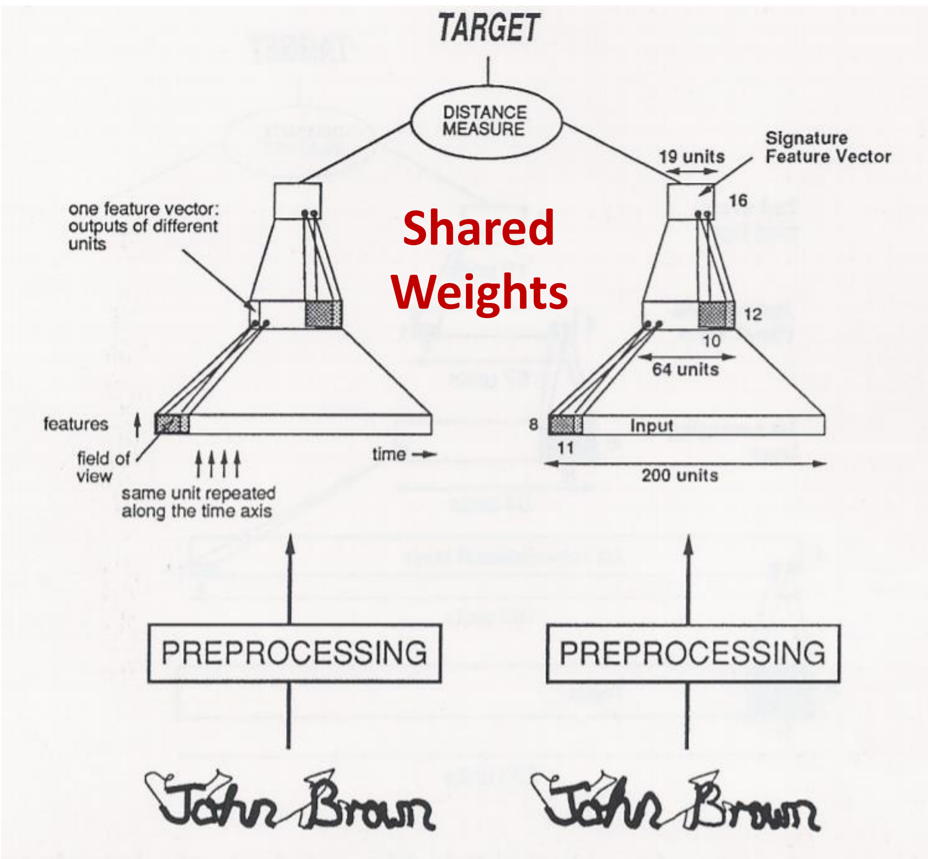
minimize



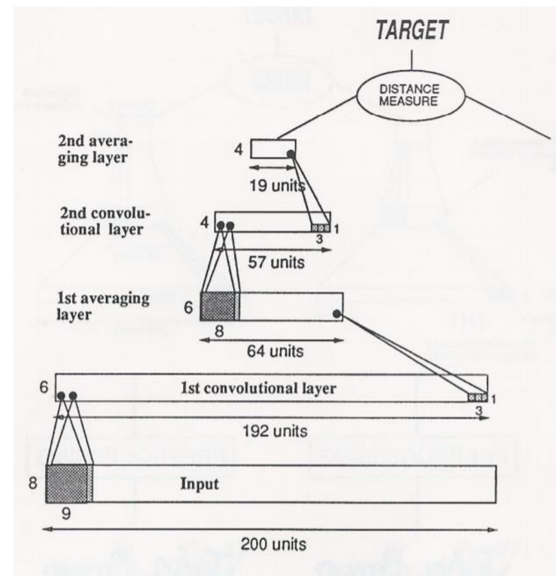
maximize

Typical Siamese CNN

- **Input**: A pair of input signatures.
- **Output (Target)**: A label, 0 for **similar, 1 else**.

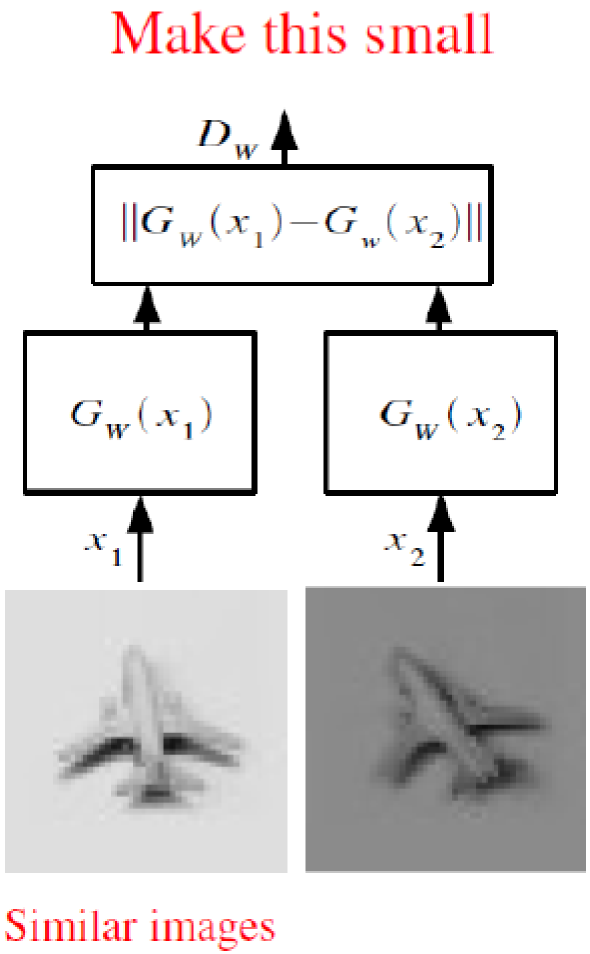


Shared Weights



Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E. and Shah, R., 1993. Signature Verification Using A "Siamese" Time Delay Neural Network. *IJPRAI*, *7*(4), pp.669-688.

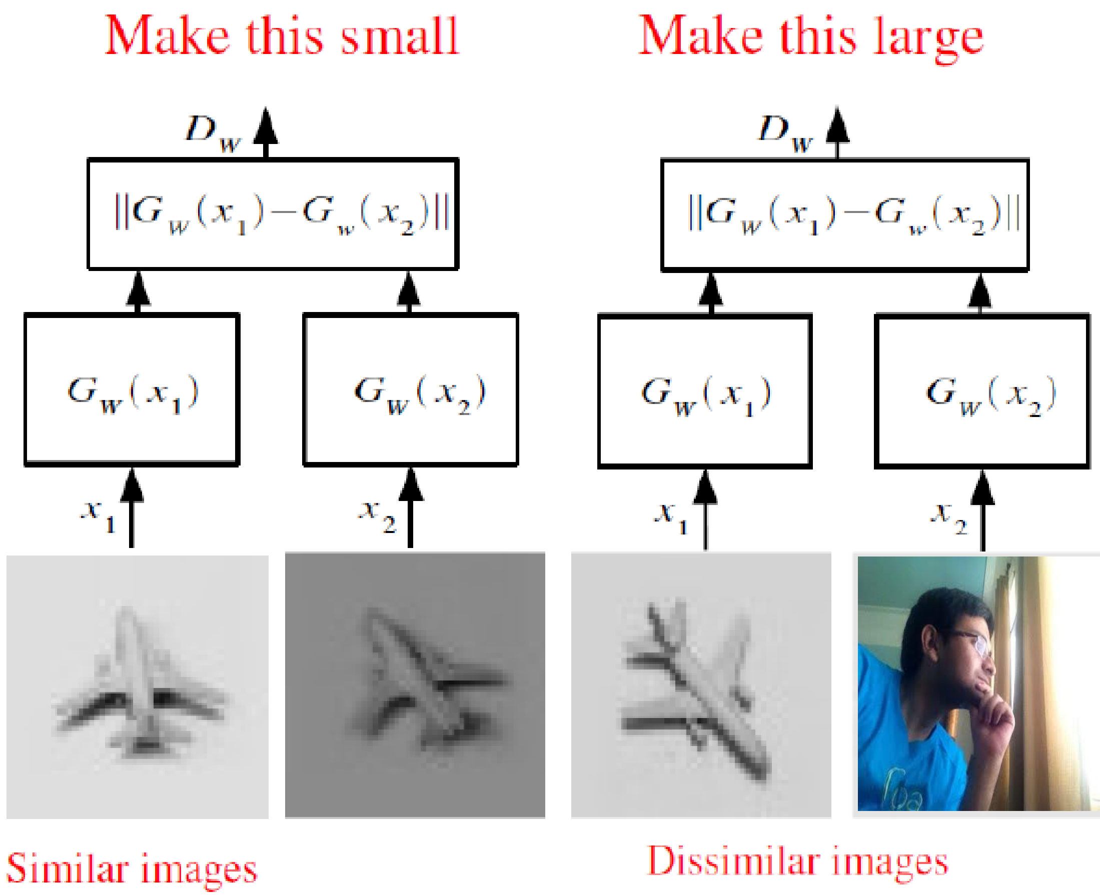
Siamese CNN – Loss Function



- **Is there a problem with this formulation?**
  - The model could learn to embed every input to the same point, i.e. predict a **constant** as **output**.
  - In such a case, every pair of input would be categorized as a positive pair.

Chopra, S., Hadsell, R. and LeCun, Y., 2005, June. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (Vol. 1, pp. 539-546). IEEE.

Siamese CNN – Loss Function



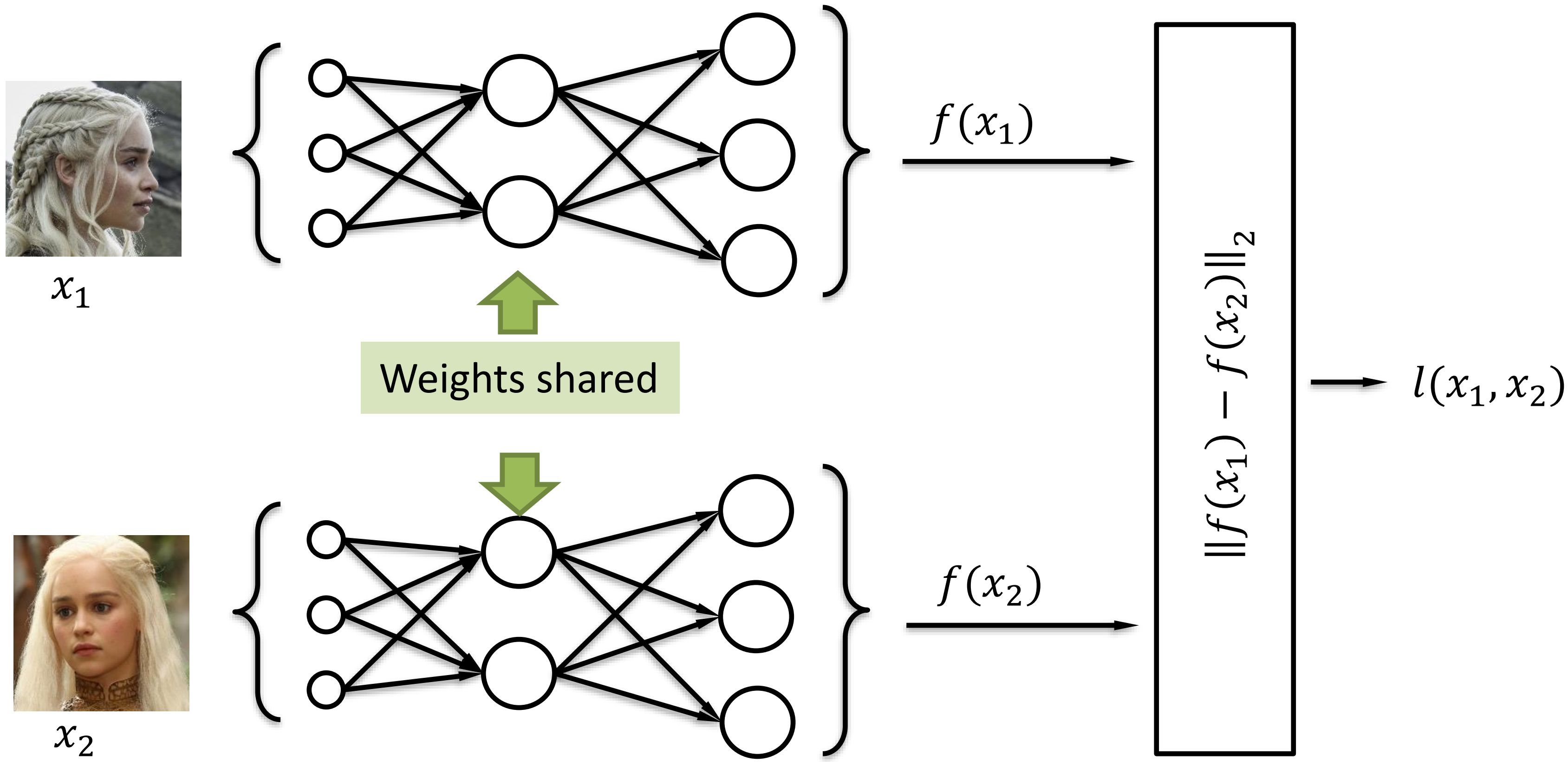
The final loss is defined as :

L = ∑ loss of positive pairs + ∑ loss of negative pairs

Chopra, S., Hadsell, R. and LeCun, Y., 2005, June. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (Vol. 1, pp. 539-546). IEEE.

Siamese networks

Two parallel feed-forward networks, with **shared weights**. It's the same network actually encoding two different inputs. Shared weights, like in RNNs.

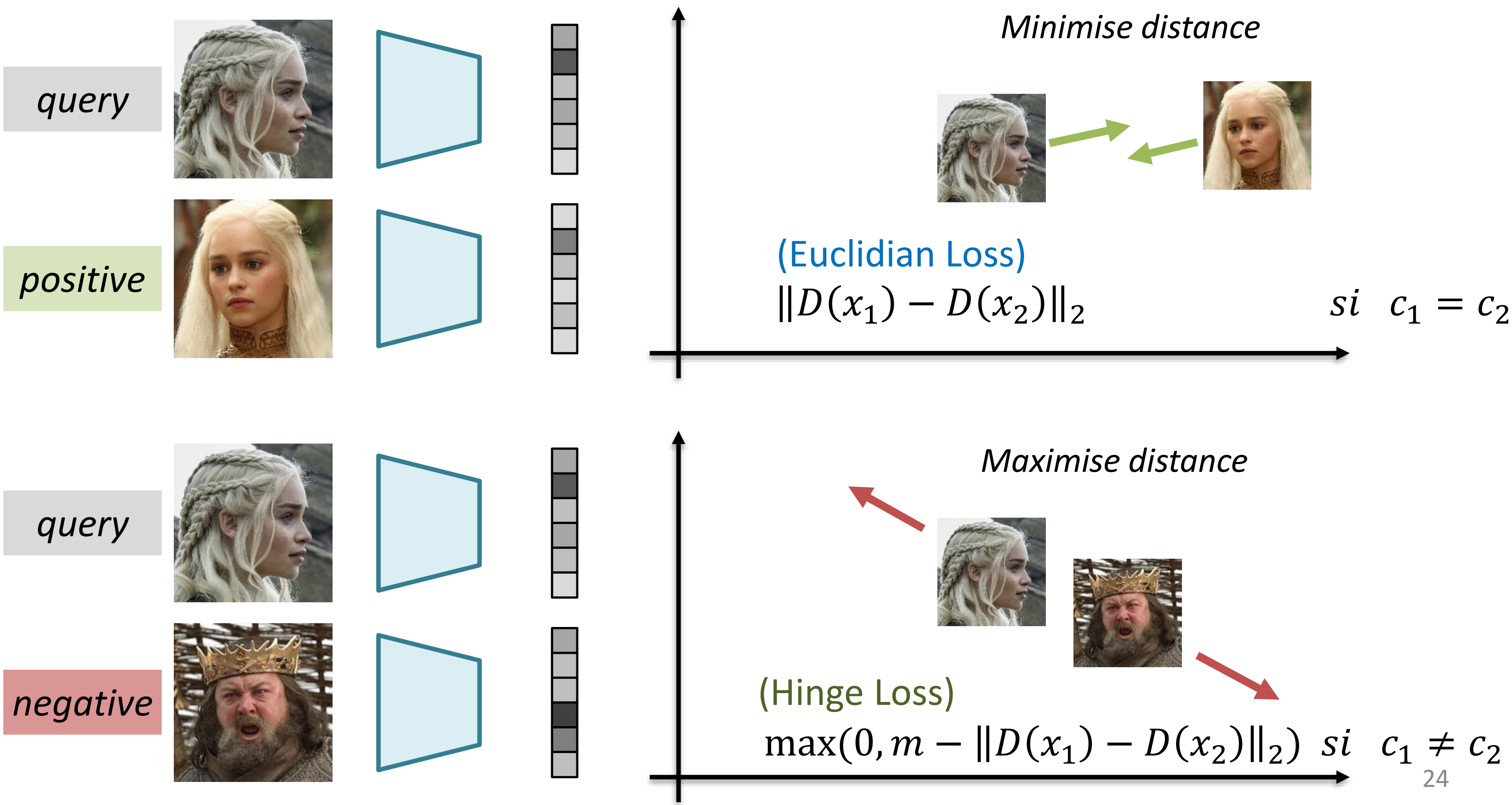


$$l(x_1,\, x_2) = \begin{cases} \|D(x_1) - D(x_2)\|_2 & si \quad c_1 = c_2 \\ \max(0, m - \|D(x_1) - D(x_2)\|_2) & si \quad c_1 \neq c_2 \end{cases}$$

Bromley, J. et al, 1993. Signature Verification Using A "Siamese" Time Delay Neural Network. IJPRAI, 7(4), pp.669-688.

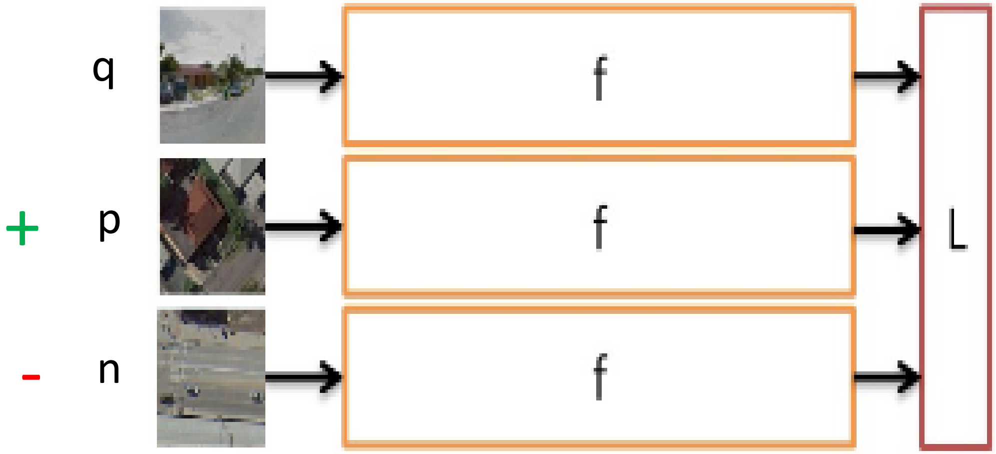
Pairwise comparisons

Given a pair of input images, we want to place them closer together if they are *similar*, and far from each other if they are not



Siamese CNN – Variants

TRIPLET NETWORK



$\textbf{Dist}(f(q), f(p)) < \textbf{Dist}(f(q), f(n))$

- Compare triplets in one go.
- Check if the sample in the **topmost** channel, is more similar to the one in the middle or the one in the bottom.
- Allows us to learn ranking between samples.

Vo, N.N. and Hays, J., 2016, October. Localizing and orienting street views using overhead imagery. In European Conference on Computer Vision (pp. 494-509).

Siamese CNN – Loss Function

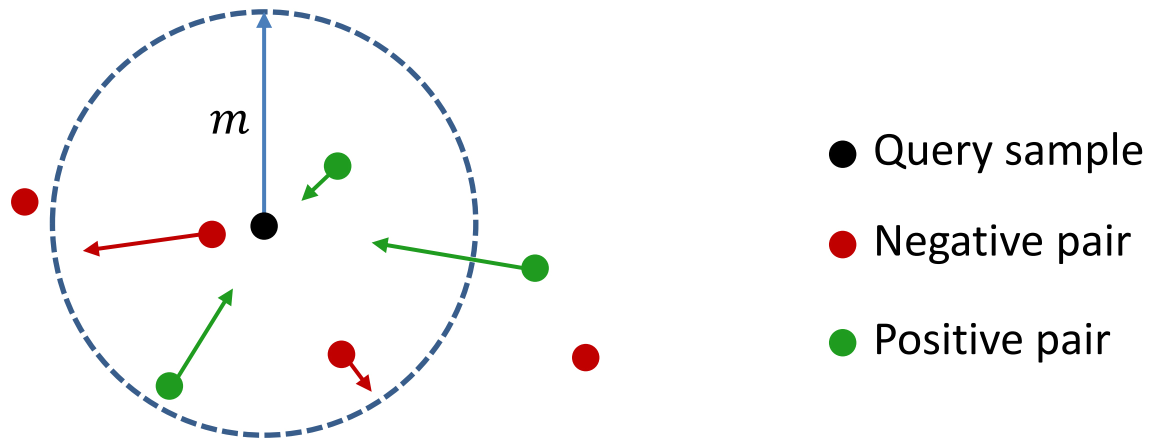
Siamese

$$l(x_q, x_a) = \begin{cases} \left\|D\left(x_q\right) - D(x_a)\right\|_2 & if\ a = p \\ \max(0, m \quad - \left\|D\left(x_q\right) - D(x_a)\right\|_2) & if\ a = n \end{cases}$$

Contrastive Loss

Triplet

$$l(x_q, x_p, x_n) = \left\|D\left(x_q\right) - D\left(x_p\right)\right\|_2 + max\left((0, m \quad - \left\|D\left(x_q\right) - D(x_n)\right\|_2\right)$$



$D(x)$: x description

MINING

Types of negatives

Easy negatives:

$$d(x_q, x_n) > d(x_q, x_p) + m$$

The negative sample is already sufficiently distant to the query (anchor) sample with respect to the positive sample in the embedding space. The loss is 0 and the net parameters are not updated.

Hard Negatives:

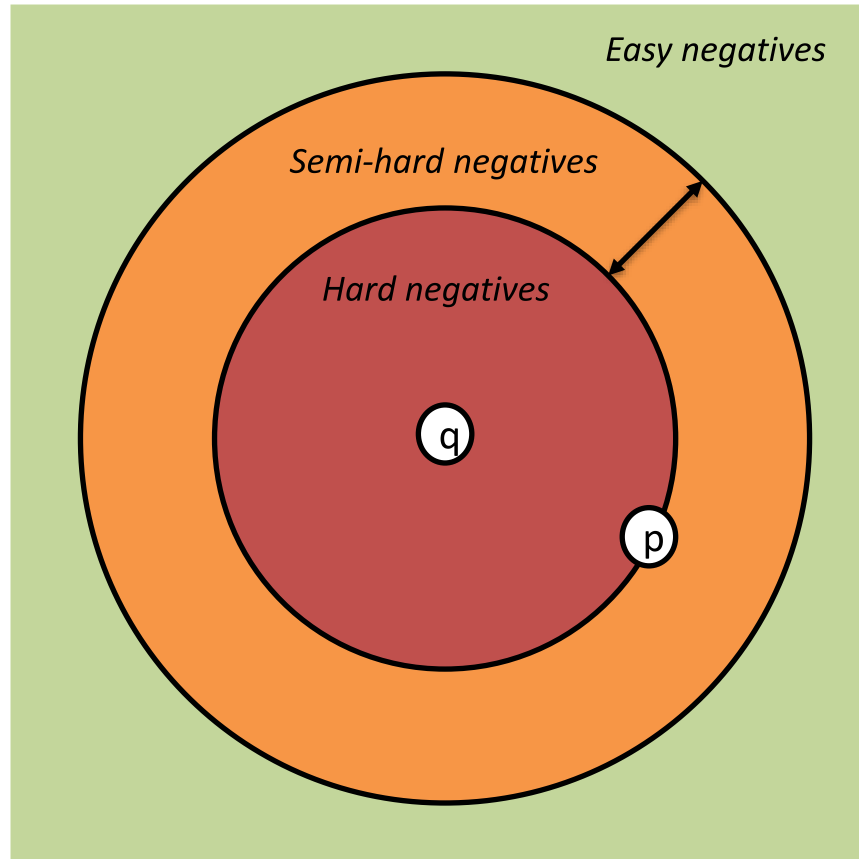
$$d(x_q, x_n) < d(x_q, x_p)$$

The negative sample is closer to the anchor than the positive. The loss is positive and greater than m.

Semi-Hard Negatives:

$$d(x_q, x_p) < d(x_q, x_n) < d(x_q, x_p) + m$$

The negative sample is more distant to the anchor than the positive, but the distance is not greater than the margin, so the loss is still positive (and smaller than m).

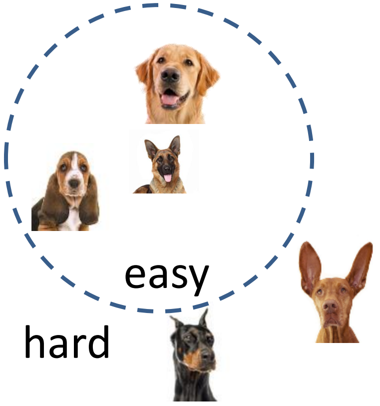


Mining training data

select a random query (e.g. dog):

Positives

Take the hardest positive samples

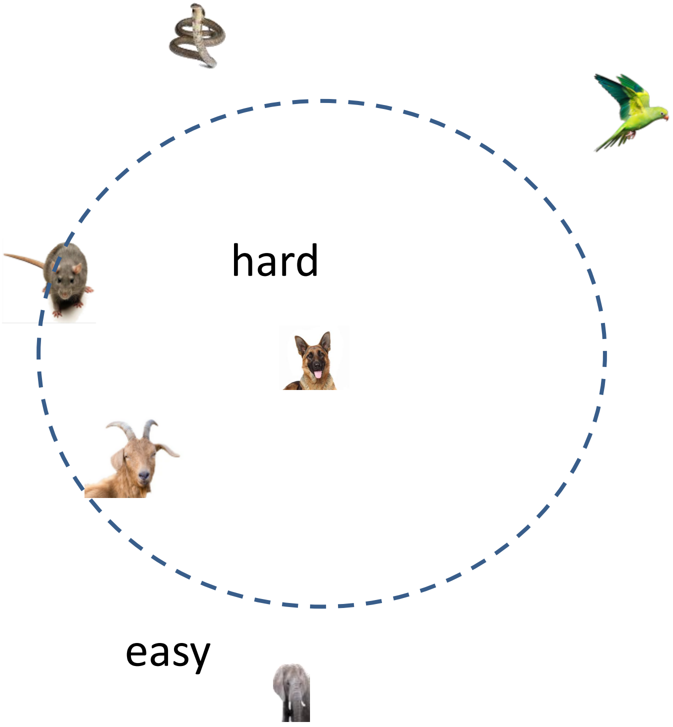


Mining training data

select a random query (dog): 

Positives 

Negatives 

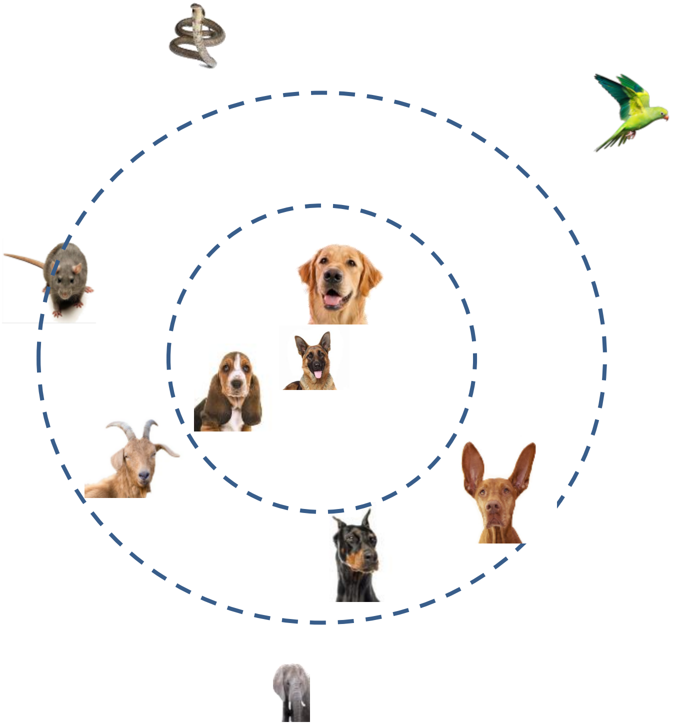


Mining training data

select a random query (dog): 

Positives 

Negatives 

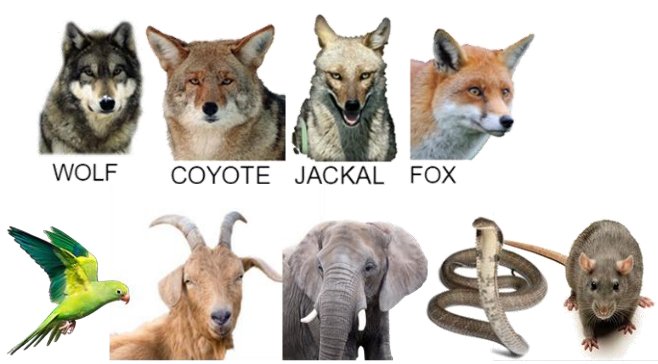


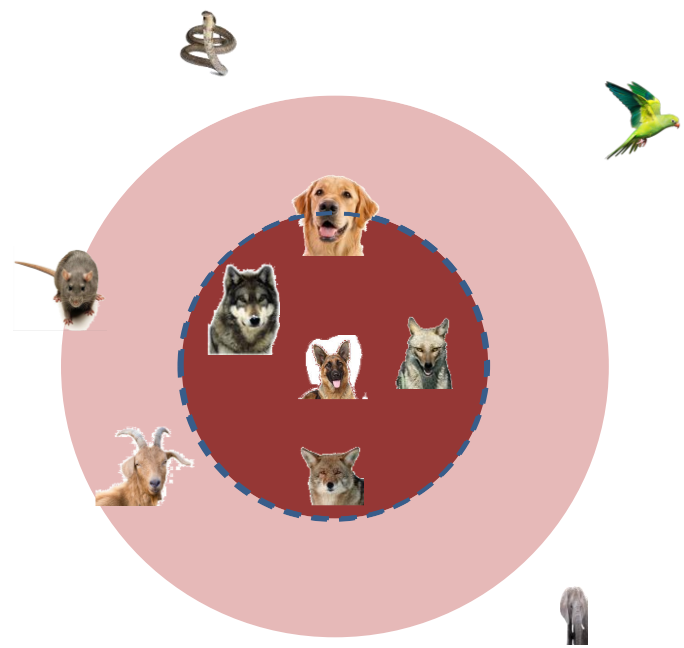
Mining training data

select a random query (dog):

Fix a positive

Negatives





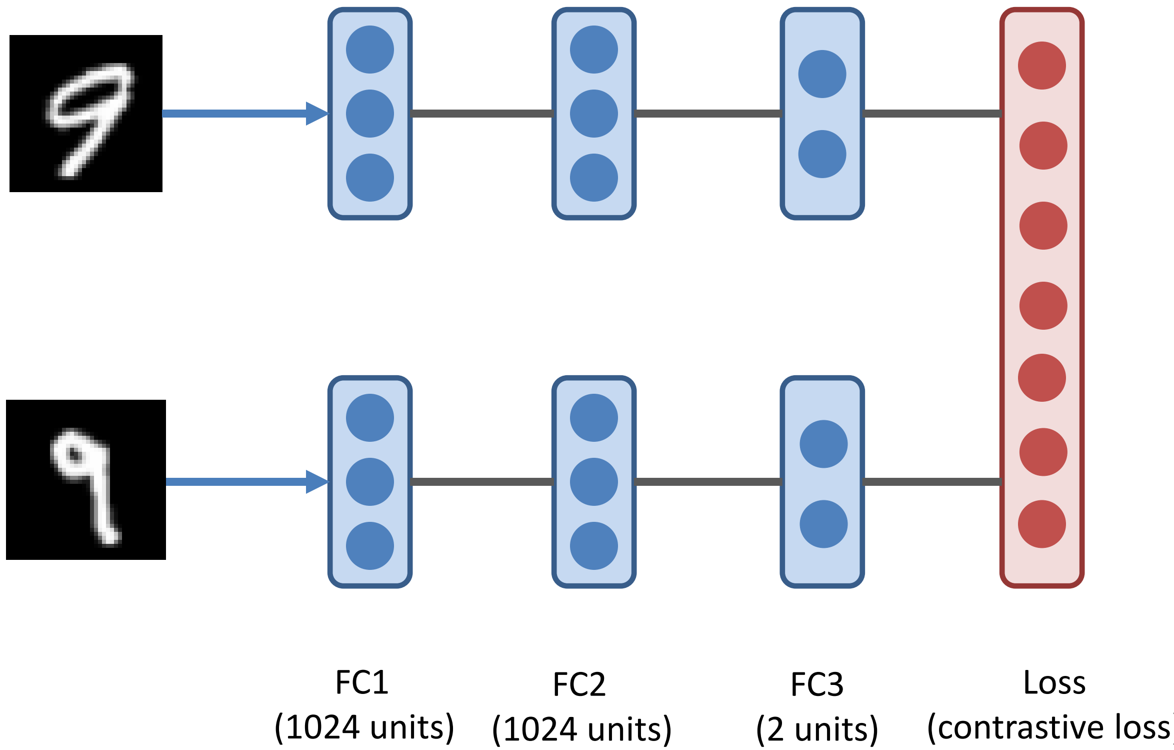
Hard negatives

Semi-hard negatives

ARE EMBEDINGS MEANINGFUL?

Metric

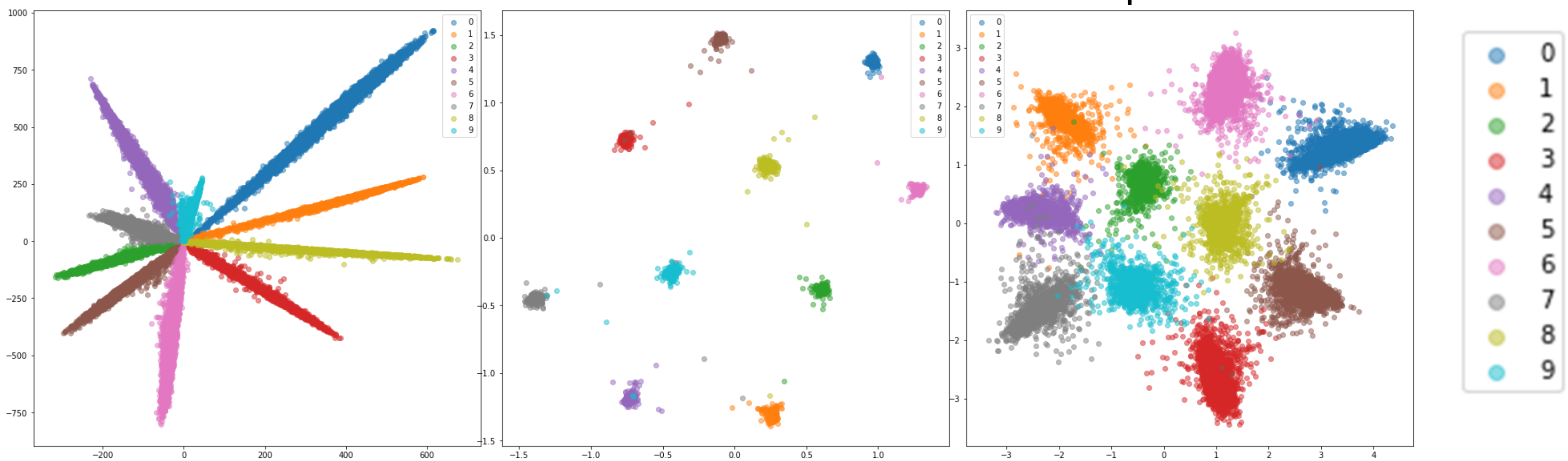
MNIST Digit Similarity Assessment



Code: @ywpkwon

Mnist test

softmax siamese triplet



Network architecture

conv 32 5x5 -> PReLU -> MaxPool 2x2 ->
conv 64 5x5 -> PReLU -> MaxPool 2x2 ->
Dense 256 -> PReLU -> Dense 256 -> PReLU -> Dense 2

https://awesomeopensource.com/project/adambielski/siamese-triplet

APPLICATIONS

Person Re-Identification

CUHK03 Dataset

Person Re-Identification

True positive

True negative

Ahmed, E., Jones, M. and Marks, T.K., 2015. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3908-3916).

MULTIMODAL LEARNING


THE
SNUGGL
EAL

Predicting the "topic" of an image

VISUAL INFORMATION

TEXTUAL INFORMATION

Scene Text Annotations

TELEPHONE

Image Captions

A big red telephone booth that a man is standing in.

A person standing inside of a telephone booth.

This is an image of a man in a phone booth.

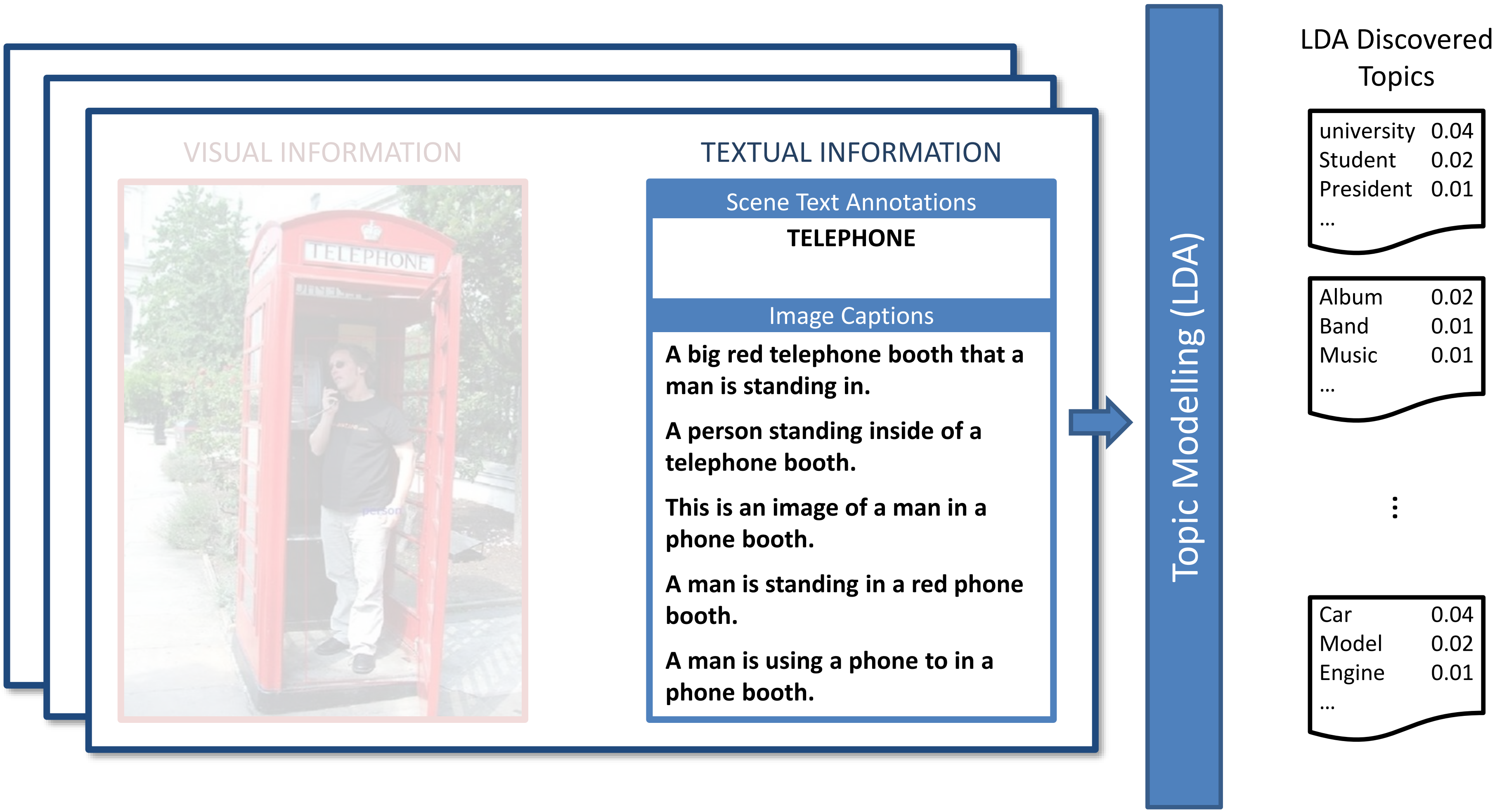
A man is standing in a red phone booth.

A man is using a phone to in a phone booth.

TRAINING SAMPLE

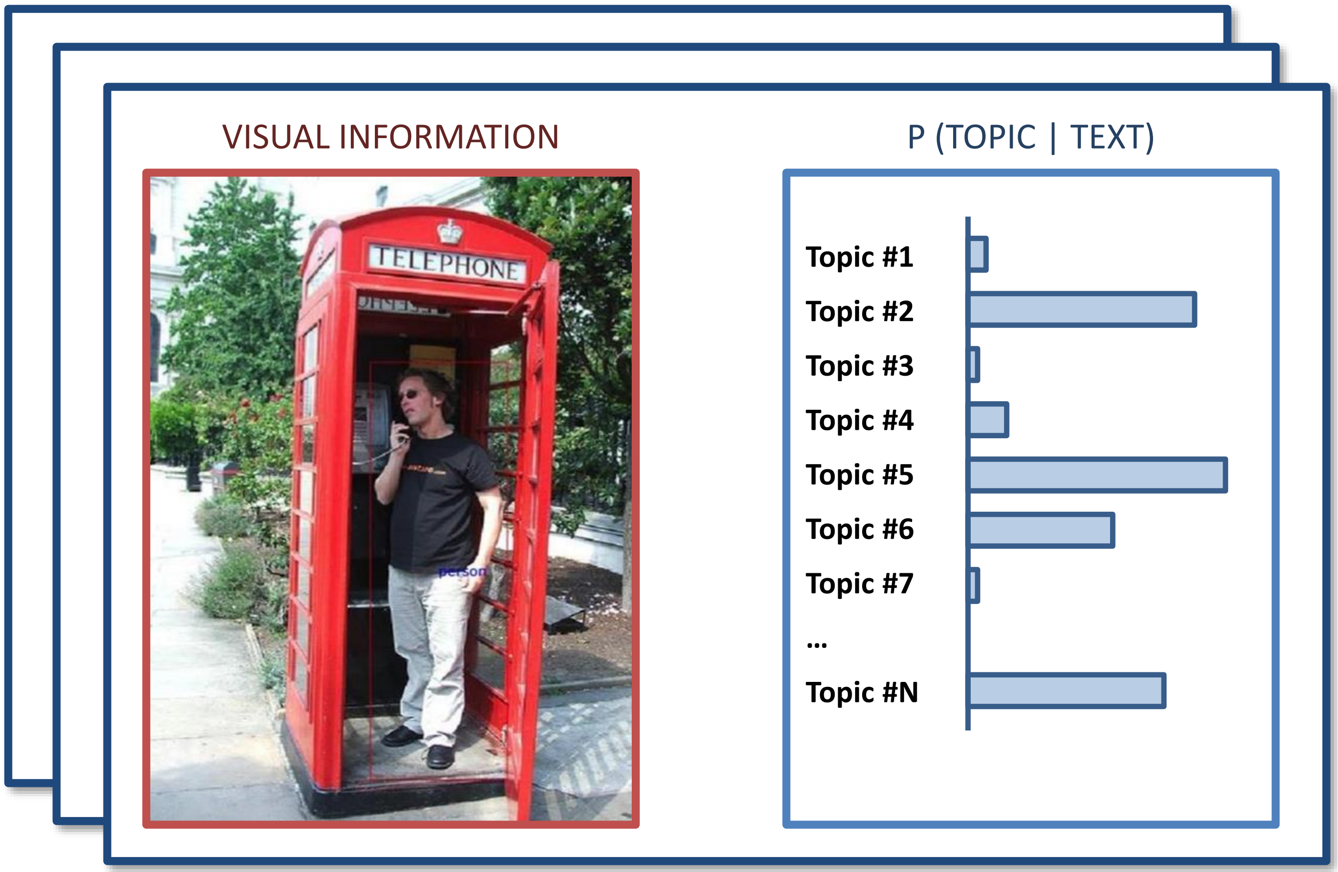
Y. Patel, L. Gomez, M. Rusiñol, D. Karatzas, "Dynamic Lexicon Generation for Natural Scene Images", IWRR 2016

Predicting the “topic” of an image



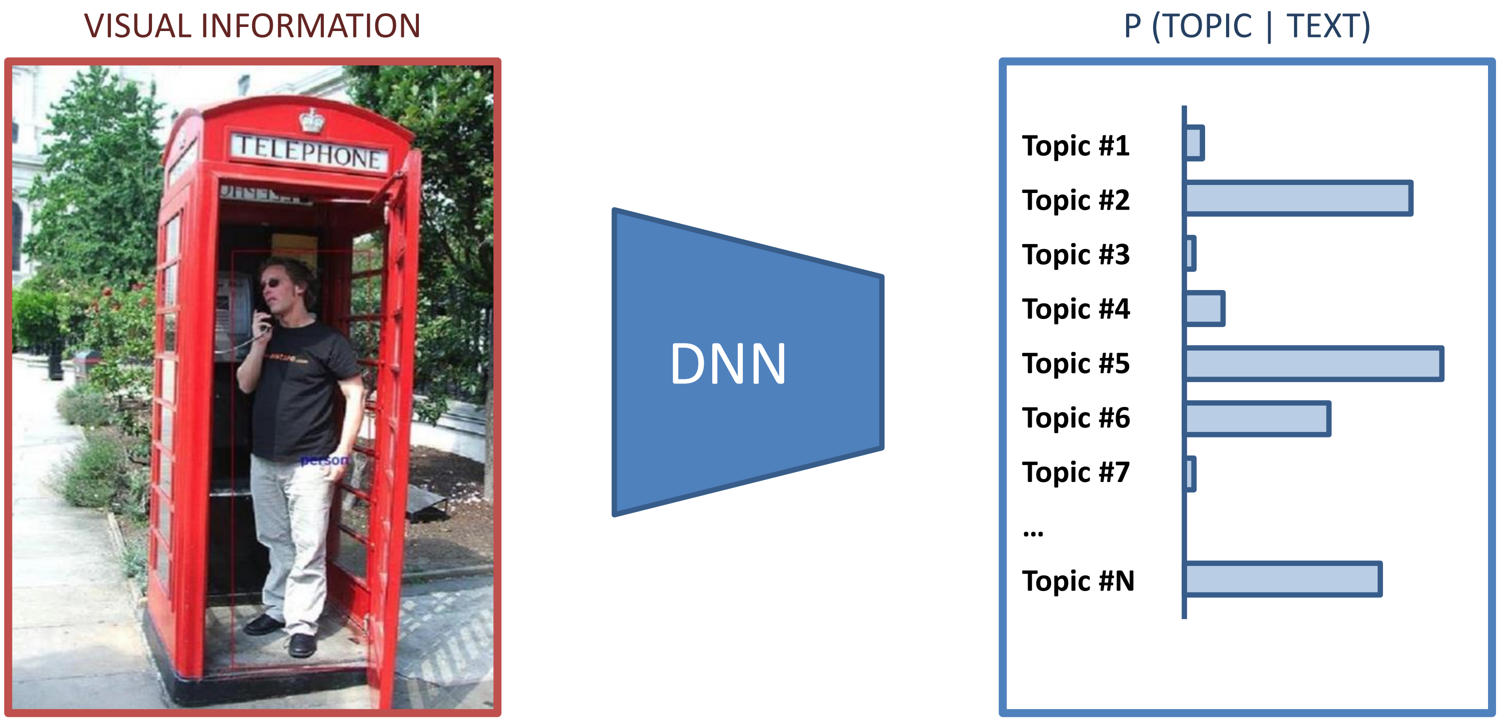
Y. Patel, L. Gomez, M. Rusiñol, D. Karatzas, “Dynamic Lexicon Generation for Natural Scene Images”, IWRR 2016

Predicting the “topic” of an image



Y. Patel, L. Gomez, M. Rusiñol, D. Karatzas, “Dynamic Lexicon Generation for Natural Scene Images”, IWRR 2016

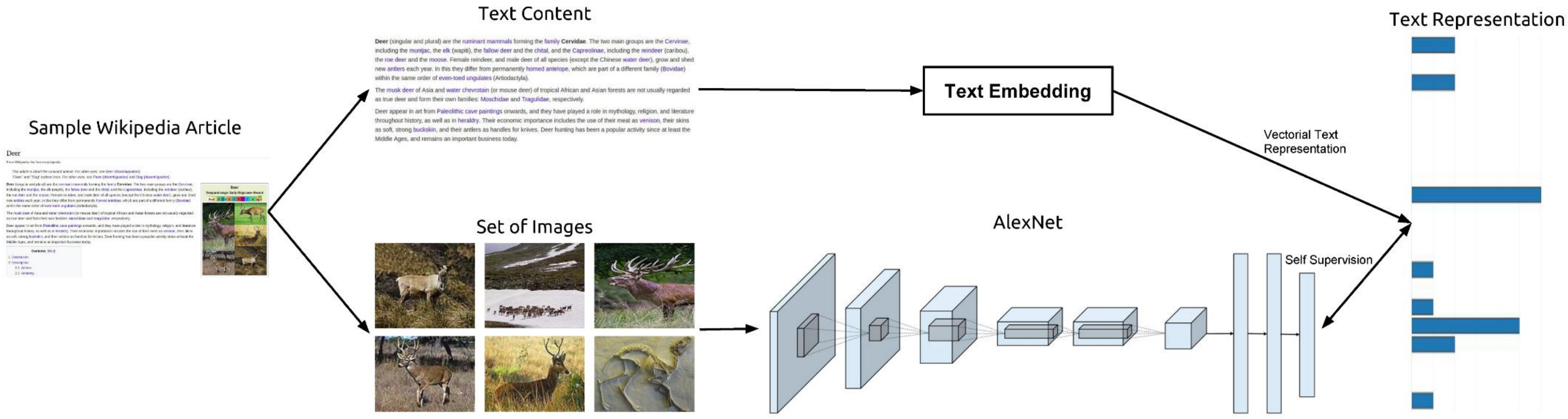
Predicting the "topic" of an image



Y. Patel, L. Gomez, M. Rusiñol, D. Karatzas, "Dynamic Lexicon Generation for Natural Scene Images", IWRR 2016

Learning to understand images by reading the Wikipedia

Task: Look at the image and predict what kind of article (topic) it illustrates



L. Gomez, Y. Patel, M. Rosinyol, D. Karatzas, "Self-supervised learning of visual features through embedding images into text topic spaces", CVPR 2017

Learning to understand images by reading the Wikipedia

Features learnt are meaningful. Units are selective to generic textures (grass, water) of shapes, objects and object-parts

L. Gomez, Y. Patel, M. Rosinyol, D. Karatzas, "Self-supervised learning of visual features through embedding images into text topic spaces", CVPR 2017

Self-supervised learning from Web Data

Wikipedia:
1.7M articles in English with 4.2M associated illusrtative images.

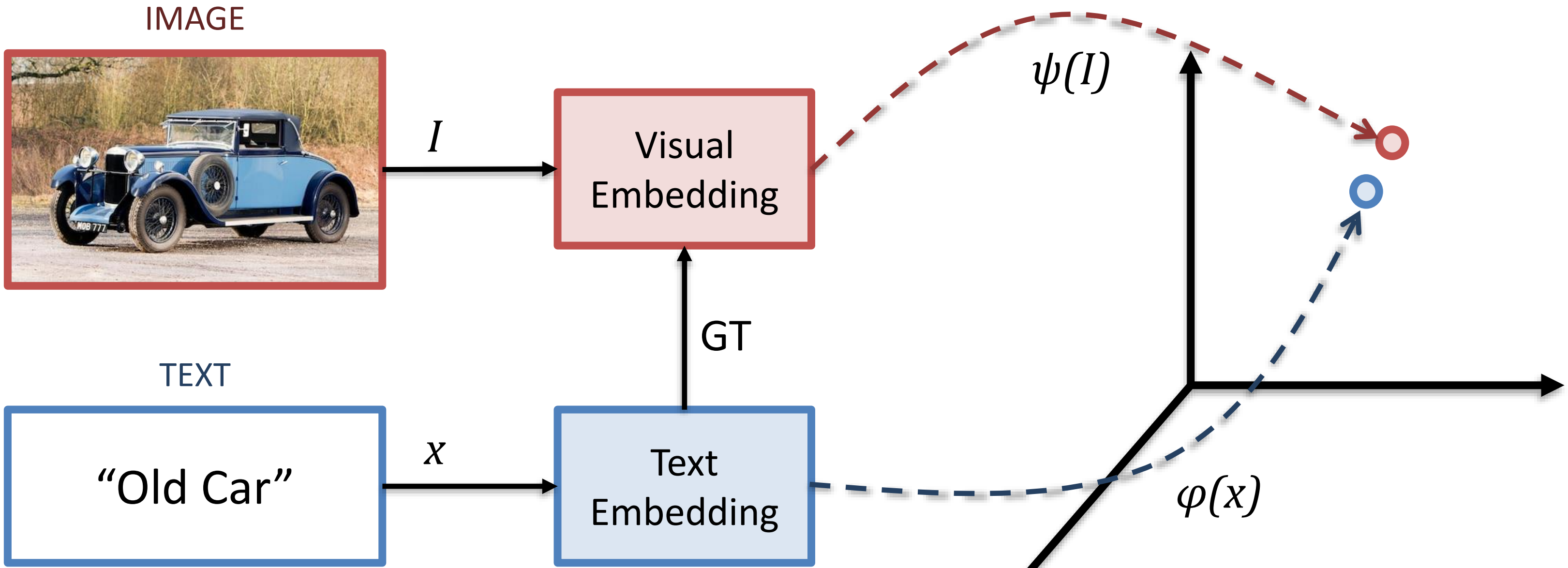
WebVision:
2.4M Flickr and Google images associated to ImageNet classes.

InstaCities1M:
1M Instagram images associated with one of the 10 most populated English speaking cities.

Gomez, R., Gomez, L., Gibert, J., & Karatzas, D. "Learning to Learn from Web Data through Deep Semantic Embeddings" MULA 2018

Gomez, R., Gomez, L., Gibert, J., & Karatzas, D. "Self-Supervised Learning from Web Data for Multimodal Retrieval", arXiv:1901.02004, 2019

Self-supervised learning from Web Data



Various sources of information

Various text embedding approaches (LDA, Word2Vec, Doc2Vec, GloVe, FastText, …).

The text embedding space **can be learned independently**.

Note that the embedding takes place in the selected text space.

This allows **exploiting properties of the text embedding space**.

Text-based semantic retrieval

“wild”

“happy”

“monday”

Model trained with Word2Vec on InstaCites1M

Text-based semantic retrieval

"haircut"

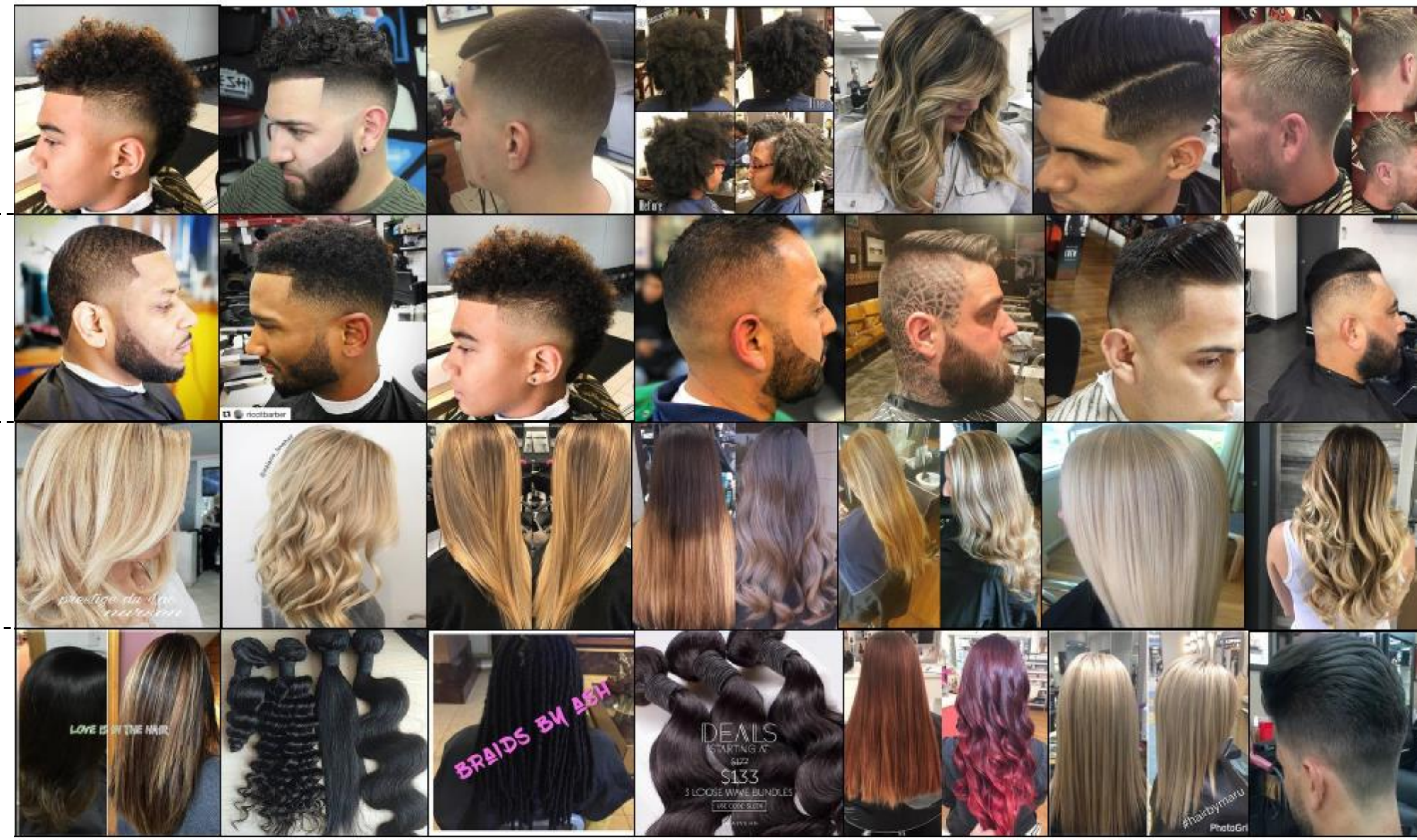
Text-based semantic retrieval

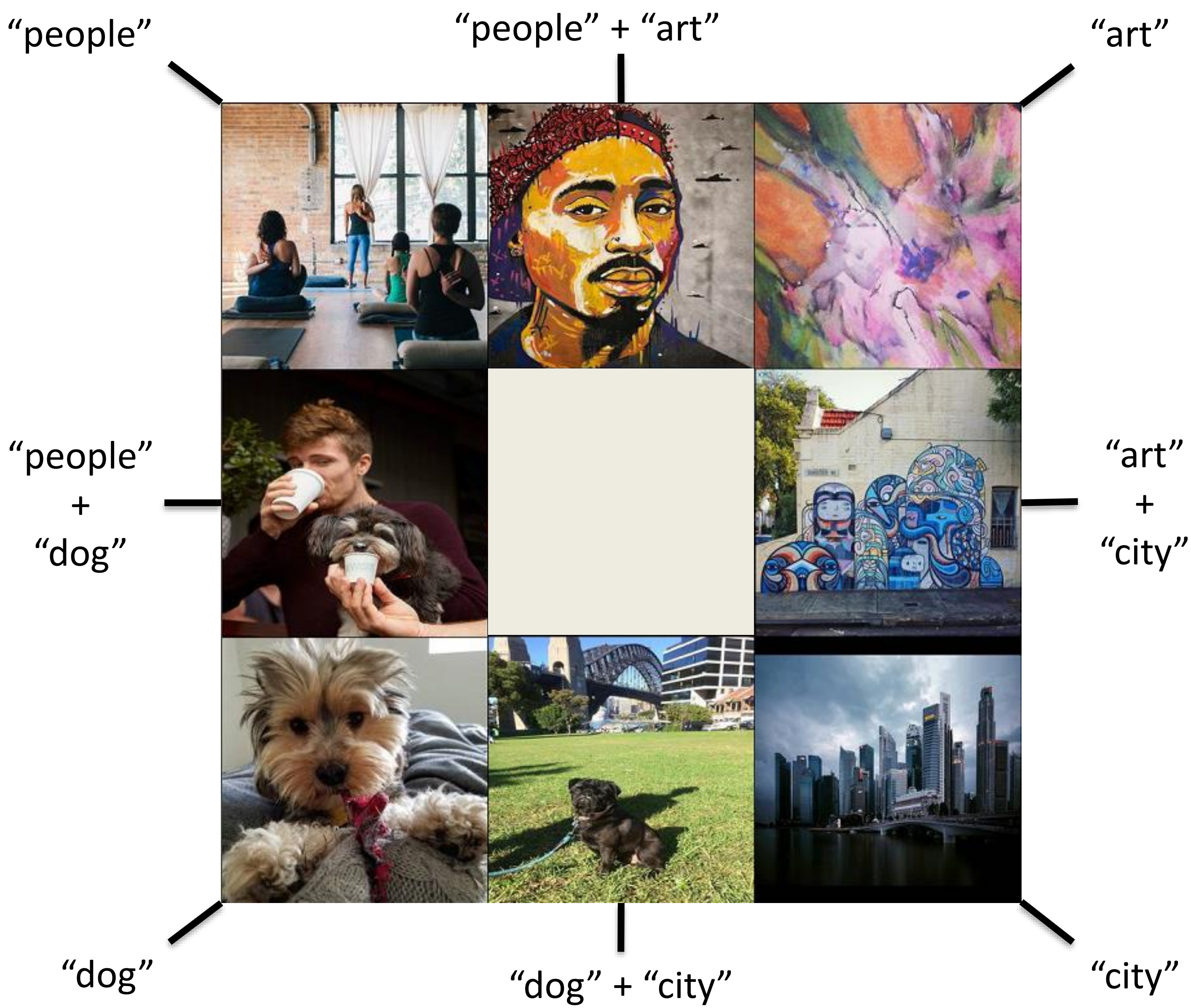
“haircut”

“haircut”
+
“man”

“haircut”
+
“woman”

“haircut”
-
“man”

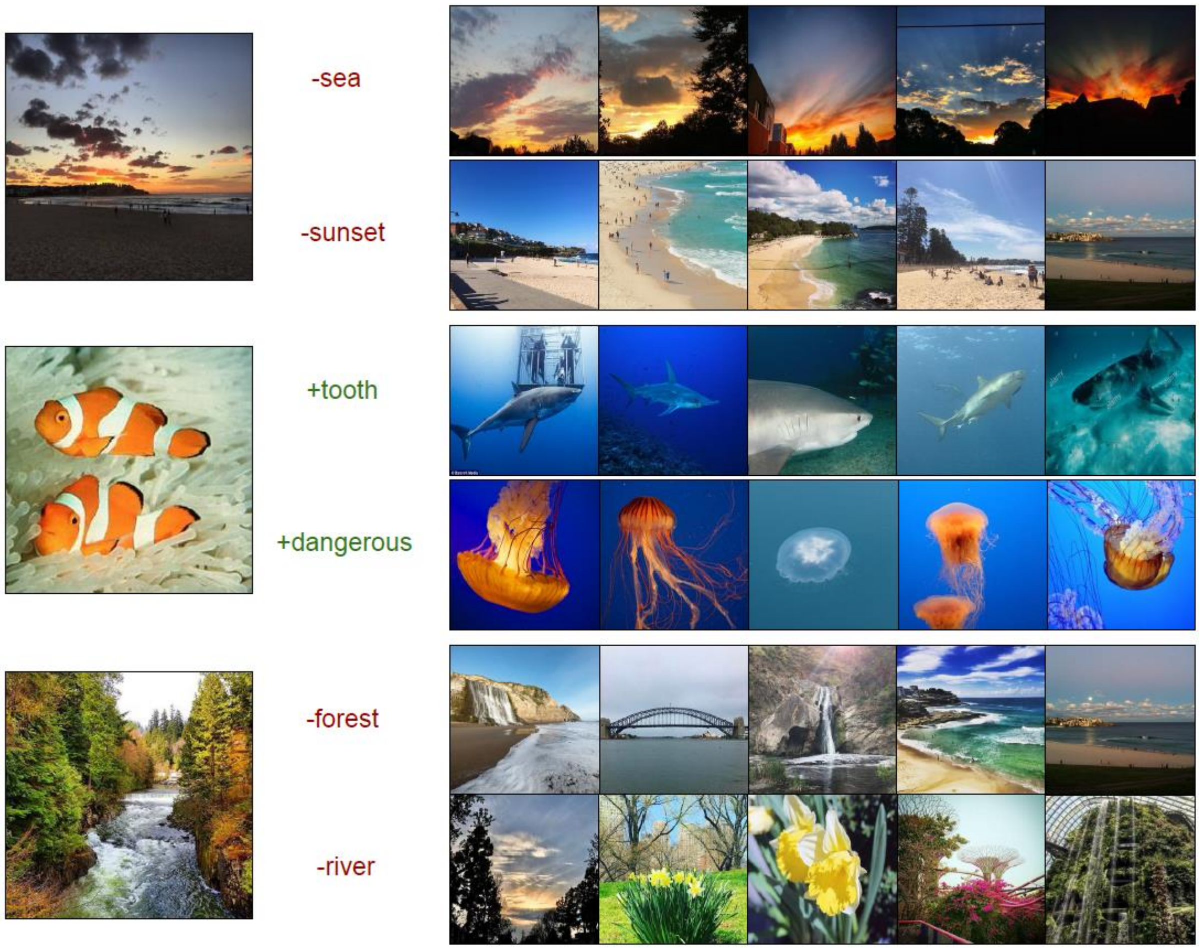



"people"
"people" + "art"
"art"
"people" + "dog"
"art" + "city"
"dog"
"dog" + "city"
"city"

-wedding

+animal

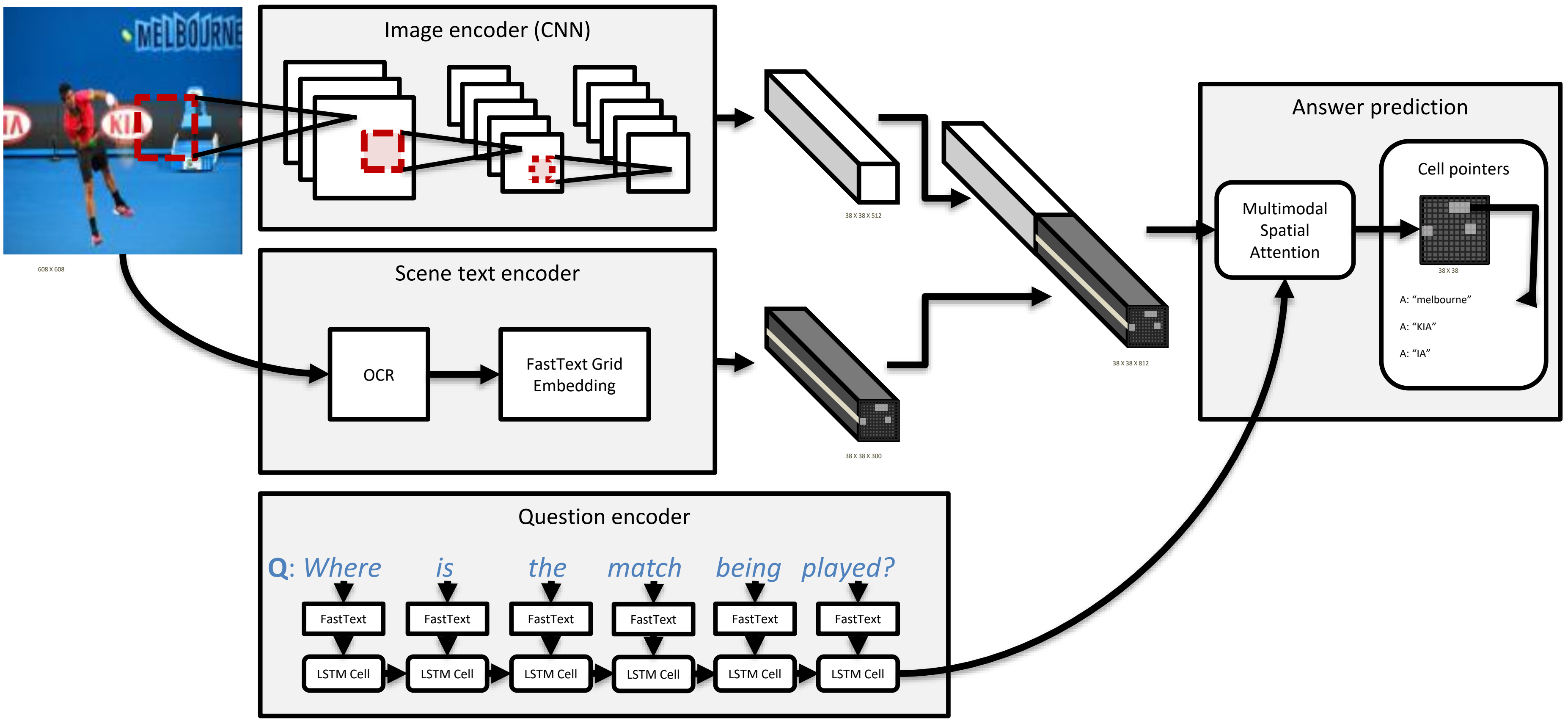

-sea
-sunset
+tooth
+dangerous
-forest
-river

Scene Text Visual Question Answering

Question	What is written on the blue shirt the boy is wearing?	What temperature does the air conditioner show?
Answer	I Think Somebody Needs A Tickle	24C
Reasoning Capacities	Object recognition, Action recognition	Prior world knowledge

A. Biten, R. Tito, A. Mafla, L. Gomez, M. Rusiñol, E. Valveny, C.V. Jawahar, D. Karatzas, "Scene Text Visual Question Answering", ICCV 2019

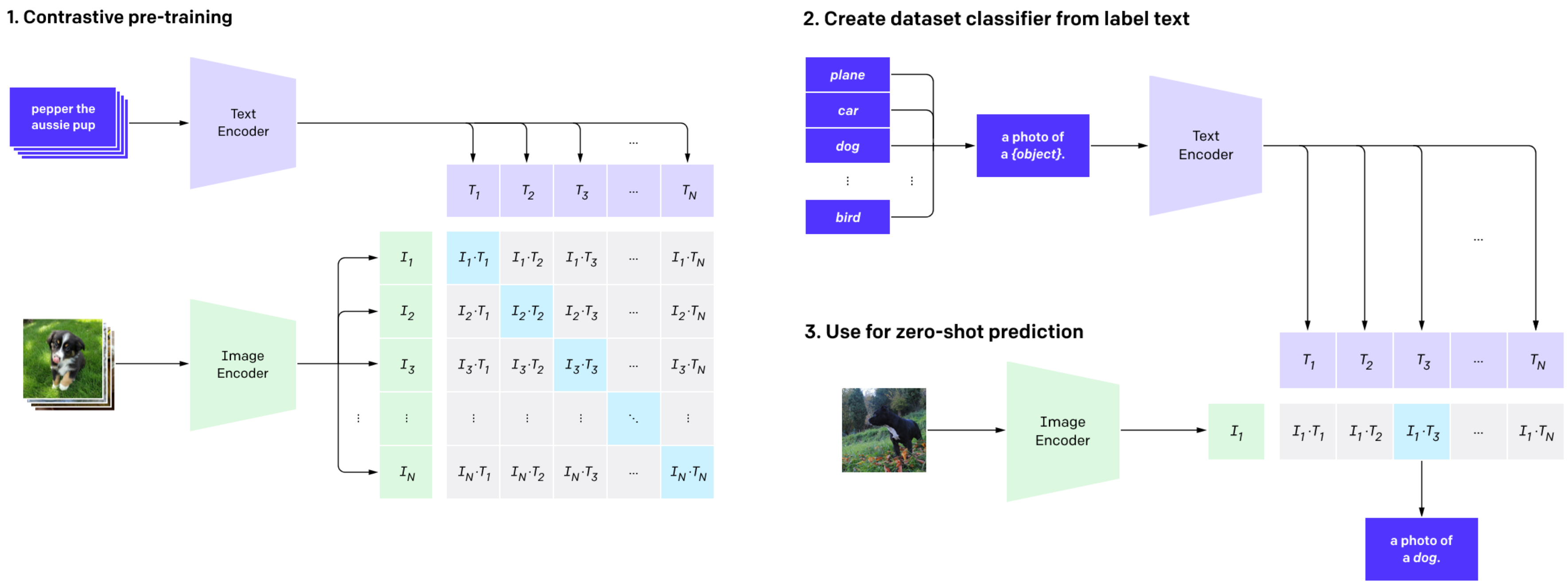
Visual Question Answering



L. Gomez, A. Biten, R. Tito, A. Mafla, D. Karatzas, "Multimodal grid features and cell pointers for Scene Text Visual Question Answering", 2021

CLIP: Contrastive Language-Image Pre-training

"Our method uses an abundantly available source of supervision: the text paired with images found across the internet"



"our best performing CLIP model trains on 256 GPUs for 2 weeks"

A. Radford et al, "[Learning Transferable Visual Models From Natural Language Supervision]", 2021