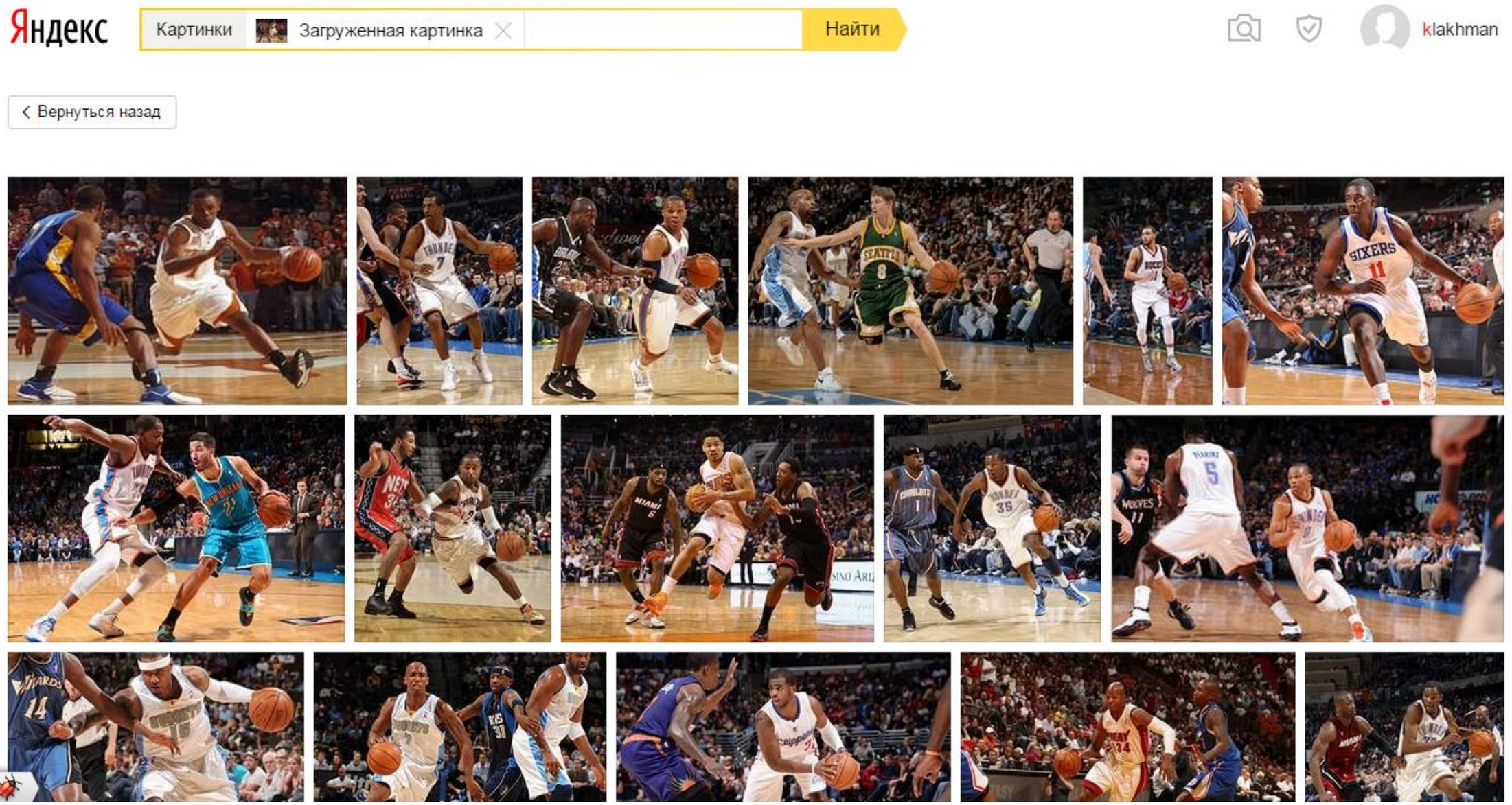
Yandex

Visual Search

Konstantin Lakhman, *Head of Computer Vision and ML-Platforms Department*

klakhman@yandex-team.ru

Similar images search



Similar images search: applications

- What/who is this?

Similar images search: applications

- What/who is this?

- Similar products search: clothes, interior, …
- Patterns search
- Hairstyle search
- ….

Similar images search: key ingredients

Similar images search: features

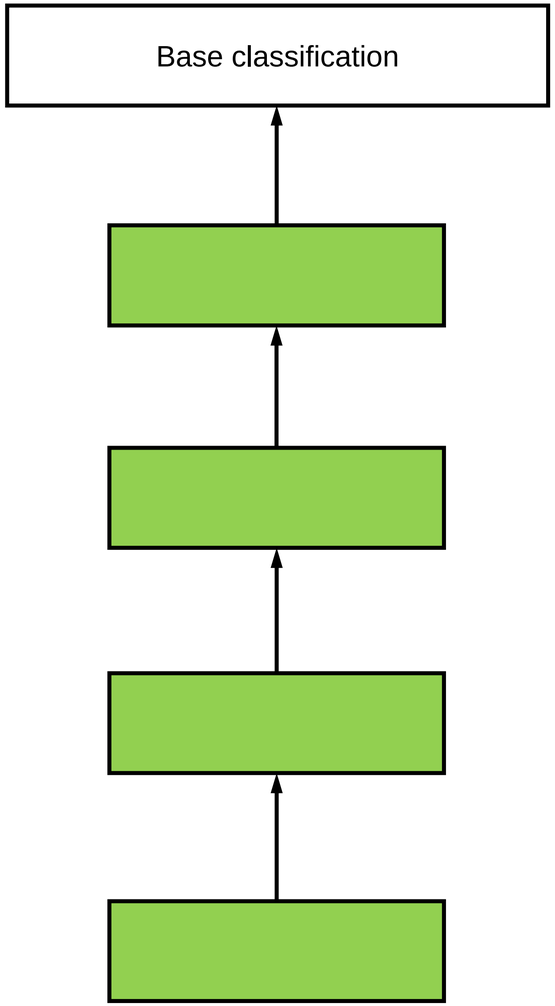
$$F(image) \rightarrow x \in \mathbb{R}^n$$

Similar images search: features

Similar images search: features space

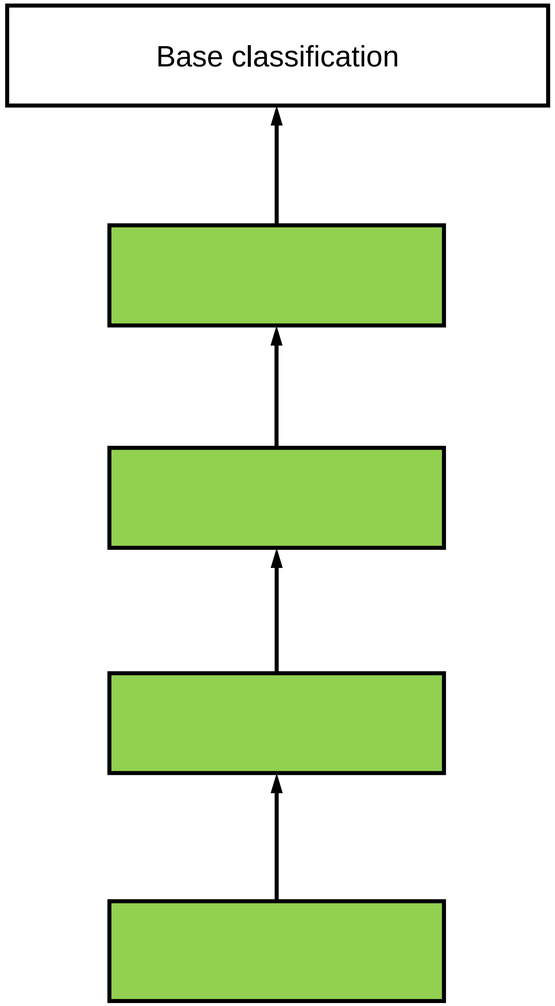
Similar images search: nearest neighbors

How to obtain image understanding?



Russakovsky et al., 2015.

Massive image classification datasets

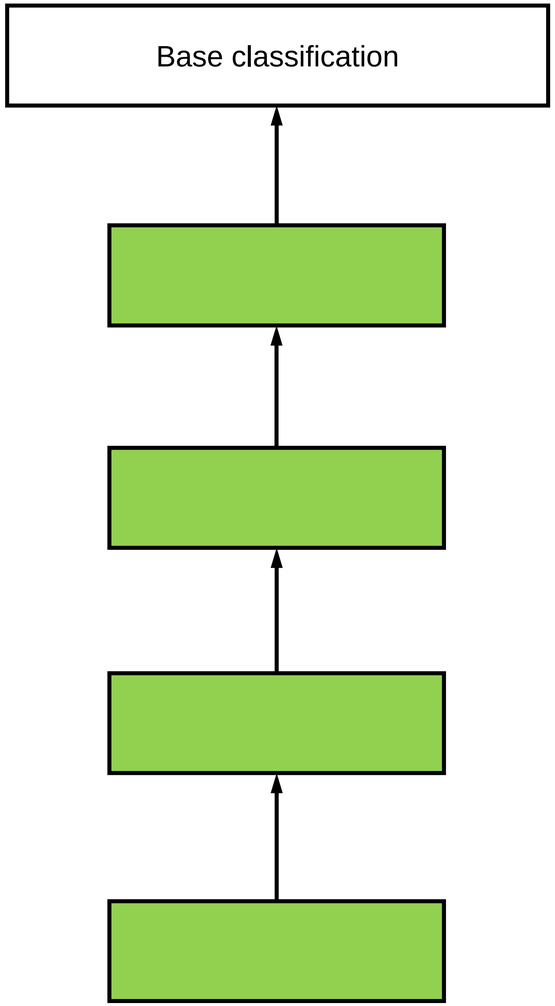


› ImageNet 1k classes

› ImageNet ~10k classes

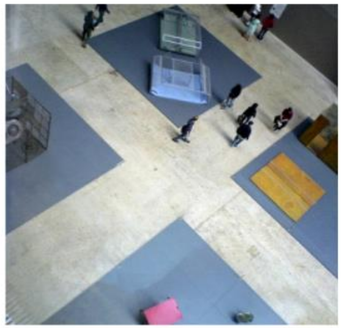
› Yahoo Flickr Creative Commons 100M
“Dataset consists of 100 million Flickr user-uploaded images and videos along with their corresponding metadata including title, description, camera type, tags, and geotags when available.”
Number of user annotated images: ~60 mln.
Number of classes: from 1k to 100k (depending on filtering and tags extraction method)

Mining datasets from weakly labeled data



the veranda hotel portixol palma

plane approaching zrh avro regional jet rj



not as impressive as embankment that s for sure

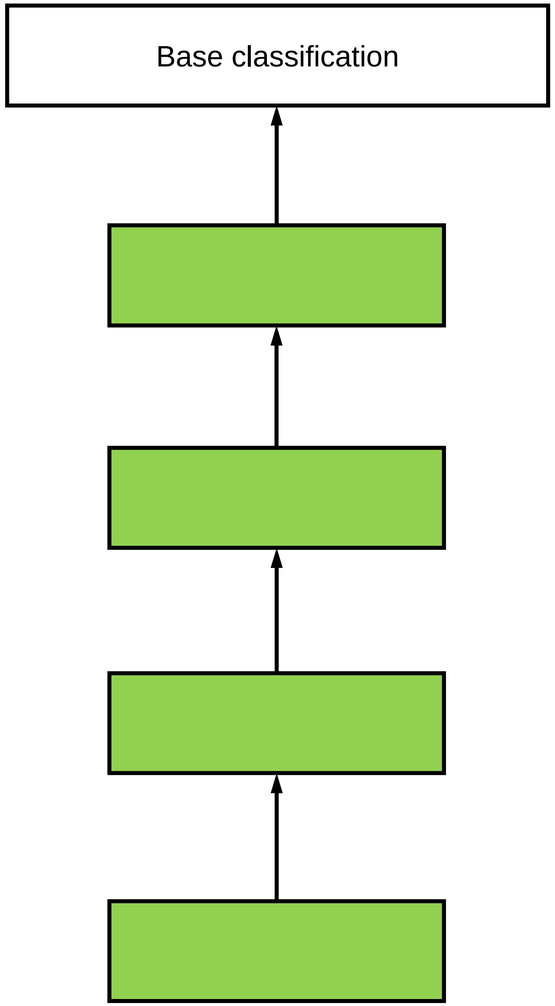
student housing by lungaard tranberg architects in copenhagen click here to see where this photo was taken

article in the local paper about all the unusual things found at otto s home

this was another one with my old digital camera i like the way it looks for some things though slow and lower resolution than new cameras another problem is that it s a bit of a brick to carry and is a pain unless you re carrying a bag with some room it s nearly x x and weighs ounces new one is x x and weighs ounces i underexposed this one a bit did exposure bracketing script underexposure on that camera looks melty yummy gold kodak film like

Joulinet al., *2015*

Mining datasets from weakly labeled data

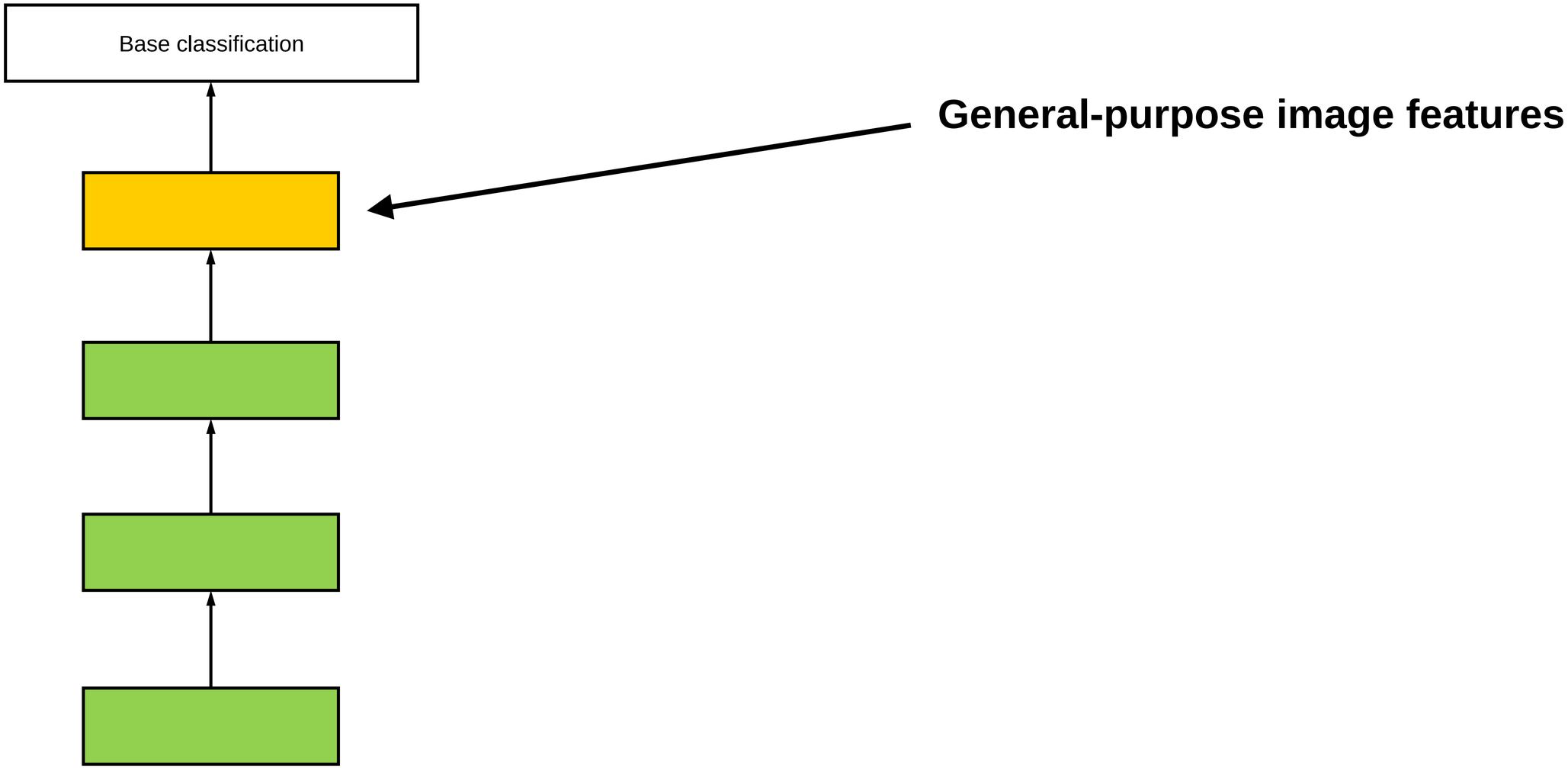


the veranda hotel
portixol palma

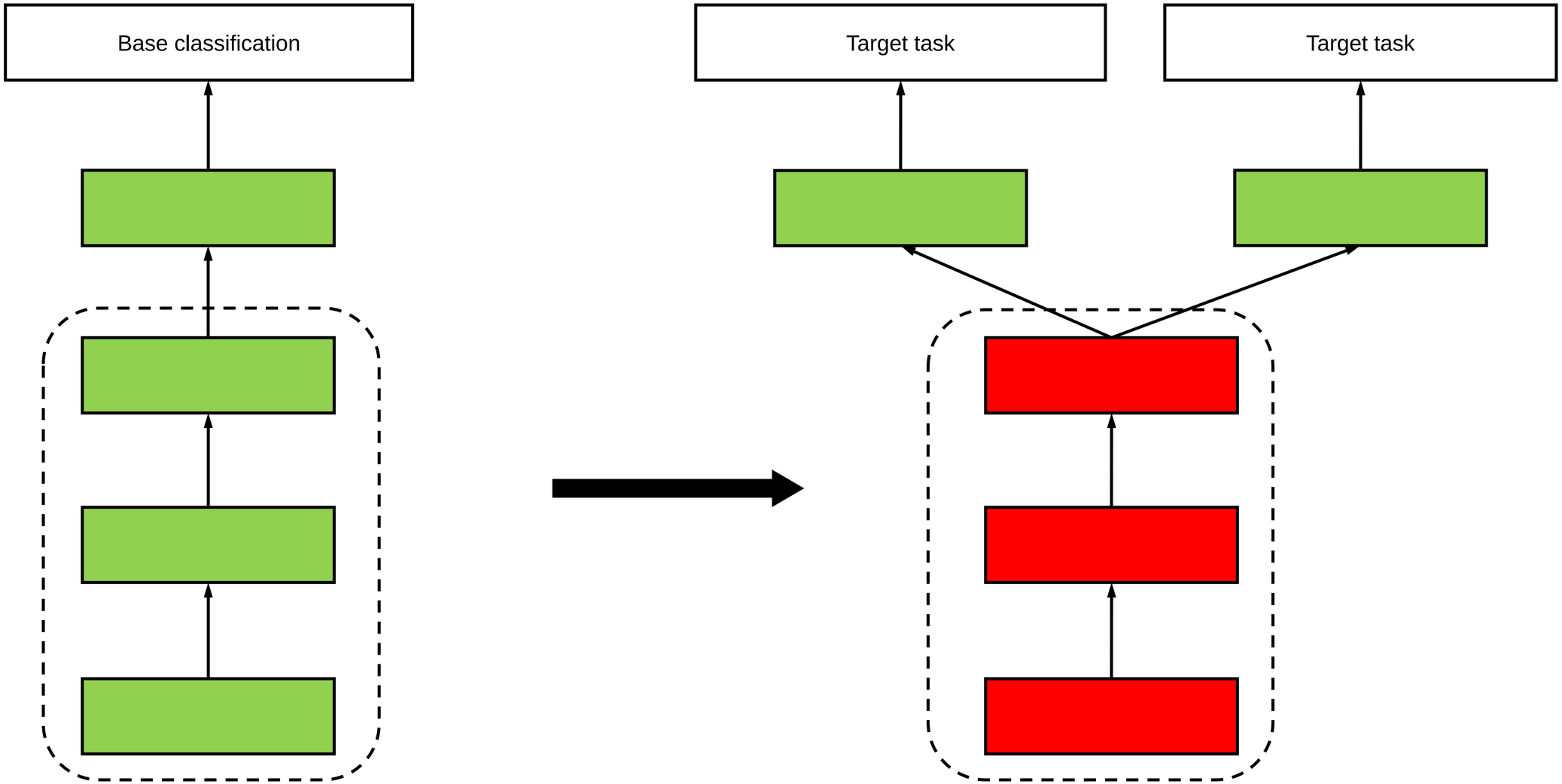
plane approaching zrh
avro regional jet rj

Joulinet al., *2015*

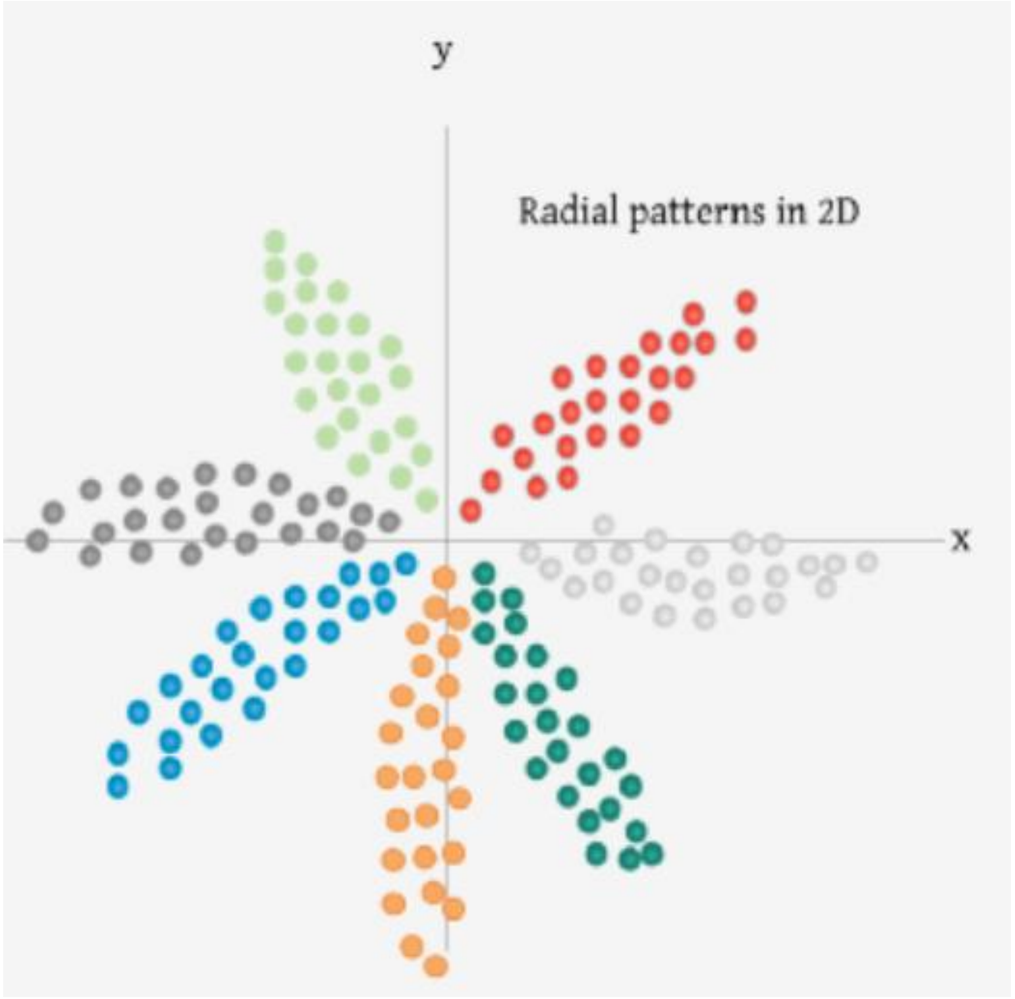
Good image features is key ingredient



Transfer learning aka multi-head networks

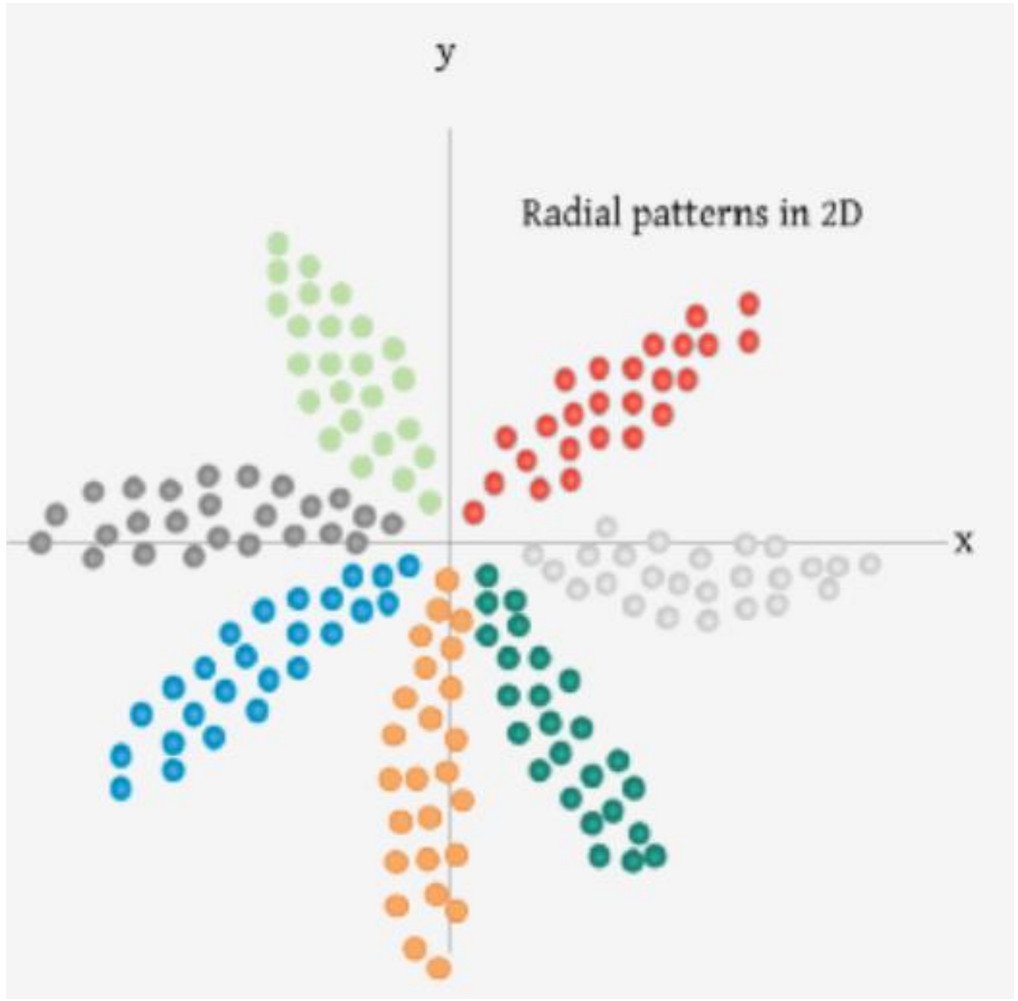


What's wrong with discriminative features?



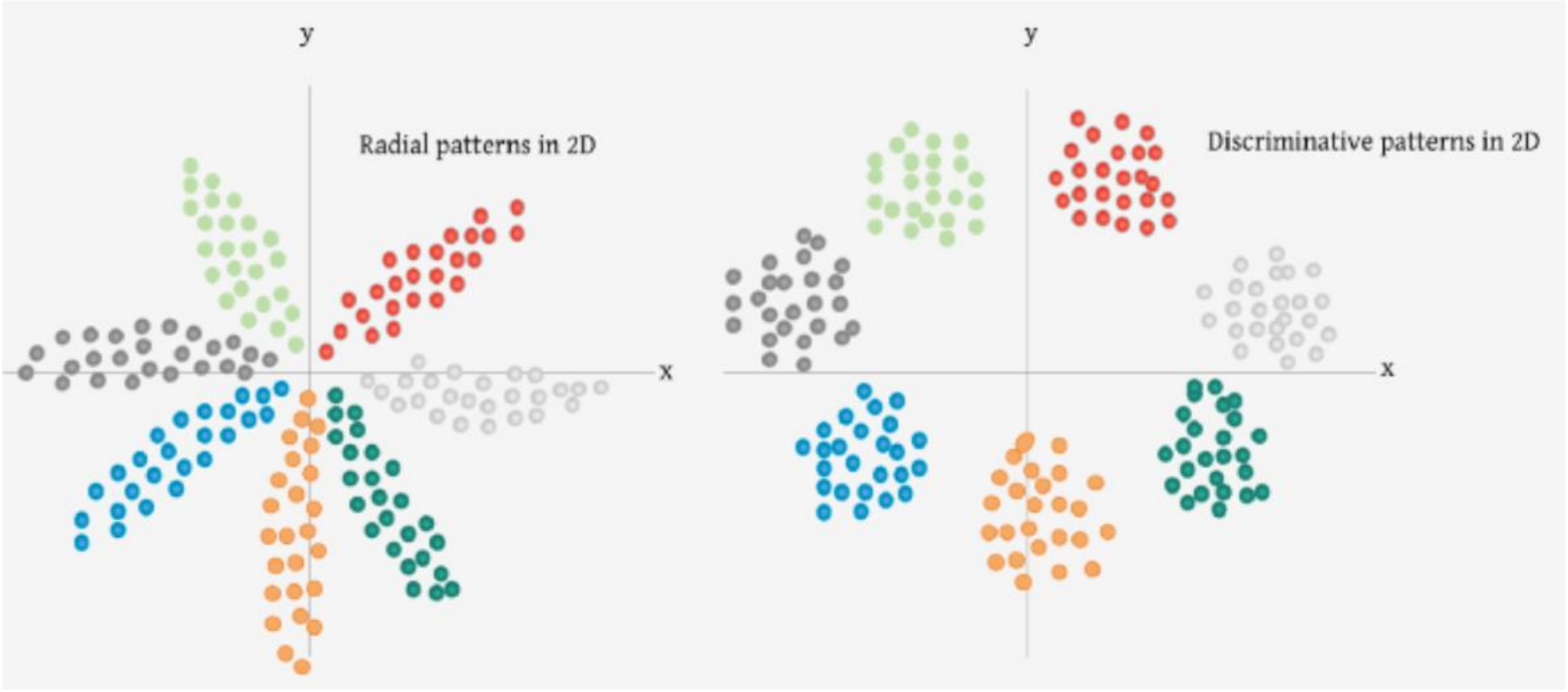
$$\sigma_i(x) = \frac{e^{w_i^T x}}{\sum e^{w_j^T x}}$$

What's wrong with discriminative features?

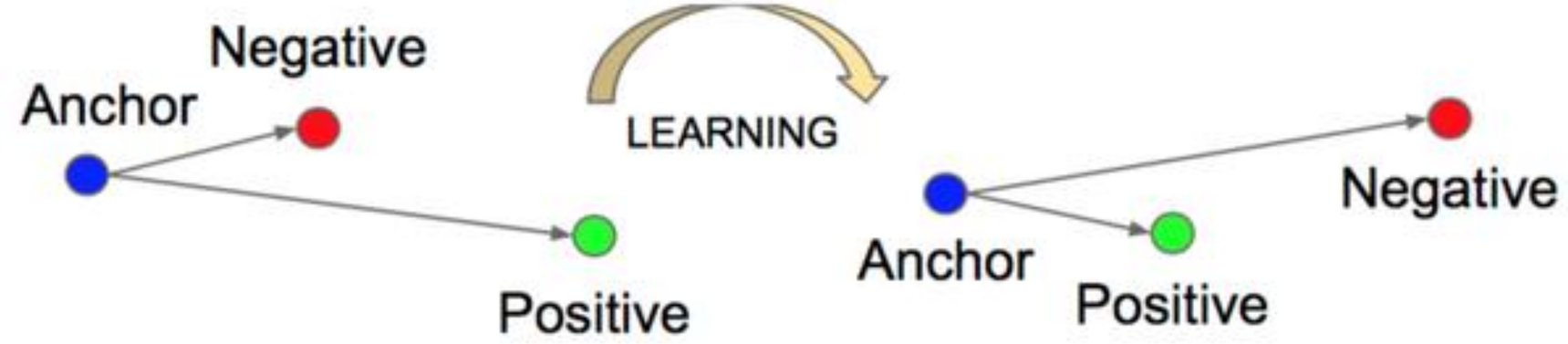


$$\sigma_i(x) = \frac{e^{w_i^T x}}{\sum e^{w_j^T x}}$$

What's wrong with discriminative features?



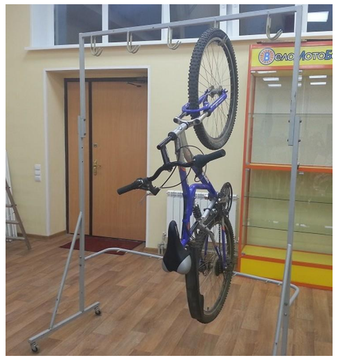
Finetune with metric learning is important piece



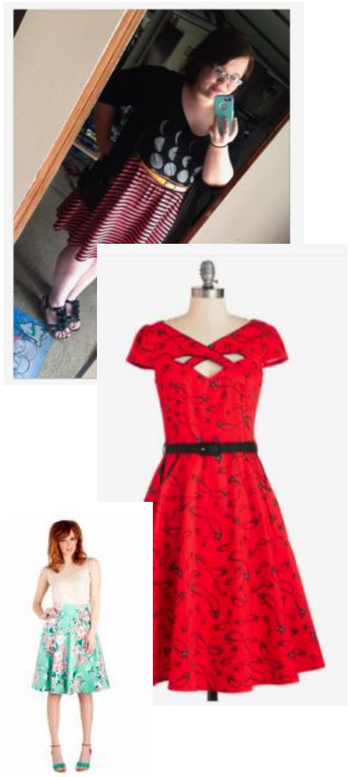
$$L = \max(0, T - F(anchor) * F(positive) + F(anchor) * F(negative)$$

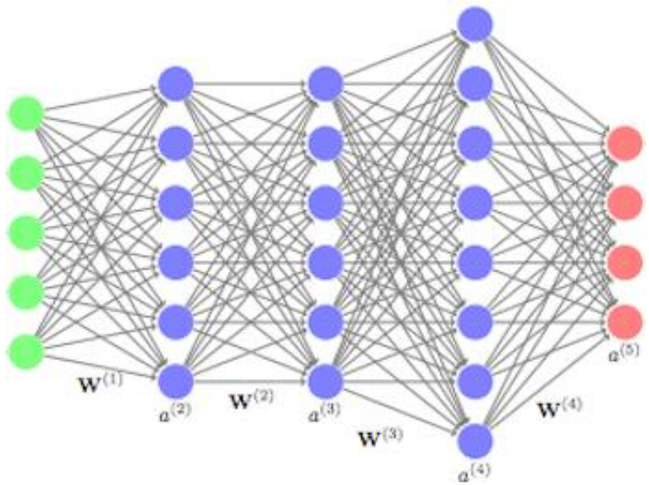
Domain shift is a problem

Domain shift is a problem



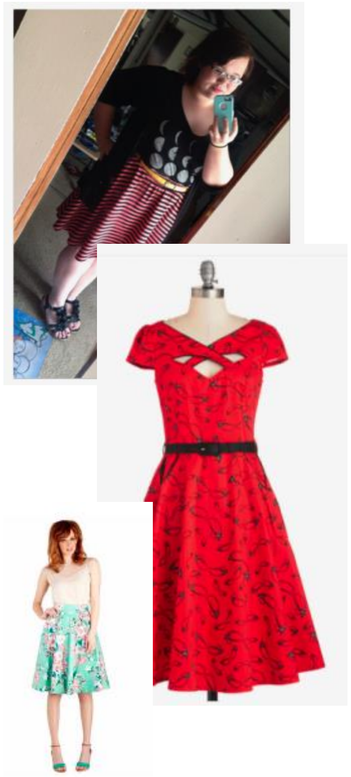
How to mitigate domain shift: algorithms

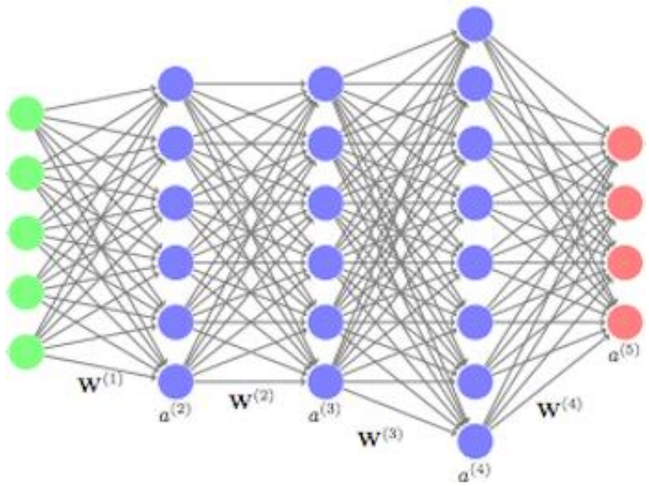




Task specific loss

How to mitigate domain shift: algorithms





Task specific loss

W → "Real" or "internet" image?

"GAN style"

How to mitigate domain shift: datasets

street2shop

M. Hadi Kiapour et al. (2015) Where to Buy It:Matching Street Clothing Photos in Online Shops

The end

The end

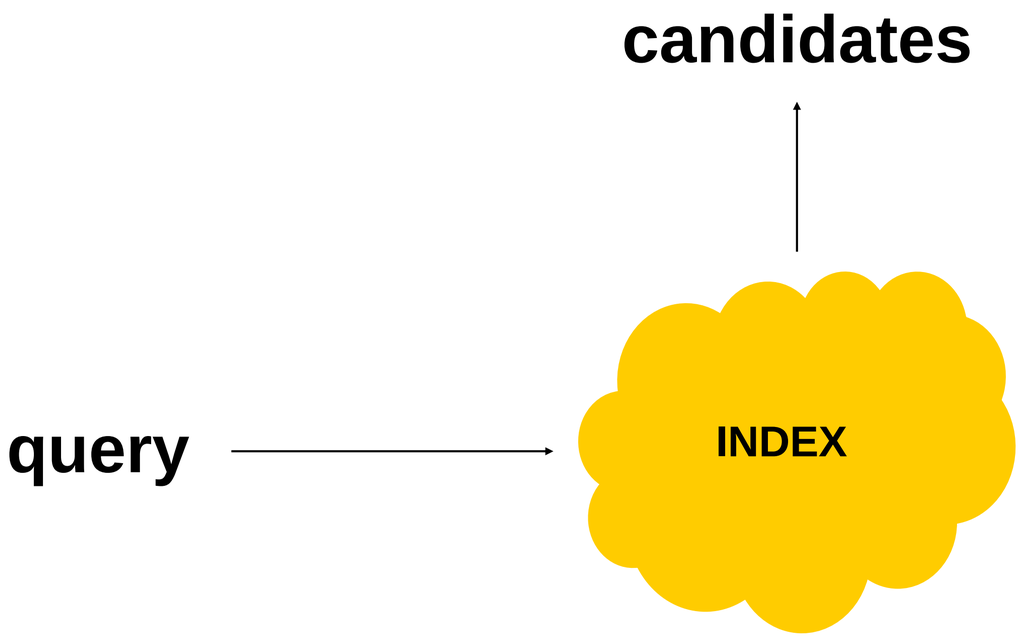
Really?

~~The end~~

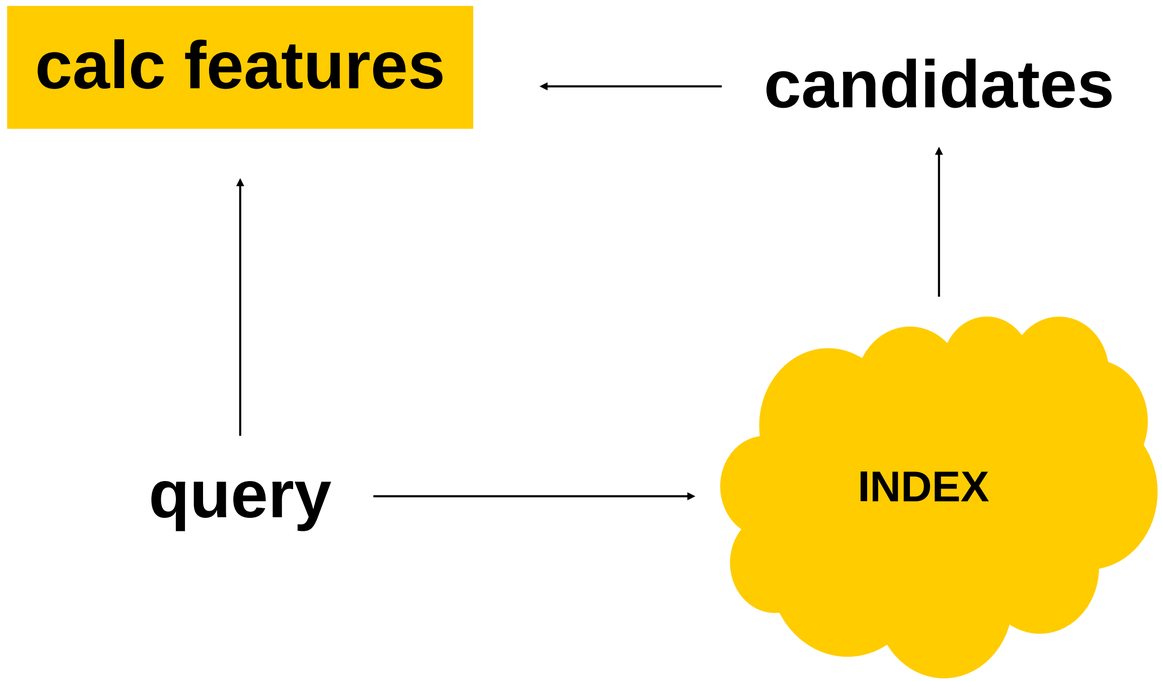
Not at all!
This type of visual similarity is only one factor.

Visual search is a place where **Computer Vision** meets **Search Engine** technologies

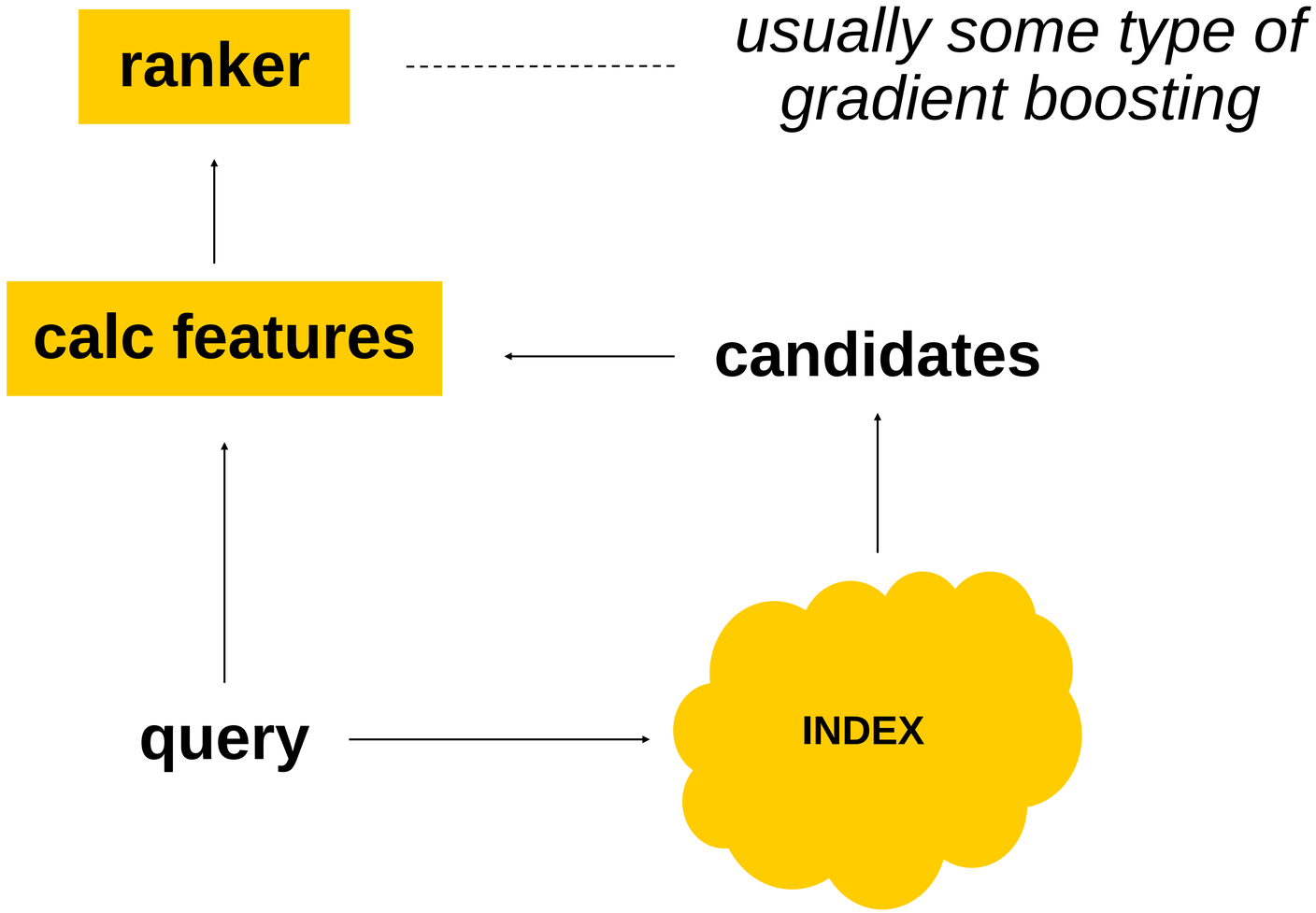
Search architecture



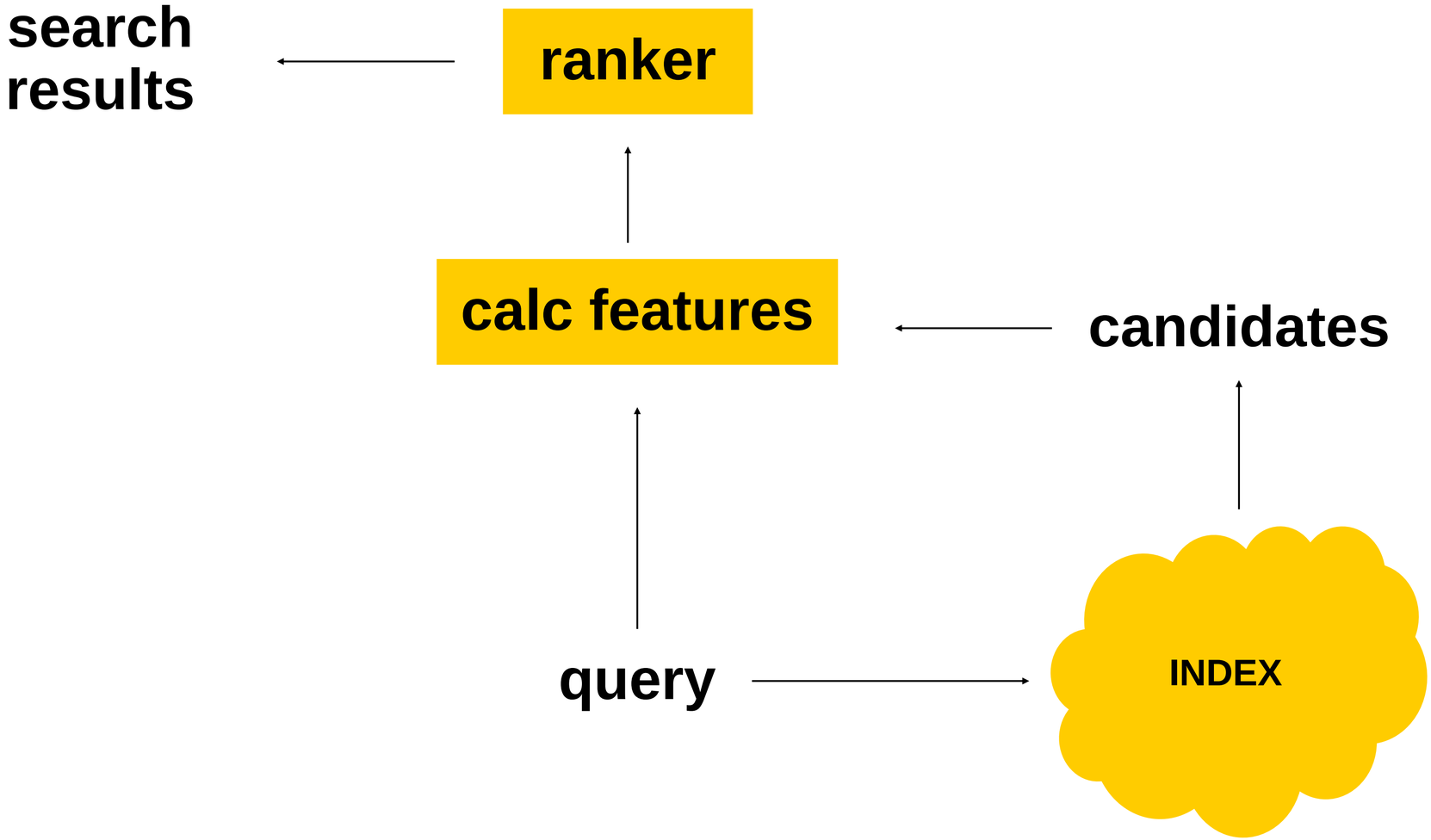
Search architecture



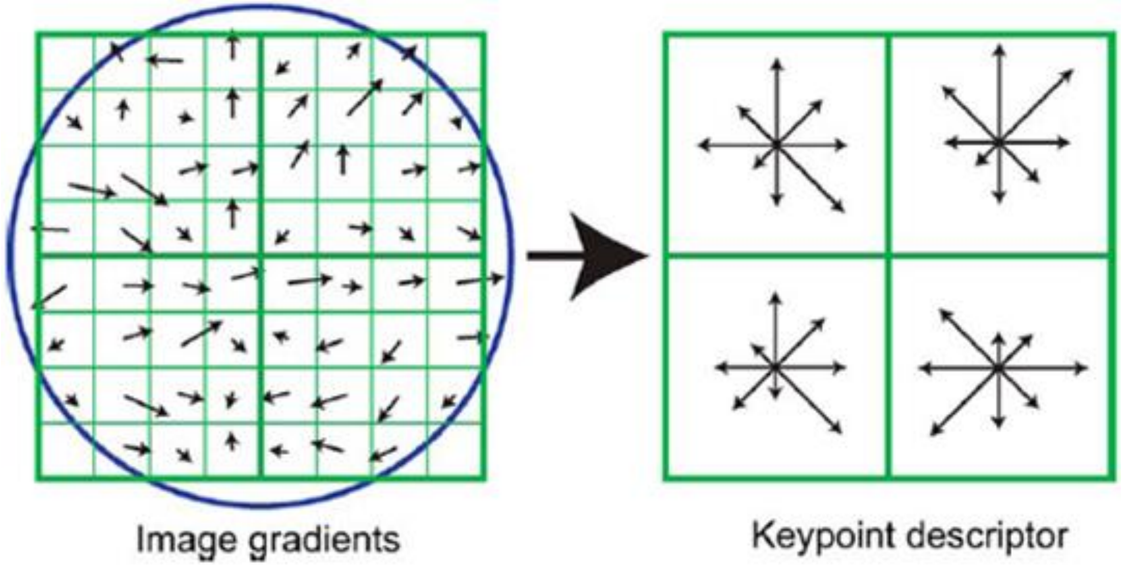
Search architecture



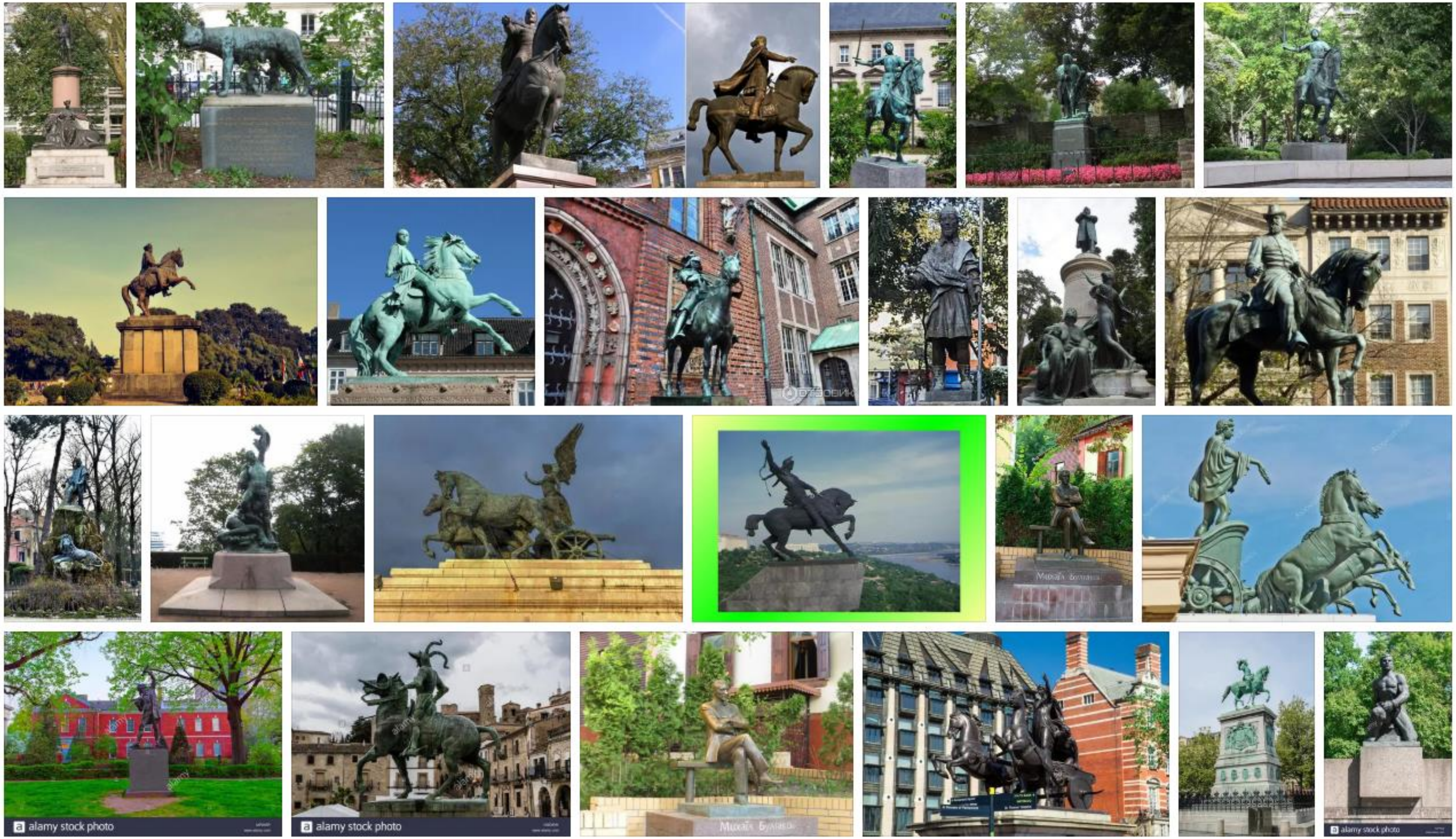
Search architecture



Local descriptors are still helpful in 2021



Local descriptors are still helpful in 2021



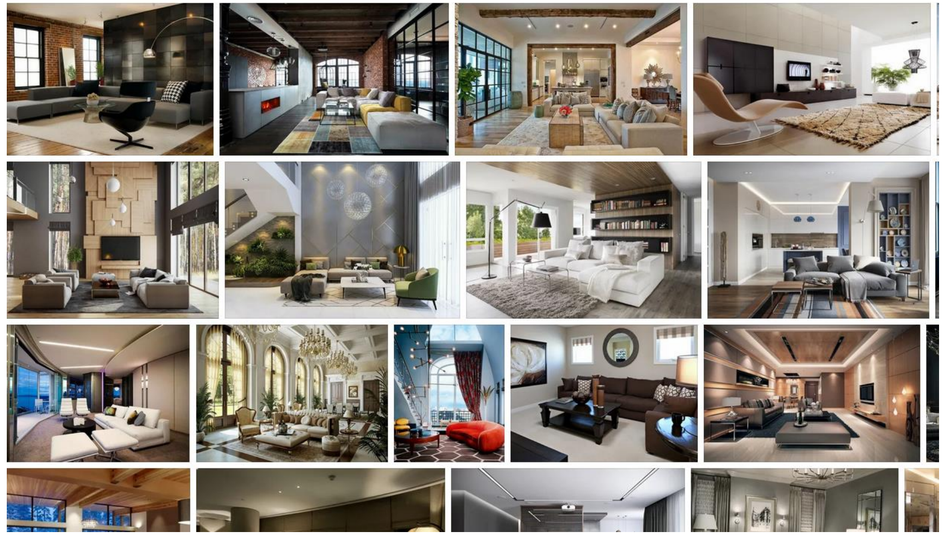
Local descriptors are still helpful in 2021

Specialized features

Faces

Fashion

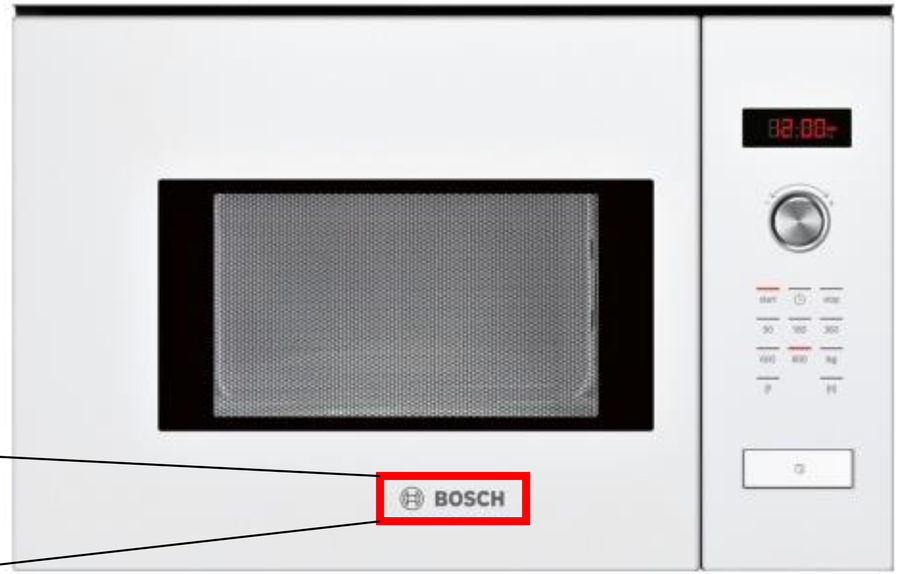
Interior



OCR

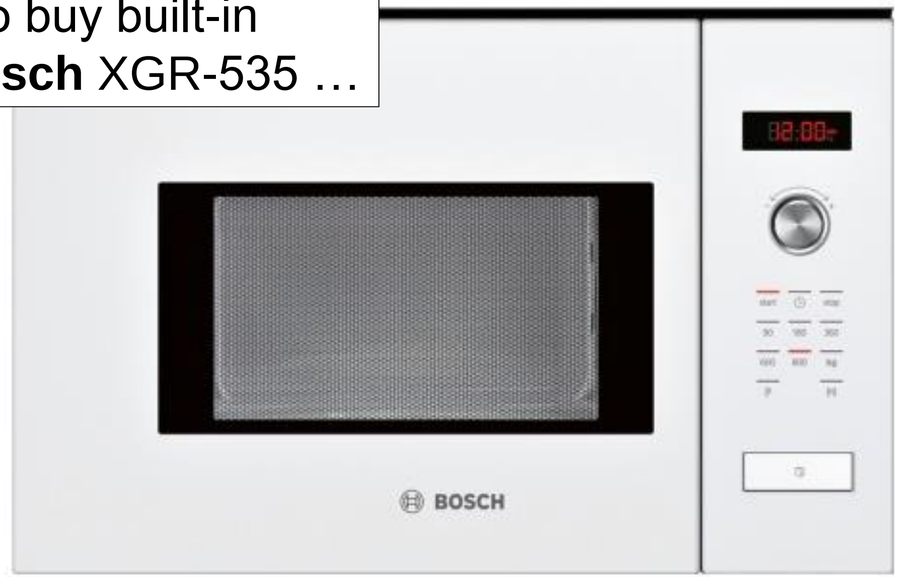
OCR

OCR



OCR vs linked text

… click here to buy built-in microwave **Bosch** XGR-535 …

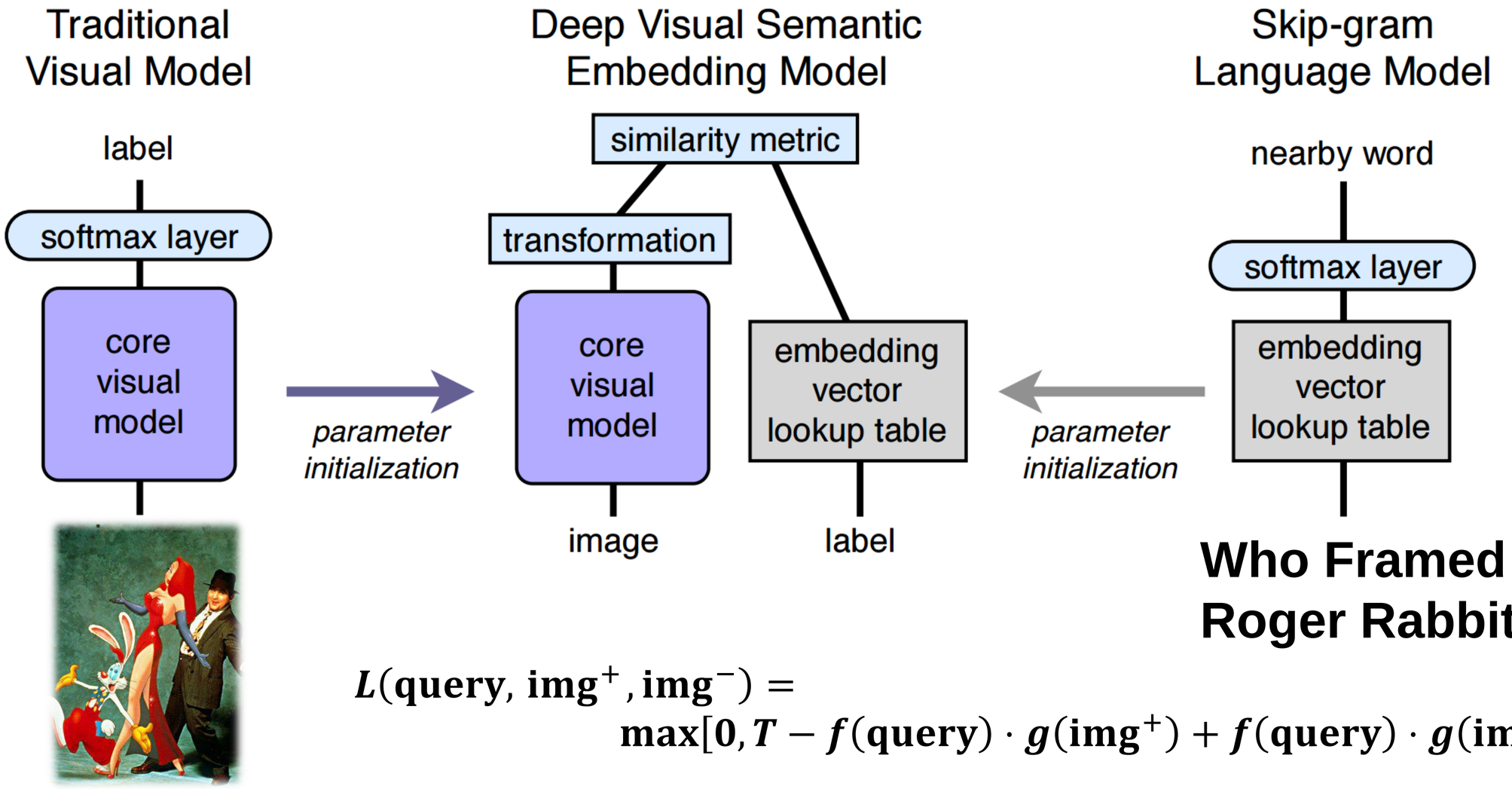


… **Hansa** is European manufacture …

Can we directly compare linked text with image?

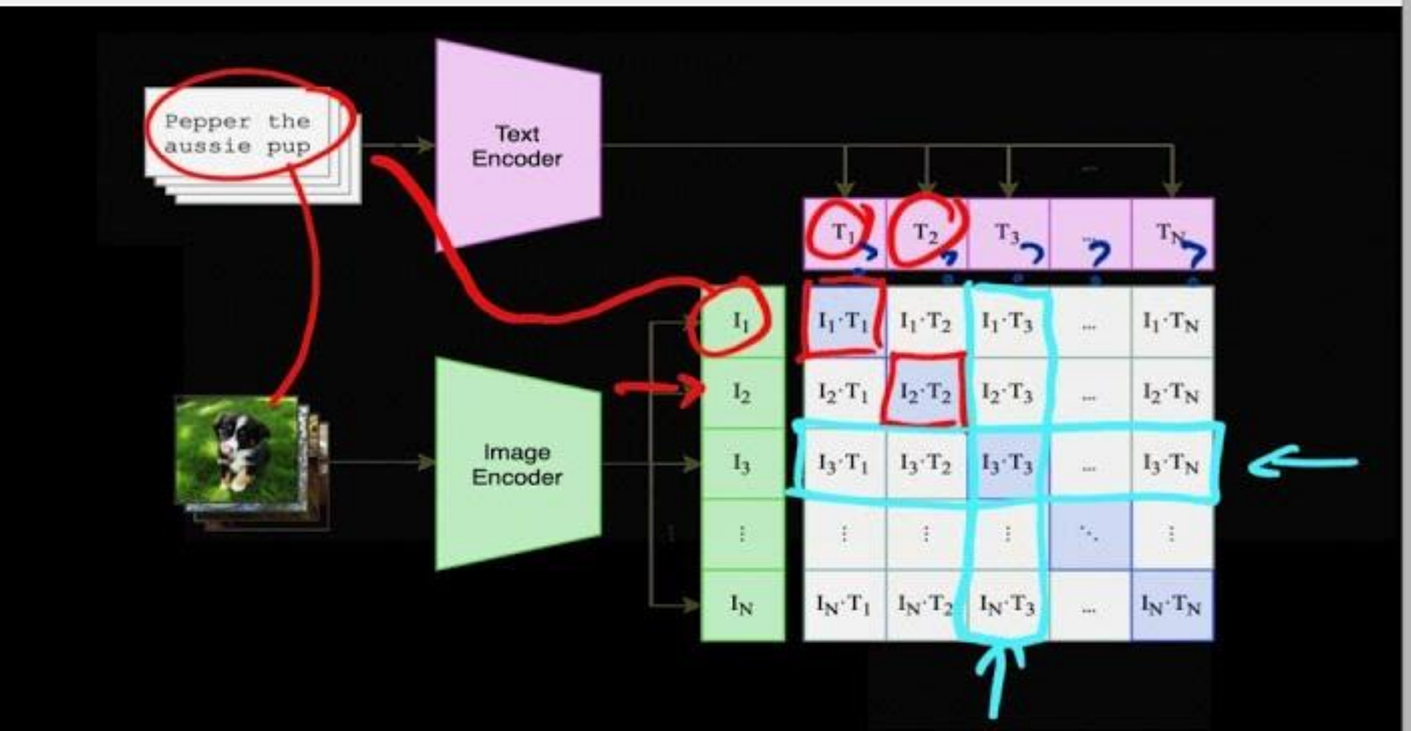
Who Framed Roger Rabbit?

Building joint semantic space



$$L(\mathbf{query}, \mathbf{img}^+, \mathbf{img}^-) = \max[0, T - f(\mathbf{query}) \cdot g(\mathbf{img}^+) + f(\mathbf{query}) \cdot g(\mathbf{img}^-)]$$

Frome A., Corrado G.S. et al., NIPS'2013



OpenAI's CLIP Connecting Text and Images

Numerous classifiers

Auto Model

IsAutomobile

IsPorn

Plant

IsFashionLook

Gruesome

IsErotic

Architectural style

Beautifulness

Similar features
Joint semantic space
Local descriptors
Numerous classifiers
OCR
Faces features
Landmarks
Text factors
…..

The end

~~**The end**~~

Not yet…

Our ultimate goal is to detect and identify an object

Natural way to query/ask someone

Флаг Аргентины

Стал государственным в 1812 году.

ru.wikipedia.org

Брэд Питт

Американский актёр и кинопродюсер. Лауреат двух премий «Золотой глобус». Обладатель премии «Оскар» как один из продюсеров фильма «12 лет рабства» - победителя в категории «Лучший фильм» на церемонии 2014 года - и за лучшую мужскую роль второго плана в картине «Однажды в Голливуде». До этого пять раз номинировался на премию «Оскар».

ru.wikipedia.org

Мачу-Пикчу

Город древней Америки, находящийся на территории современного Перу, в 6 километрах от посёлка Агуас-Кальентес, на вершине горного хребта на высоте 2400 метров над уровнем моря, господствуя над долиной реки Урубамбы. В 2007 году удостоен звания Нового чуда света.

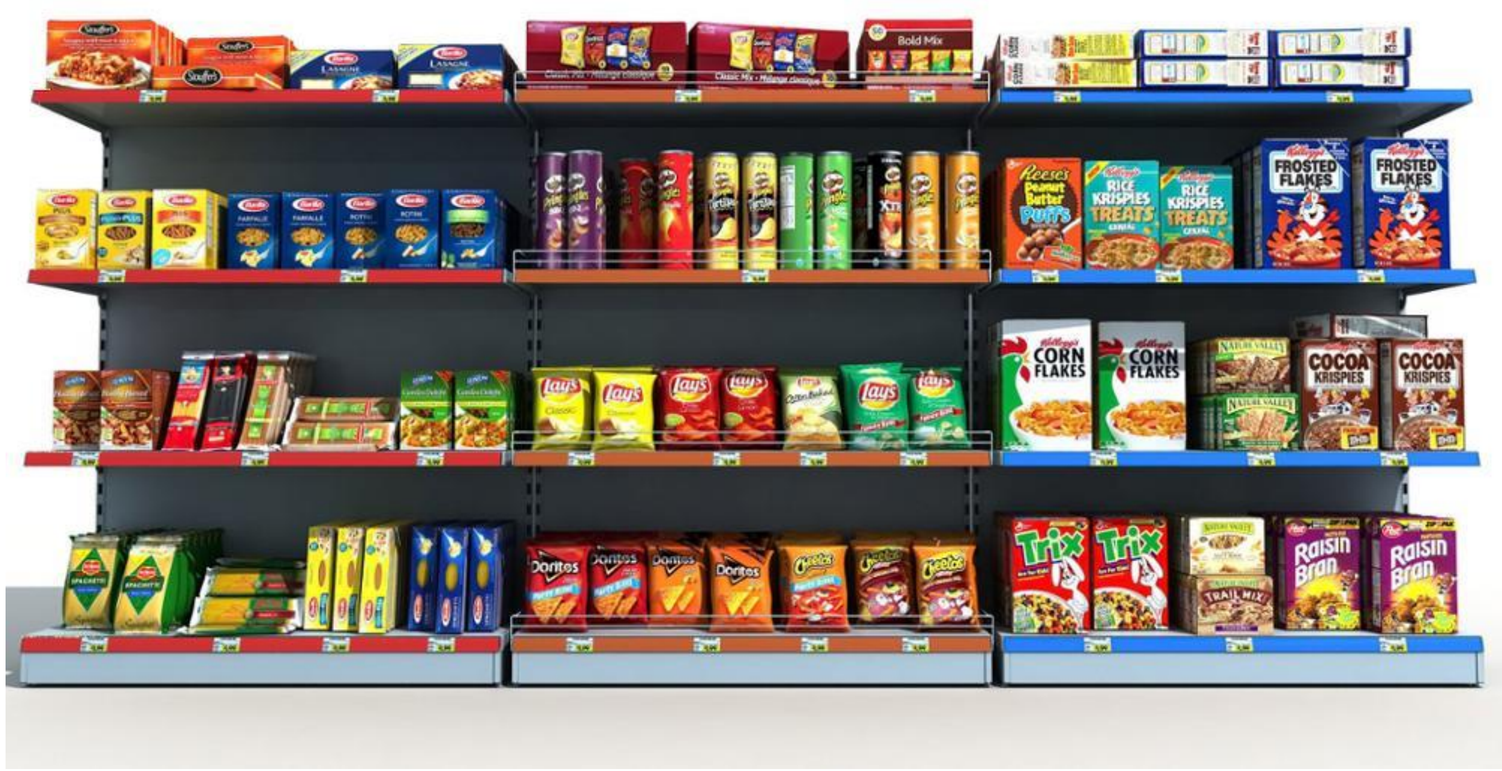
ru.wikipedia.org

Дендробиум

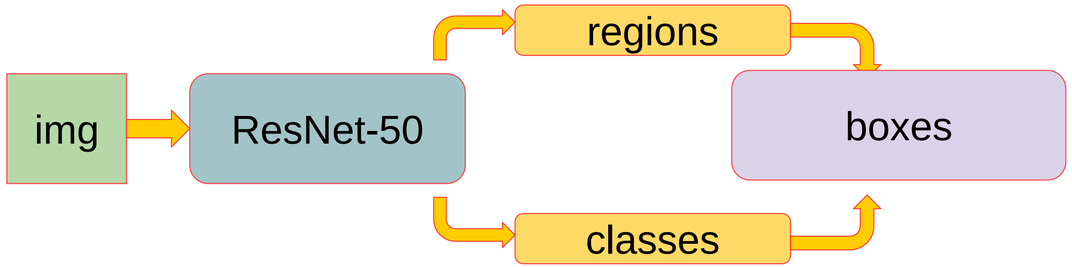
Род многолетних травянистых растений семейства Орхидные. Аббревиатура родового названия используемая в любительском и промышленном цветоводстве - Den.

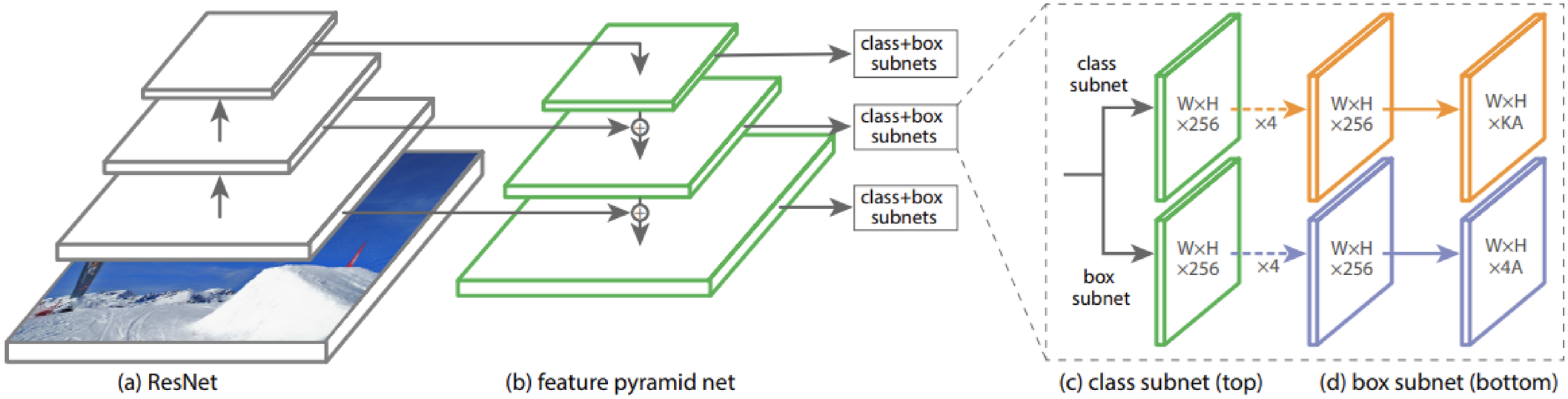
ru.wikipedia.org

Object detection as first stage



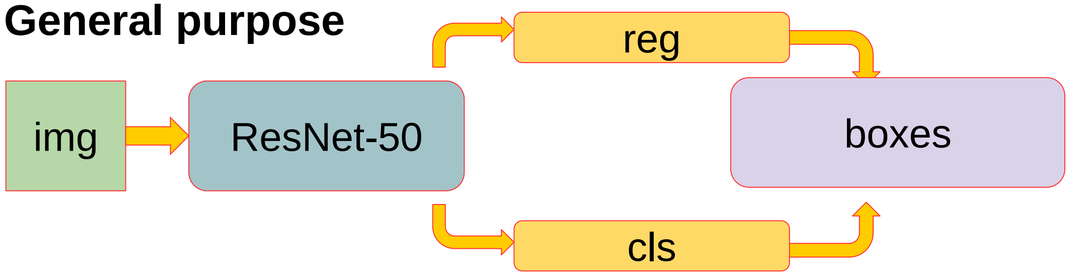
Detector


regions
img
ResNet-50
boxes
classes


class+box subnets
class+box subnets
class+box subnets
class subnet
box subnet
W×H ×256
×4
W×H ×256
W×H ×KA
W×H ×256
×4
W×H ×256
W×H ×4A
(a) ResNet
(b) feature pyramid net
(c) class subnet (top)
(d) box subnet (bottom)

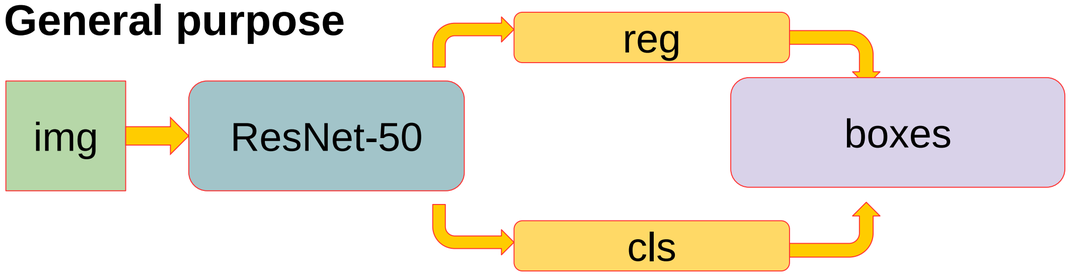
Detector’s zoo

General purpose

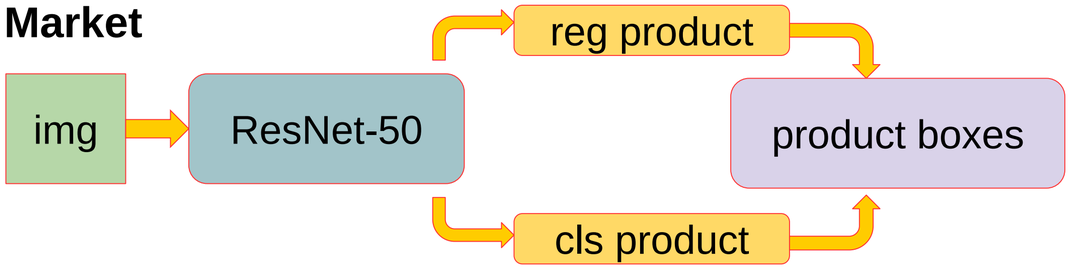


Detector’s zoo

General purpose

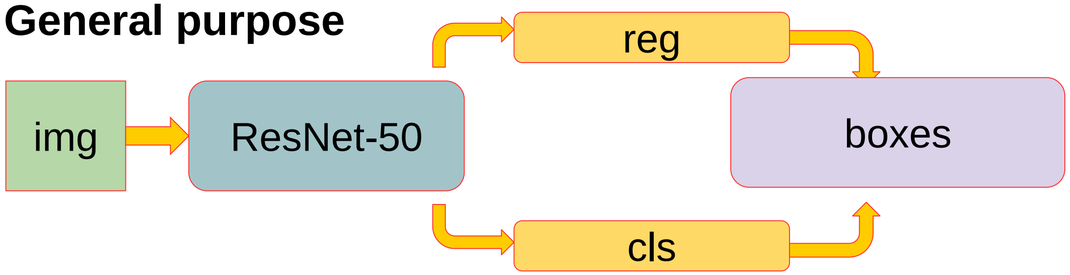


Market

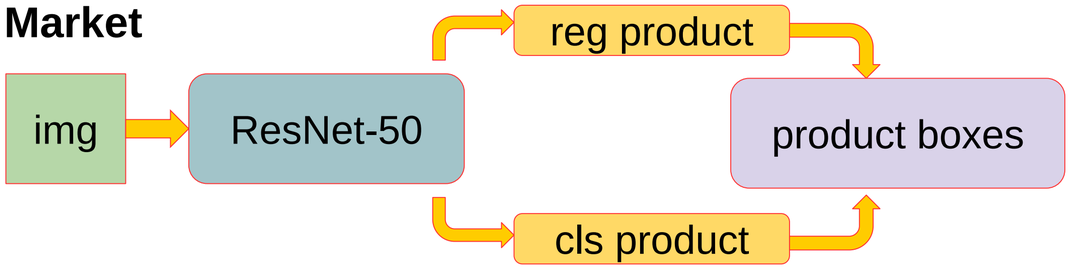


Detector’s zoo

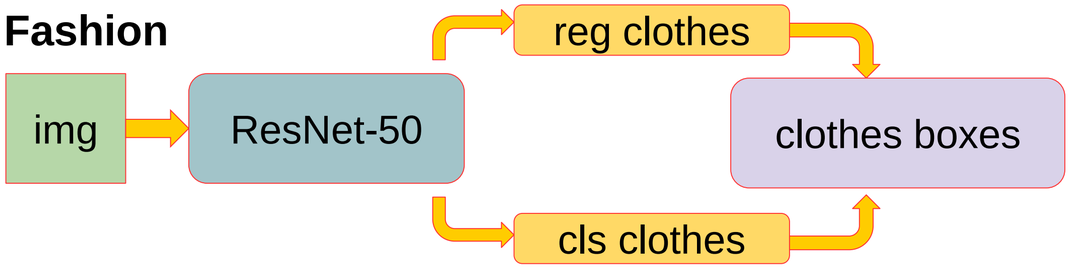
General purpose



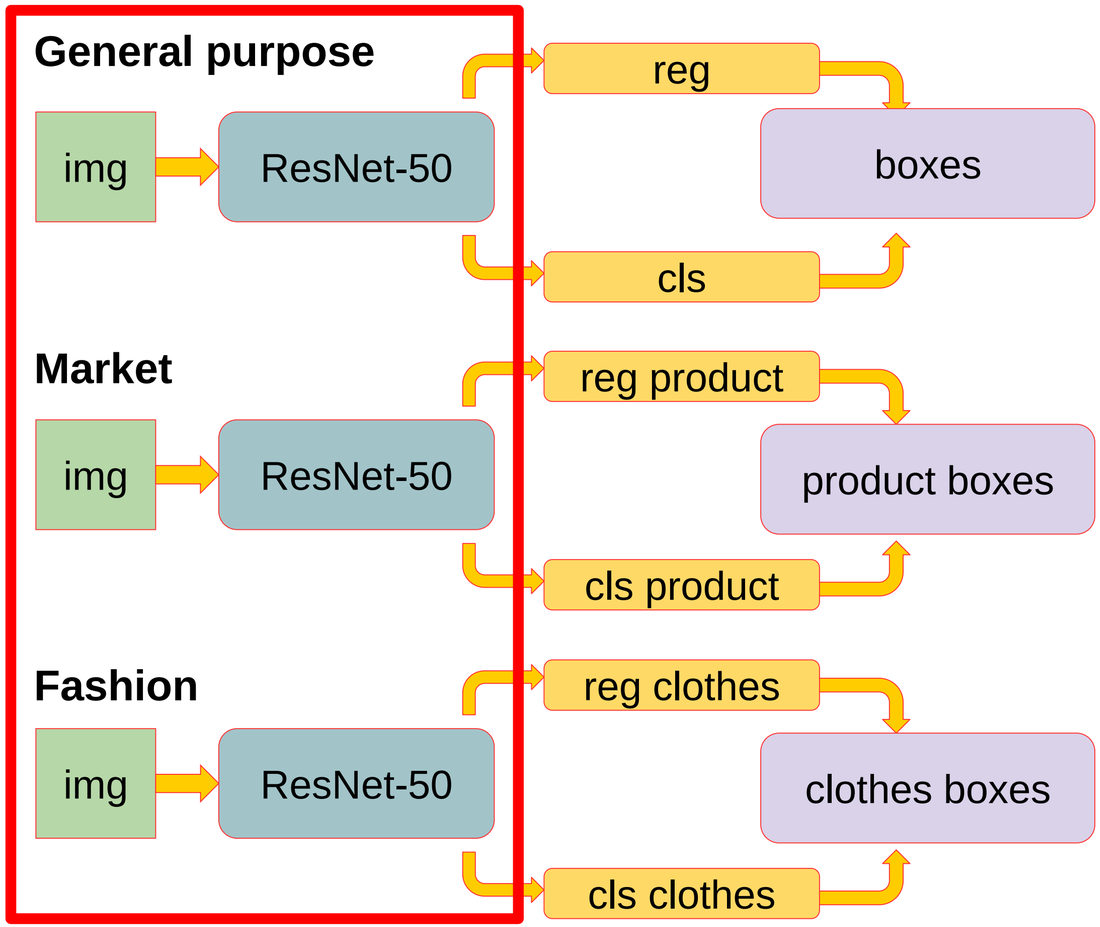
Market



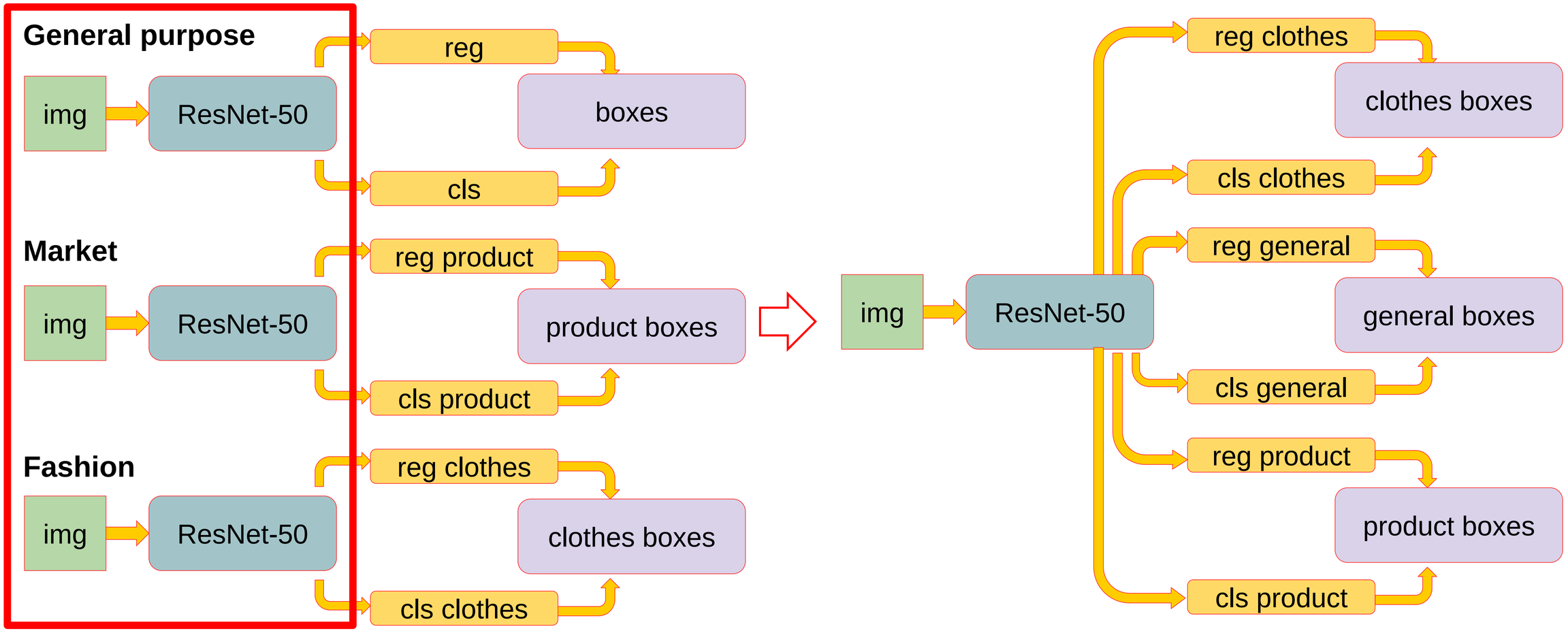
Fashion



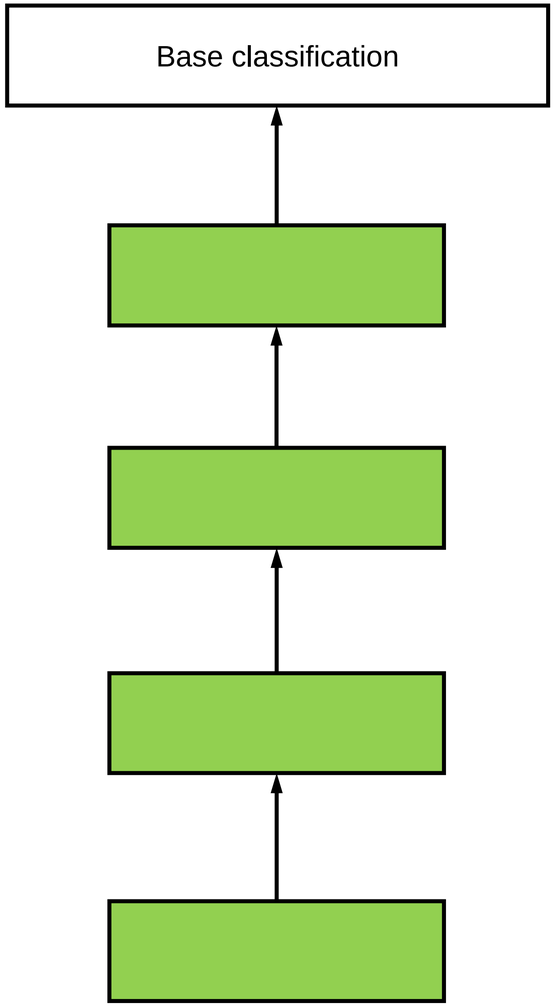
Detector’s zoo



Multi-head detector



How to tag image?

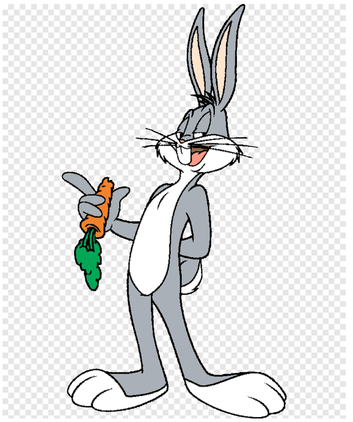


From similar images to tagging

From similar images to tagging

From similar images to tagging via linked text

…look at this lovely bugs bunny figure…



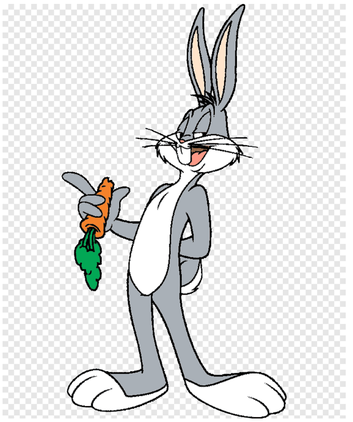
…looney tunes: bugs bunny…

…plush pinky rabbit toy…

…click to buy bugs bunny toy…

Tag aggregation

…look at this lovely bugs bunny figure…

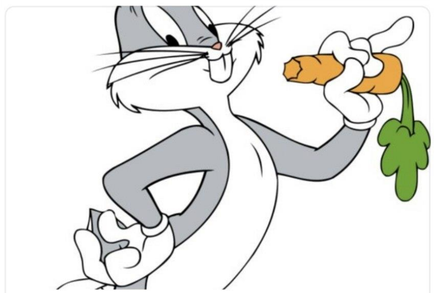


…looney tunes: bugs bunny…

…plush pinky rabbit toy…

…click to buy bugs bunny toy…

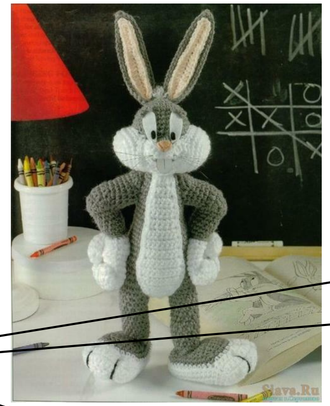
Object identification



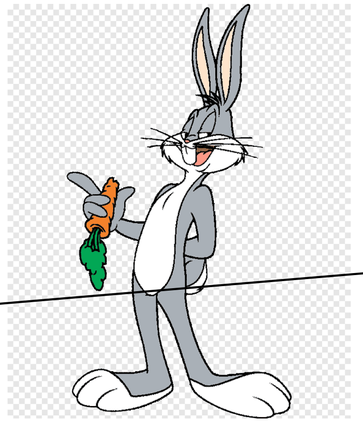
Багз Банни

Герой американских мультфильмов и комиксов; находчивый, бесстрашный и немного ехидный кролик. Из друзей его любят все, кроме Даффи. Создан творческим дуэтом режиссёра Текса Эйвери и аниматора Боба Гивенса на студии Warner Brothers, хотя общественное мнение до сих пор приписывает авторство Чаку Джонсу. В настоящий момент Багз Банни является эмблемой компании, особенно в области анимационной продукции. Согласно его биографии, он «родился» в 1938 году в Бруклине, Нью-Йорк. Знаменит приключениями, в которых легко побеждает любых врагов, а также бруклинским акцентом и фразой «В чём дело, Док?»...

ru.wikipedia.org



…look at this lovely bugs bunny figure…



…looney tunes: bugs bunny…

…plush pinky rabbit toy…

…click to buy bugs bunny toy…

A lot of remaining questions

- How to identify user’s goal?

A lot of remaining questions

- How to identify user’s goal?
- How to measure end-to-end quality of such a wide variety of possible results?

A lot of remaining questions

- How to identify user’s goal?
- How to measure end-to-end quality of such a wide variety of possible results?
- What to do with similar but not exactly matched tags?

A lot of remaining questions

- How to identify user’s goal?
- How to measure end-to-end quality of such a wide variety of possible results?
- What to do with similar but not exactly matched tags?
- How to filter trash linked text?

A lot of remaining questions

- How to identify user’s goal?
- How to measure end-to-end quality of such a wide variety of possible results?
- What to do with similar but not exactly matched tags?
- How to filter trash linked text?
- How to make list-wise ranking instead of point-wise?

A lot of remaining questions

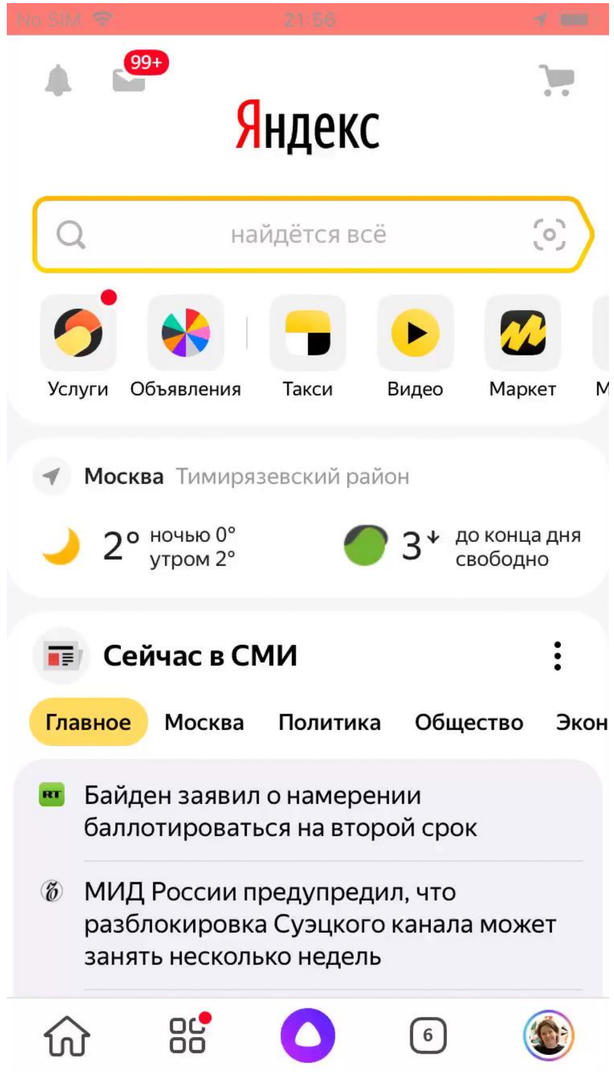
- How to identify user’s goal?
- How to measure end-to-end quality of such a wide variety of possible results?
- What to do with similar but not exactly matched tags?
- How to filter trash linked text?
- How to make list-wise ranking instead of point-wise?
- ... and many many more …

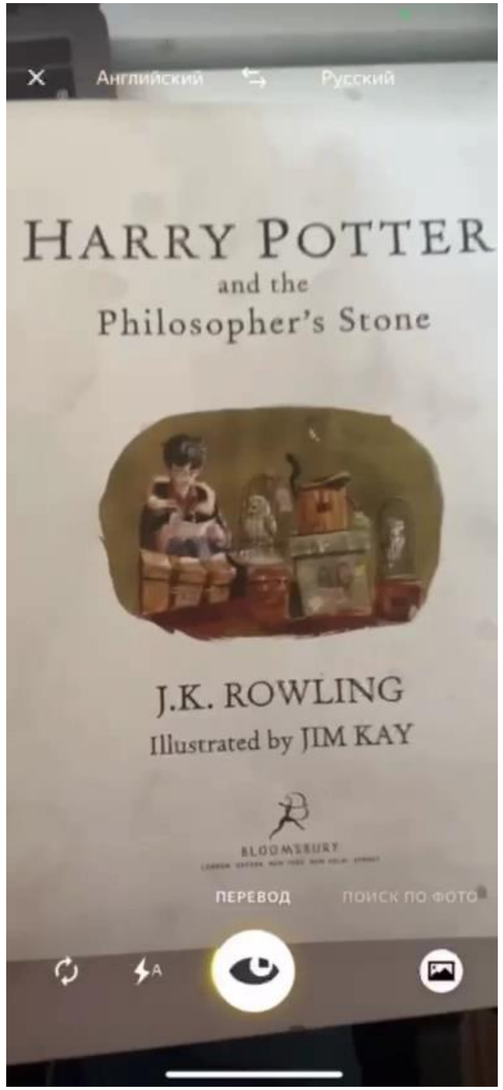
The end

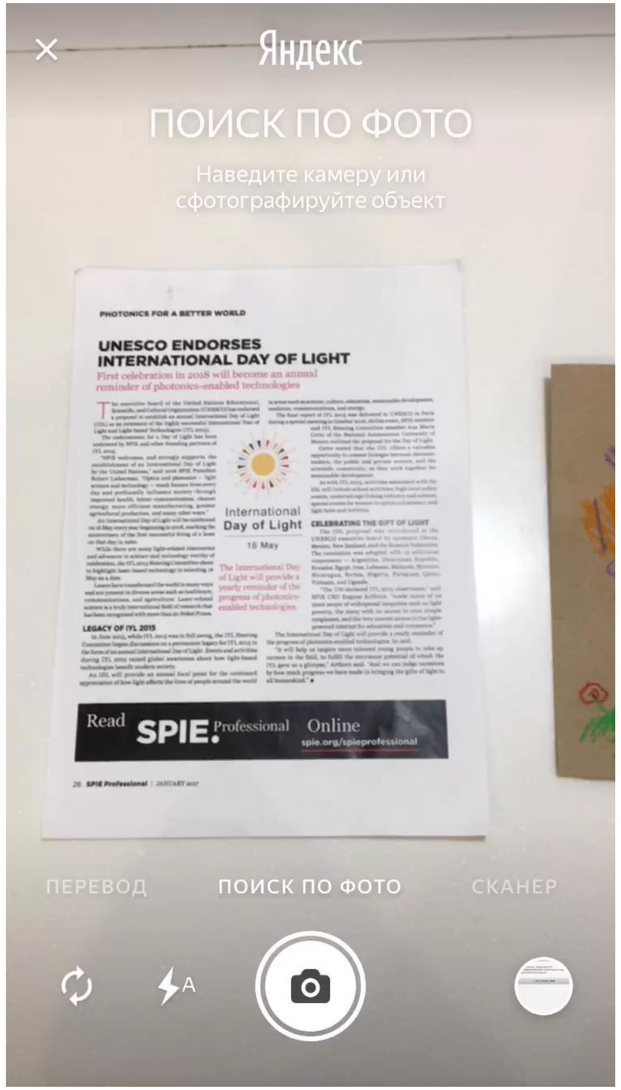
~~The end~~

Almost done…

Smart camera: beyond visual search







Quick recap

- Good basic image features is key ingredient
- Visual search is a full-stack search engine with many stages and factors
- More factors → more quality
- Open-ended tagging mechanism is based on similar images search

Yandex

Thank you!

Konstantin Lakhman, *Head of Computer Vision and ML-Platforms Department*

klakhman@yandex-team.ru