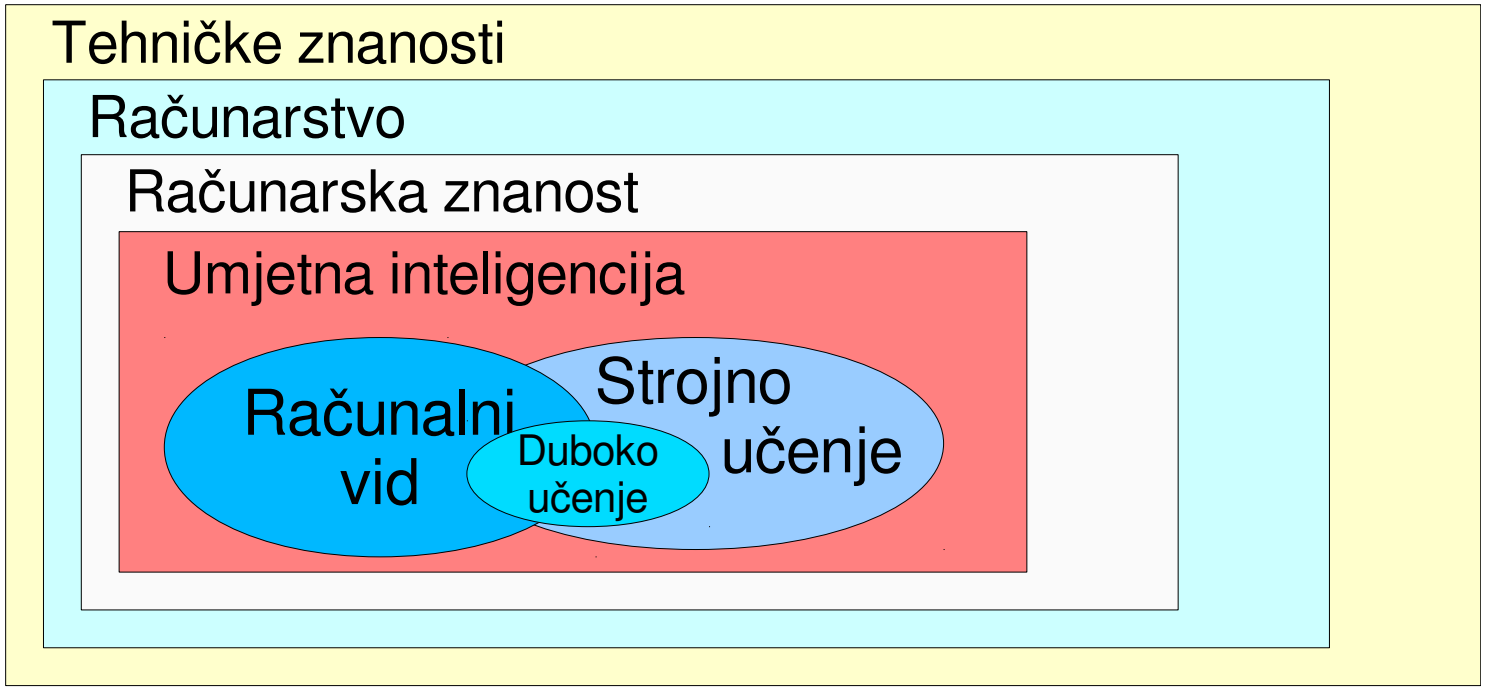
Modeli za predstavljanje slike i videa

Uvodno predavanje

Siniša Šegvić
Fakultet elektrotehnike i računarstva
Sveučilište u Zagrebu

O predmetu: karta pojmova



Umjetna inteligencija:

- tehnički sustavi za poslove koje ljudi obavljaju bolje od strojeva

Računalni vid:

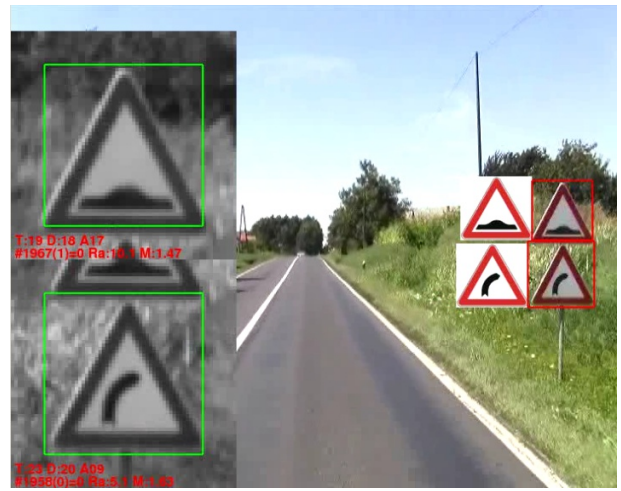
- razumjeti procese strojne percepcije i spoznaje
- stvoriti novu vrijednost kroz napredne usluge

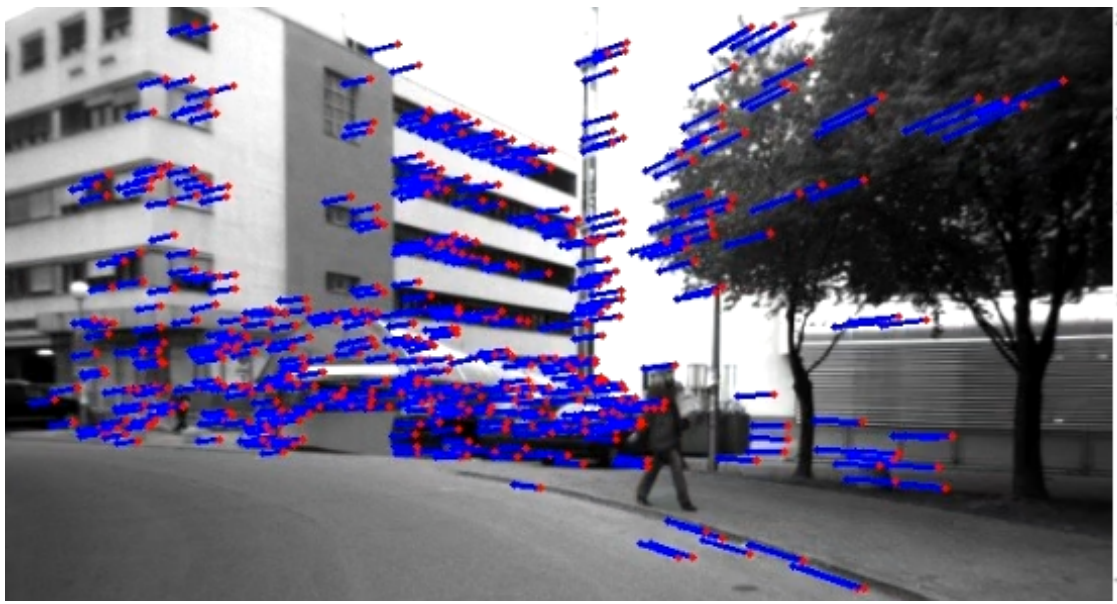
O predmetu: računalni vid

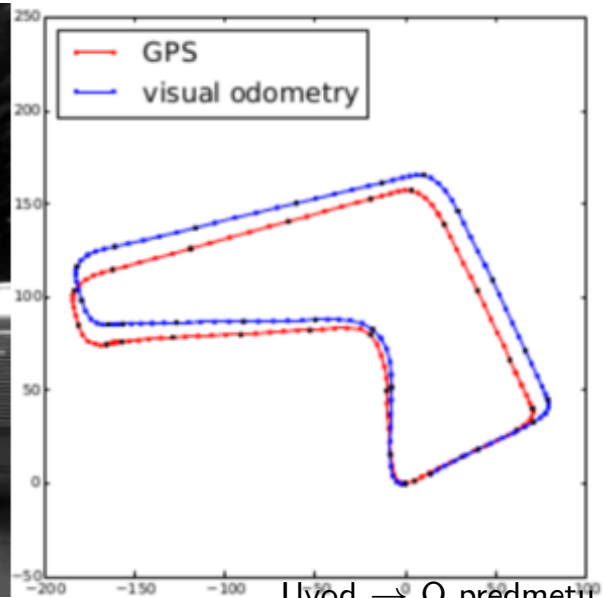
Proučavamo tehnike za izlučivanje korisnih informacija iz slika

Dva glavna pristupa prema tom cilju:

- **raspoznavanje**: naučiti izlučiti simboličku informaciju: ima li objekta na (x,y)?
- **rekonstrukcija**: procijeniti numeričku informaciju vezanu uz geometriju: struktura scene, gibanje kamere, ...







O predmetu: metode

Oba pristupa koriste naprednu optimizaciju:

- **raspoznavanje**: optimirati preslikavanje slika u vjerojatnosti
  - kriterij: uspjeh na skupu za učenje (+ regularizacija...)
  - optimizacija tijekom učenja, (relativno) jednostavno zaključivanje
  - pronalaženje objekata, klasificiranje objekata, klasificiranje slika
  - krajnji cilj kognitivnog vida: razumjeti sliku
- **rekonstrukcija**: optimirati geometrijski model
  - kriterij: slaganje s mjerenjima u slici
  - nema učenja, optimizacija se provodi tijekom "zaključivanja"
  - često trebamo više od jedne slike

Fokus ovog kolegija je **raspoznavanje**:

- ključ: izlučiti kvalitetne slikovne reprezentacije
- danas to provodimo modelima **strojnog učenja**

O predmetu: vrijeme

Sve donedavno, raspoznavanje u "divljini" je bilo akademska diciplina

- primjenljivi rezultati dobivani su samo u laboratorijskim uvjetima
- međutim, vrijedan rad doveo je do fascinantnog napretka

Danas je kognitivni vid tehnologija s jasnom industrijskom vrijednošću

- Google, Facebook, Intel, Microsoft, Nvidia, i mnogi kineski giganti u područje ulažu milijarde dolara
- Google kupio DeepMind for 0.6 B$ (2014)
- Intel kupio MobileEye for 15.3 B$ (2017)
- NVidia investirala 20 M$ in TuSimple (2017)

Pogledajmo zašto je kognitivni vid toliko zanimljiv!

Primjene: autonomna vozila

https://www.youtube.com/watch?v=9ydhDQaLAqM

Društvo automobilskih proizvođača definira 6 razina autonomije (2014)

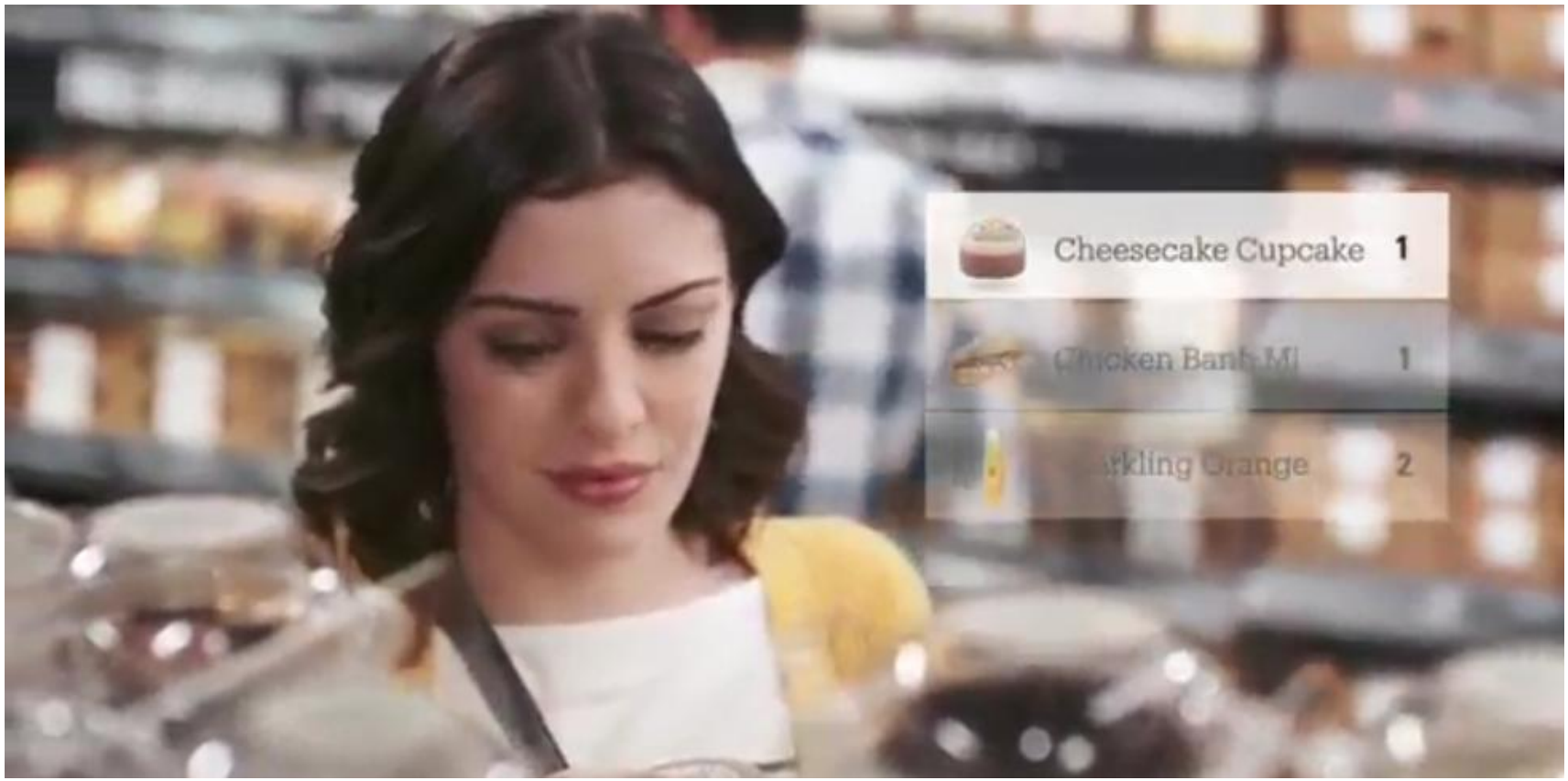
- Teslin autopilot zadovoljava razinu 2 (2016).
- Prototipi razine 4 već su na cestama: Google, Tesla, Uber, Audi, Renault

Primjene: dijagnostika bez liječnika

http://www.nature.com/nature/journal/v542/n7639/abs/nature21056.html

Esteva et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118 (02 February 2017)

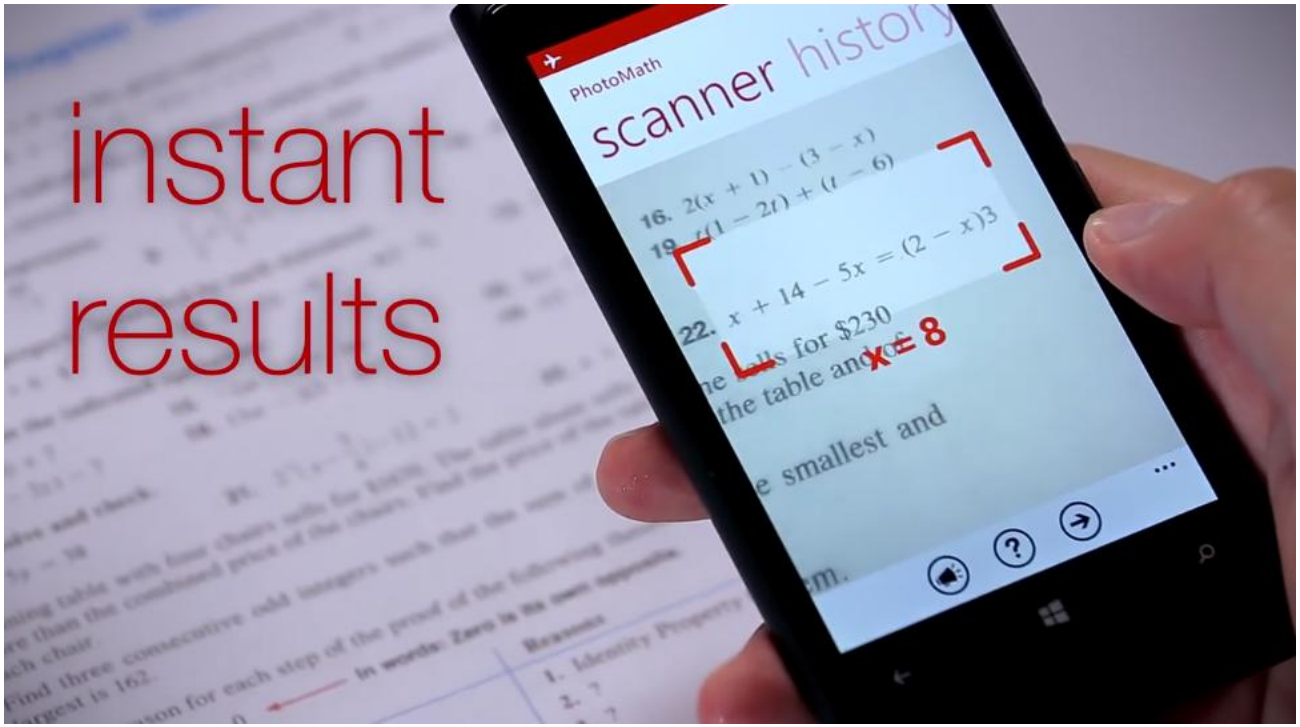
Primjene: dućan bez blagajni

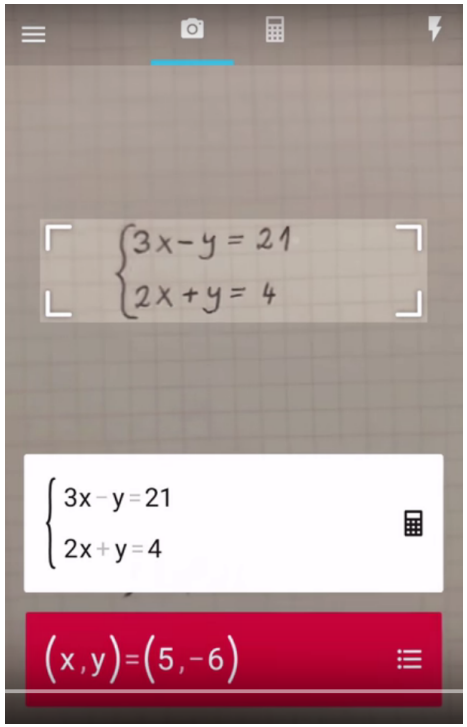


https://www.youtube.com/watch?v=Jg0pQaV44I8

Amazon Go, 2131 7th Ave, Seattle, WA.

Primjene: uplatnice bez šaltera...



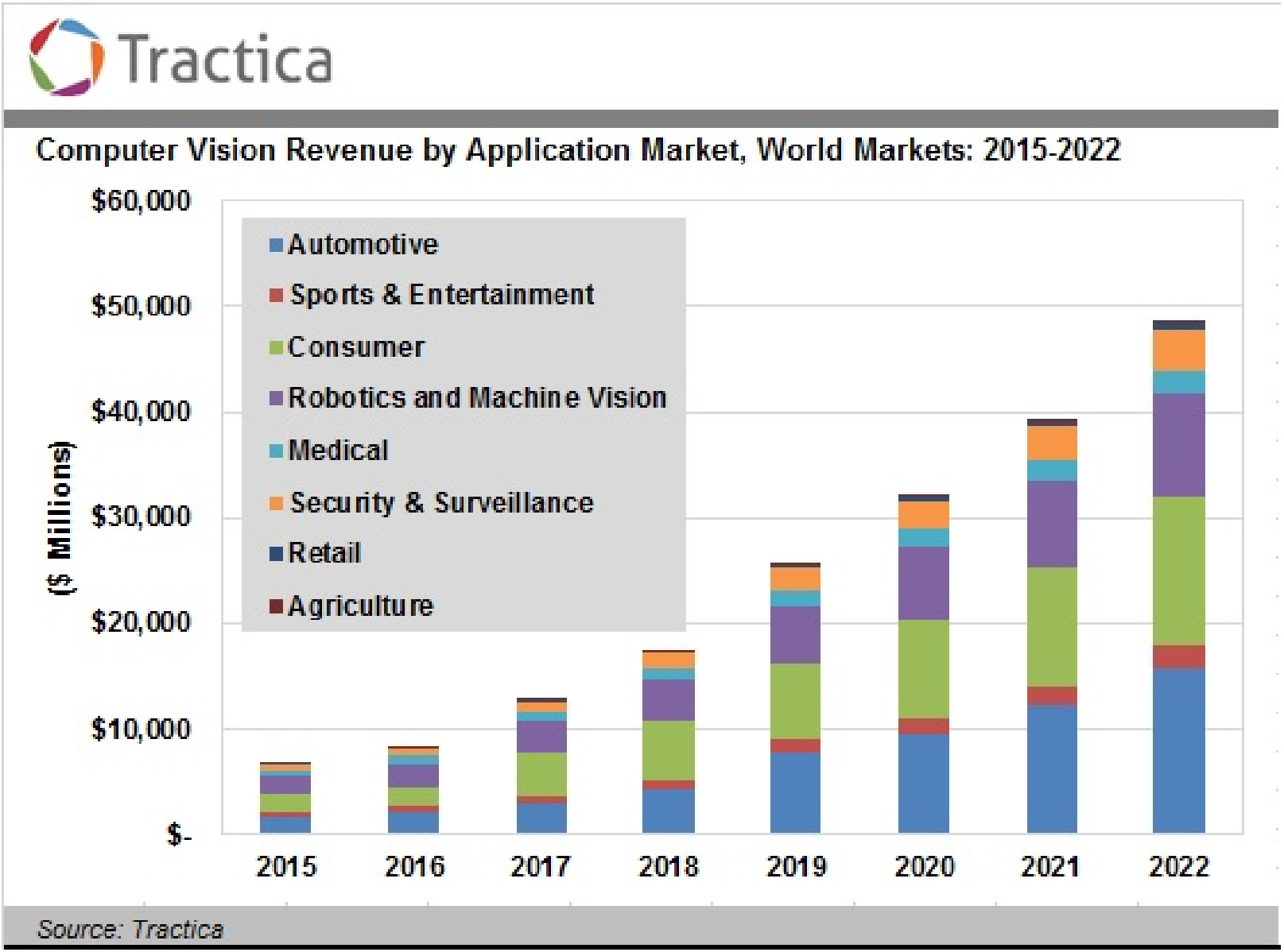


https://www.youtube.com/watch?v=oXn9NuUmhDM

https://www.youtube.com/watch?v=XlbVB50mIh4

...i rješavanje matematičkih zadataka (s postupkom!)

Primjene: projekcije i prognoze



https://www.tractica.com/

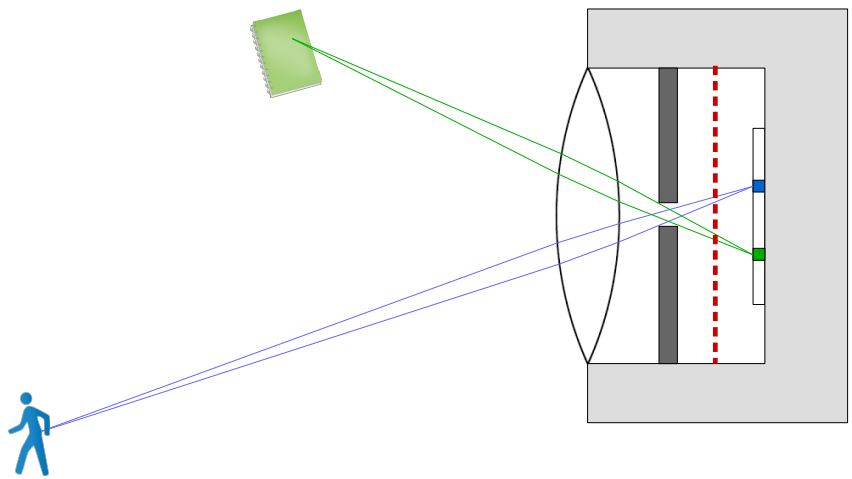
Plan

1. Uvod i motivacija
   - pristupi strojnog gledanja, primjene
2. Pogled unatrag: klasični računalni vid
   - o digitalnim slikama, važna postignuća
3. Klasifikacija slike
   - pristupi, prednosti, nedostatci
4. Duboki konvolucijski modeli
   - arhitekture, slojevi, kod
5. Sklopovska podrška, trendovi razvoja
6. Zaključak

Digitalne slike: stvaranje

Digitalna kamera ima četiri glavna elementa:

1. senzor: a pravokutno polje osjetilnih elemenata ($\rightarrow$ **pixeli**!)
2. otvor: rupica koja propušta svjetlost na senzor
   - osjetilne elemente pobuđuju odgovarajući uski snopovi svjetlosti
3. leća: uređaj za prikupljanje svjetla (osigurava jači signal)
4. zatvarač: osigurava da ekspozicija traje djelić sekunde



Za slike u boji treba nam malo složeniji sustav

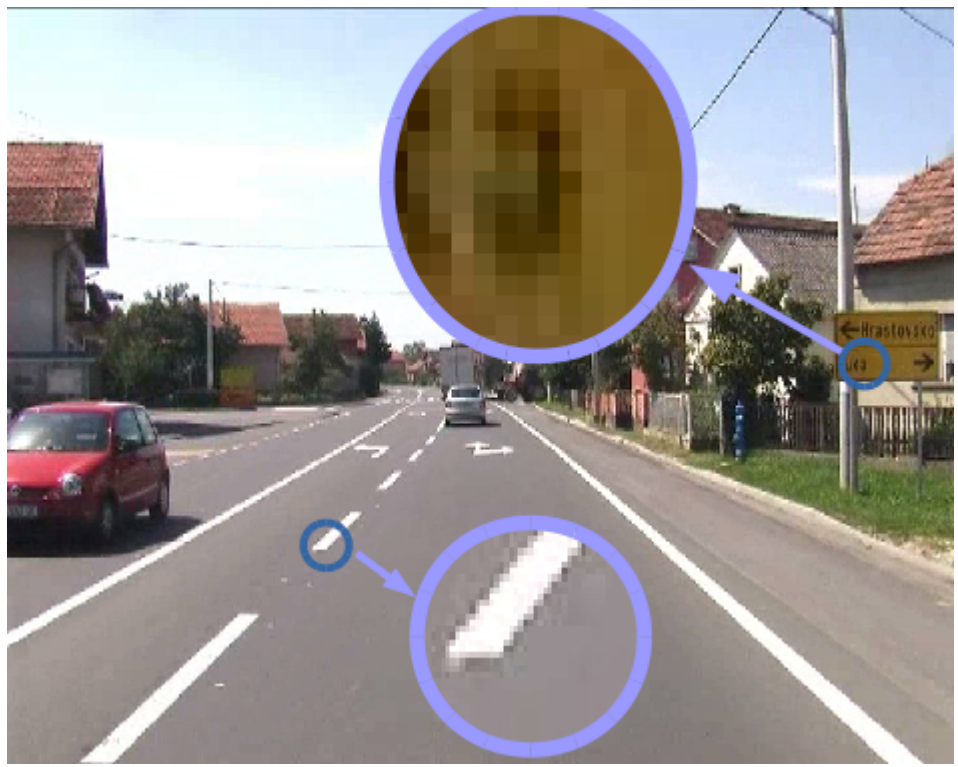
- npr. tri prepletena filtra, tri senzora i sustav ogledala

Digitalne slike

Tako, svaka digitalna slika je pravokutno polje piksela

- dimenzije (retci $\times$ stupci) ovise o senzoru

Ako sliku dovoljno povećamo, pikseli postaju vidljivi:

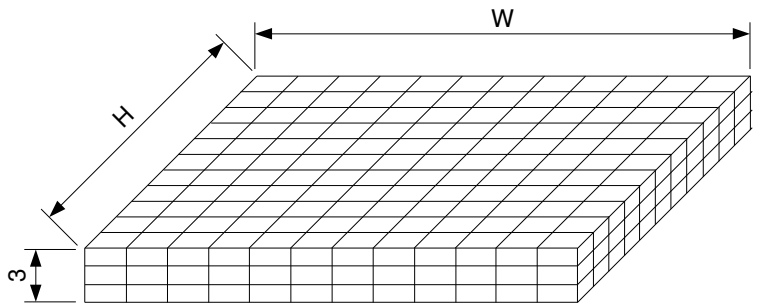


Ako se to ne dogodi, znači da vas program za prikazivanje slika vara :-)

Digitalne slike: struktura

Ako se radi o slici u boji, svaku piknju opisujemo trojkom (R, G, B).

Takvu slike predstavljamo tenzorom $H \times W \times 3$:



Ispeglani tenzor možemo smjestiti u memoriju kao polje $H \cdot W \cdot 3$ brojeva:

R_{11}	G_{11}	B_{11}	R_{12}	G_{12}	B_{12}	...	...	...	R_{HW}	G_{HW}	B_{HW}

Sliku od 200×200 piksela možemo predstaviti 120000-D vektorom.

Pogled unatrag: početak (1966)

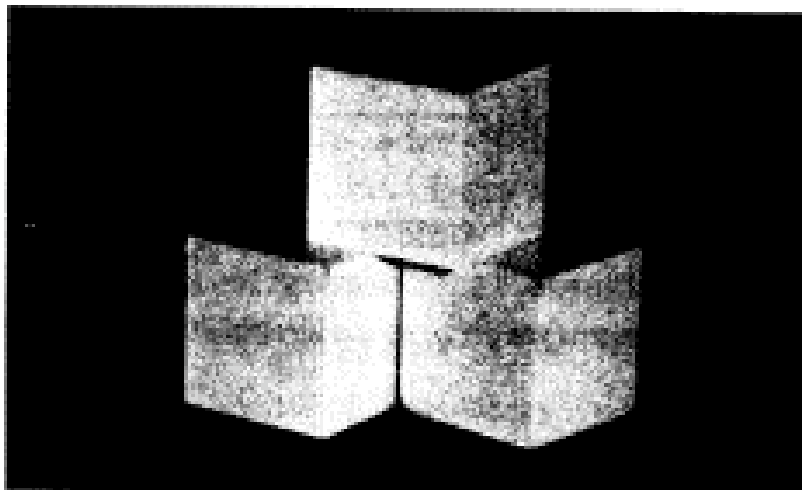
Odmah nakon pojave digitalnih slika postavilo se logično pitanje:

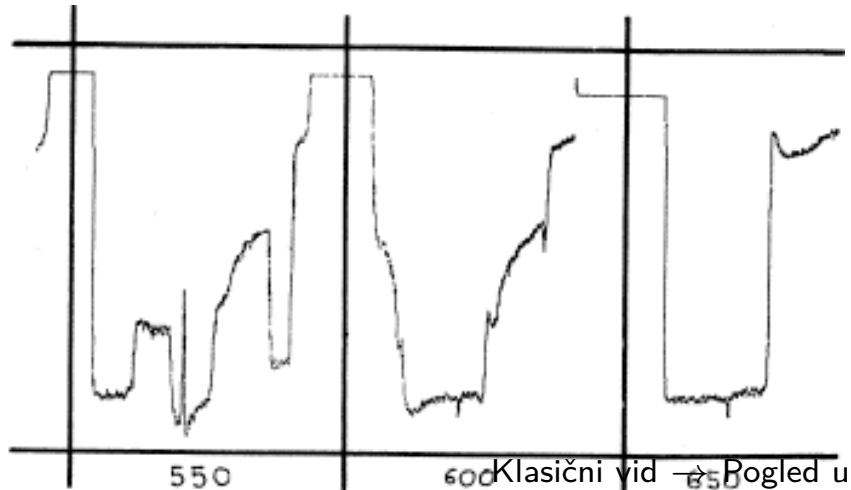
- može li računalo analizirajući sliku razumjeti svijet?

1966 godina: jedan američki profesor predlaže diplomandu da napiše program koji će ispisati što ima u slici.

Taj zadatak je bio ispred svog vremena: problem je mnogo godina ostao neriješen unatoč vrijednom radu mnogih ljudi...

Razumijevanje problema raslo je polako ali ustrajno, ukratko ćemo ponoviti glavna postignuća

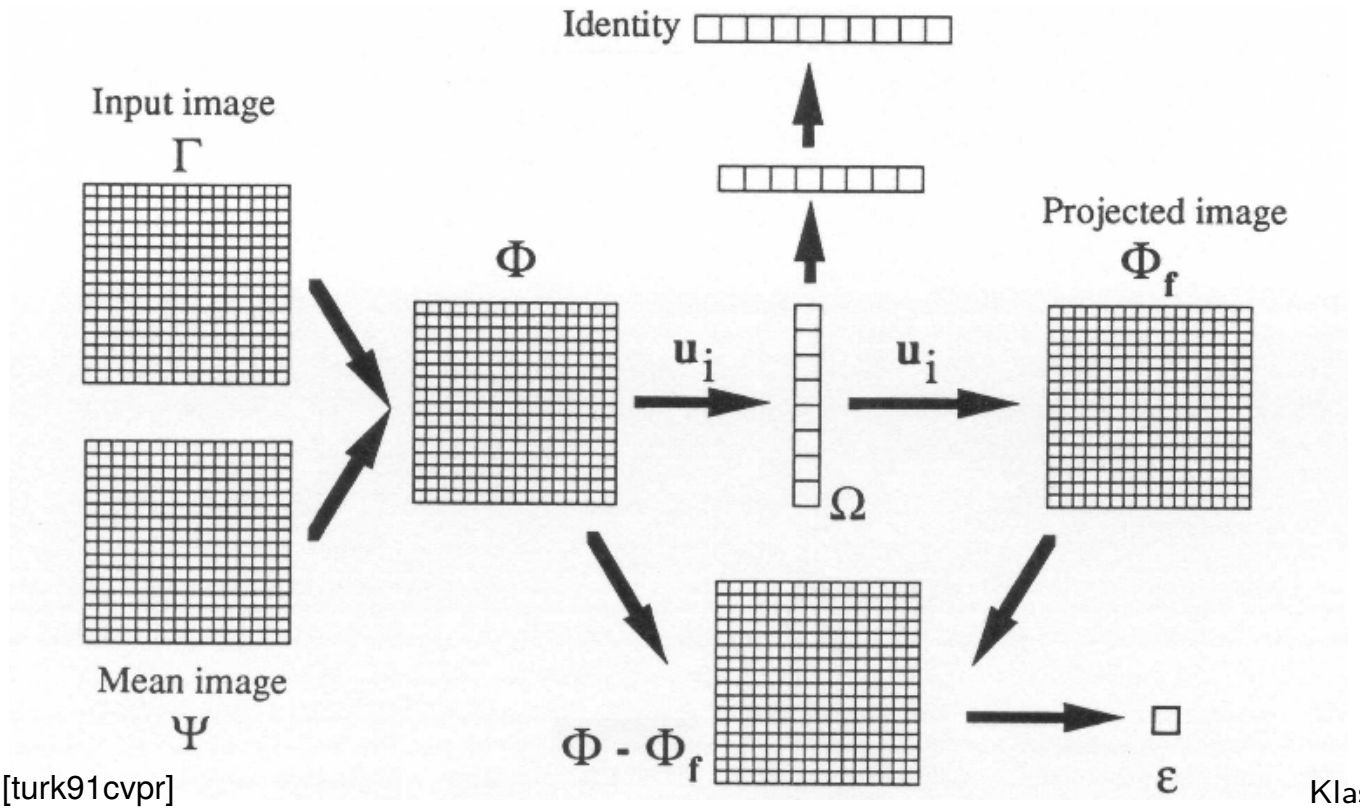




Pogled unatrag: klasifikacija objekata (1991)

Recept za klasificiranje poravnatih slika lica [turk91cvpr]:

- slike se tretiraju kao visokodimenzionalni podatci
- hypotetizira se da slike nastanjuju nižedimenzionalni **prostor lica**
- uči se linearna projekcija na prostor lica (podatci za učenje!)
- provodi se raspoznavanje na reprezentaciji slike u prostoru lica

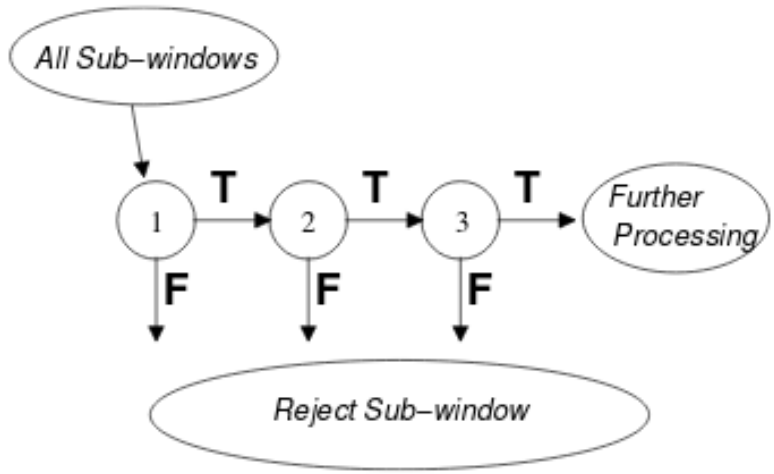


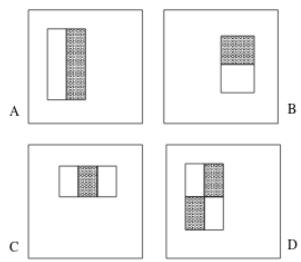
[turk91cvpr]

Pogled unatrag: pronalaženje objekata (2001)

Jednostavni objekti, stvarno vrijeme [viola01cvpr]:

- pronalaženje se tretira kao binarna klasifikacija u pomičnom oknu
- klasifikator se uči kao kaskada sve strožih detektora
- pojedine razine kaskade kombiniraju **jednostavne** značajke
- mogućnost klasificiranja preko 100000 okana na 25 Hz



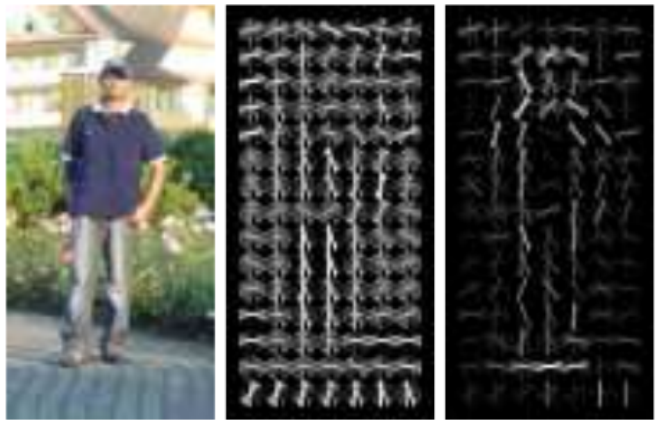


[viola01cvpr]

Pogled unatrag: pronalaženje objekata (2005)

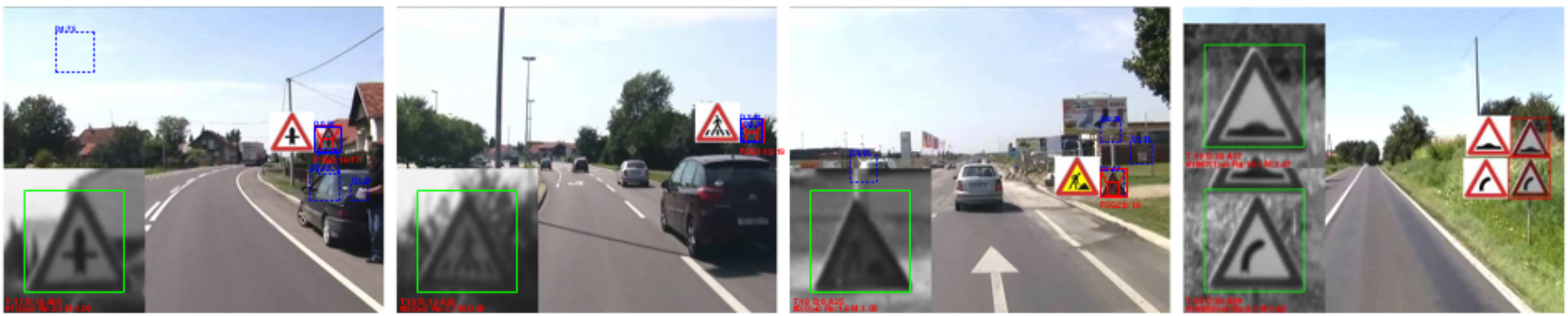
Nešto složeniji objekti [dalal05cvpr]:

- reprezentirati okna ručno dizajniranim lokalnim opisnikom (HOG)
- linearna klasifikacija okana
- moćniji i sporiji od [viola01cvpr]



[dallal05cvpr]

Povezivanje posljednjih dvaju pristupa omogućava pouzdano pronalaženje prometnih znakova:

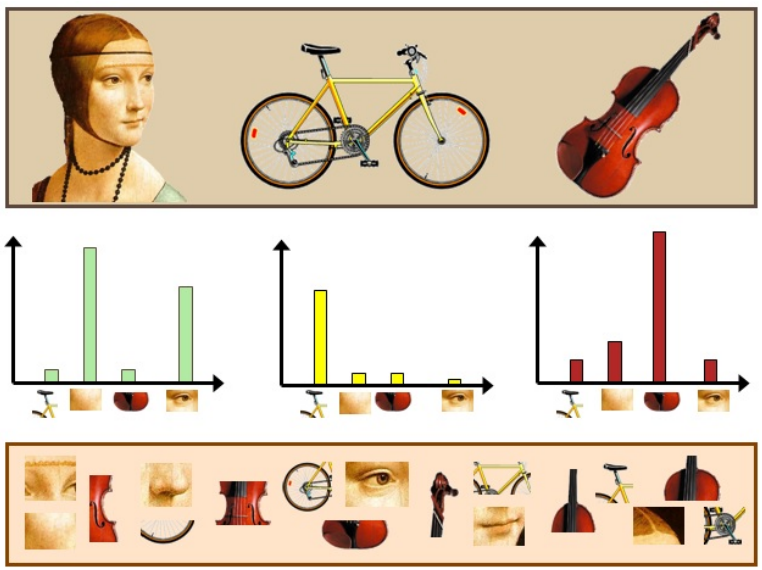


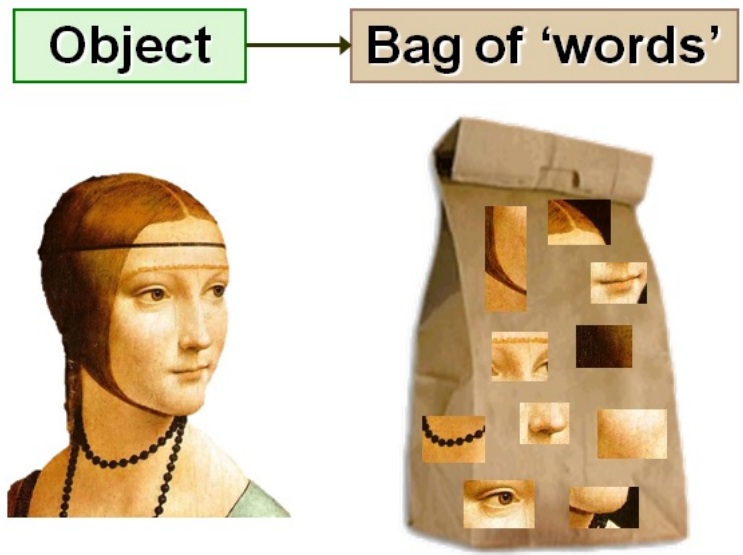
[segvic14mva]

Pogled unatrag: klasifikacija slika (2004)

Recept za raspoznavanje neporavnatih slika [csurka04eccv]:

- kodirati izgled okana ručno dizajniranim značajkama (SIFT)
- naučiti diskretni rječnik opisnika okana (slikovne riječi)
- reprezentirati slike kao histograme slikovih riječi
- klasificirati slikovne reprezentacije plitkim klasifikatorom





[fergus09iccv]

Moderno vrijeme: jučer

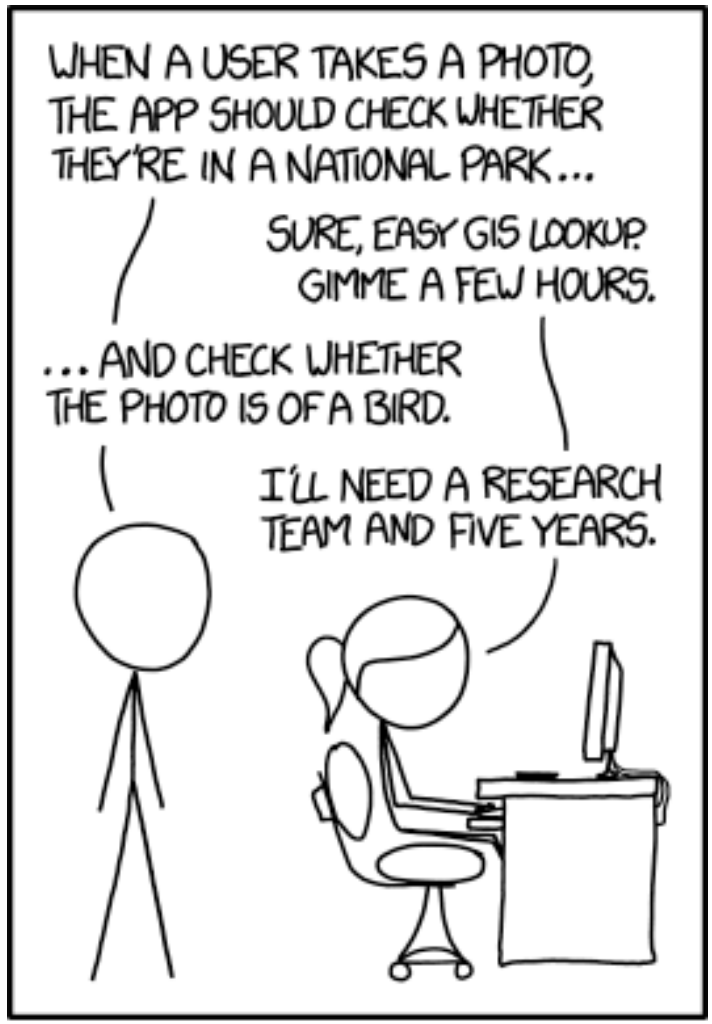
Konačno, svi su saznali da je vid težak...

2014: popularna kultura predstavlja računalni vid kao nemoguću misiju...

- zaključiti sadrži li slika pticu?
- pa, s istraživačkom ekipom i pet godina - možda...

Međutim, od 1966 smo prevalili veliki put:

- pouzdana detekcija ptica moguća još 2010 (s istraživačkom ekipom)
- današnja metodologija i alati omogućuju nam da takve probleme riješimo kod kuće!

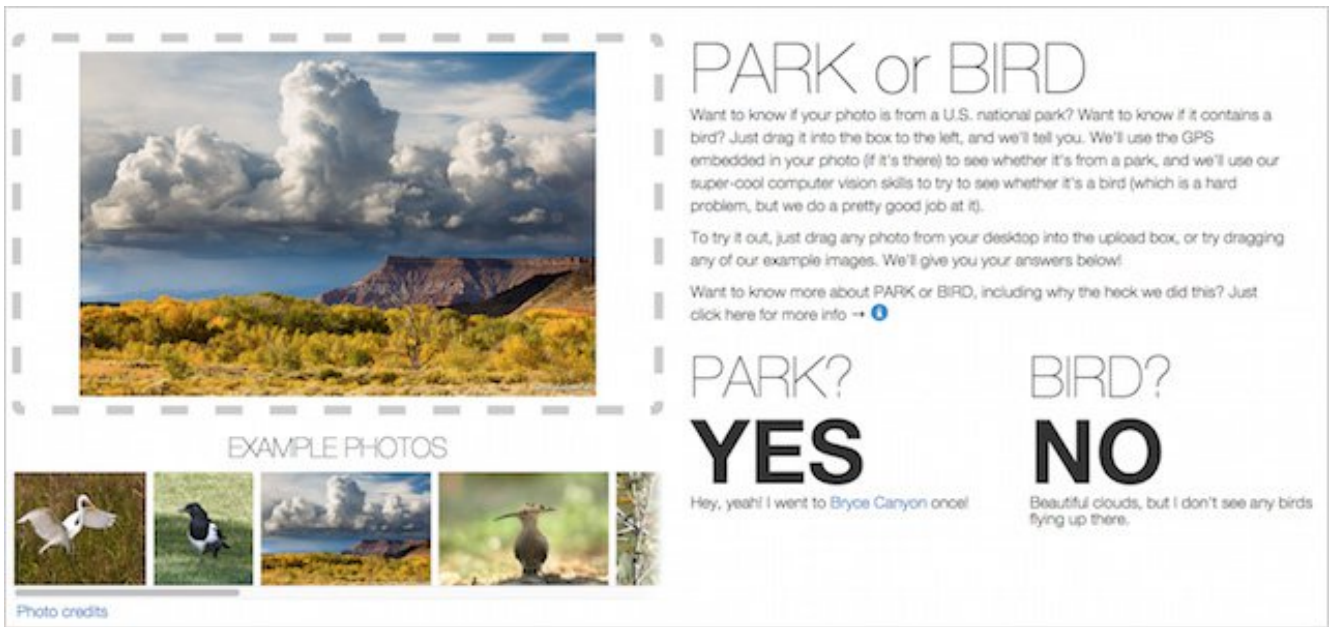


https://xkcd.com/1425 24.09.2014

Moderno vrijeme: danas

Tako, mjesec dana nakon XKCD 1425, Flickr ponosno javlja:

- riješili smo problem "park or bird"...
- ...u manje od pet godina :-)



https://code.flickr.net/2014/10/20/introducing-flickr-park-or-bird/

Probleme u računarskoj znanosti lako je i potcijeniti i precijeniti :-)

Klasifikacija slika: problem

Image classification is the ultimate recognition problem

Consider the problem of discriminating bison from oxen:

[image-net.org]

Hard because we do not know where is the defining object

Especially hard when intra-class variance is large and inter-class variance is small

Klasifikacija slika: rule-based approach

We could try to solve image classification by a rule-based system:

1. concentrate on image regions projected from four legged animals
2. oxen have longer horns or no horns at all
3. bison are dark brown, etc, etc

It's neat (no learning!), but nobody succeeded to make that work!

[image-net.org]

Klasifikacija slika: learning

Hence, we look up approaches which are able to learn functionality from the data

A machine learning approach would roughly follow the following steps:

1. express the program with many free parameters
   - the parameters determine a transformation which we call the model
2. fit parameters on the training set
3. evaluate performance on the test set

Success depends on the model, training set and processing power

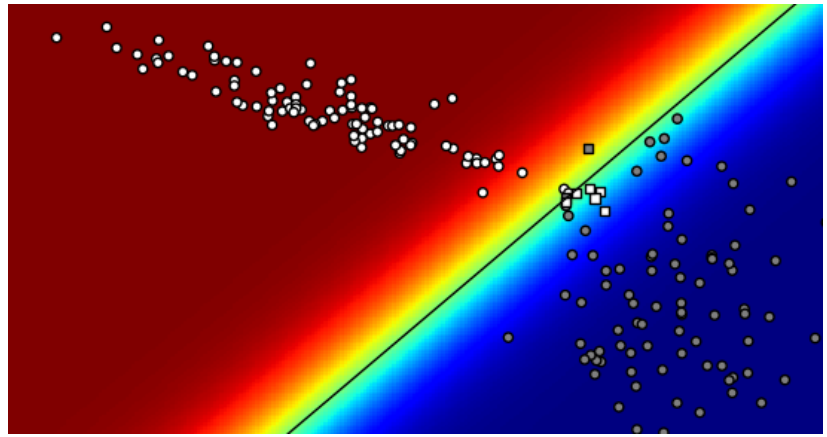
1. model may have insuficient (or excessive) capacity
2. the training set may be too small or not representative enough
3. insufficient processing power $\Rightarrow$ the training may not converge

Klasifikacija slika: shallow classification

Transformation: data $\rightarrow$ decision

1. one linear projection
2. optional squashing non-linearity

An example in 2D: $y_i = \sigma(\mathbf{w}^\top \mathbf{x} + b)$



Advantages of shallow models:

- the best solution is guaranteed and fast
- this is the best approach when classes are linearly separable

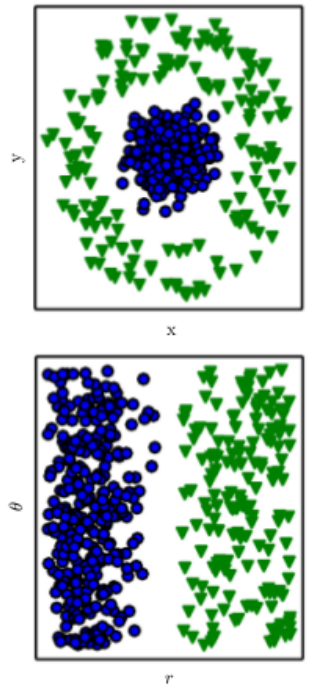
Unfortunately, shallow models are a poor fit for image classification:

- the model can only learn a lookup table
- applicable only for very simple tasks (eg. 88% on MNIST)
- insufficient capacity, tendency to underfit

Klasifikacija slika: data representation

Classification can profit from good representation:

- it would be easy to discriminate bison from oxen if some magical algorithm converted the input image into a binary vector:
  [fur?, small horn?, wilderness?, ...]
- most bison would be: [1, 1, 1, ...]
- most oxen would be: [0, 0, 0, ...]



[goodfellow16]

Hand crafting quality representation is hard:

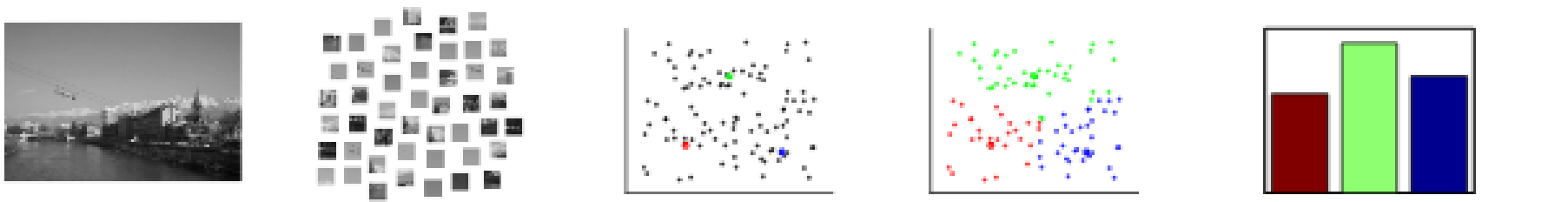
- Greek and Romans did not invent 0 in 1000 years of civilization
- that significantly hindered the development of maths
  - MCMLXXI + XXIX = ?
  - MXXIV : LXIV = ?

A lot of hard work of smart people went into feature engineering

Klasifikacija slika: bag of words

Early image classification approaches (BOW) had three layers:

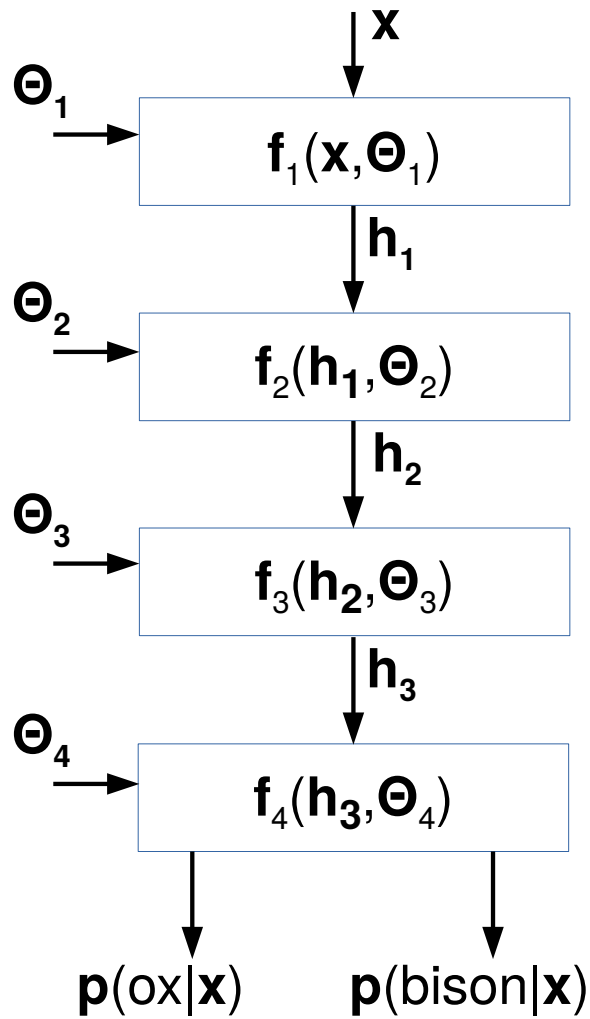
1. describe patches with hand-crafted feature descriptors
2. learn a visual dictionary (a collection of visual words)
3. map patches into visual words, and aggregate them into a histogram
4. classify image descriptors with a shallow model
5. can be viewed as a kernel approach with explicit embedding



[krapac11phd]

Klasifikacija slika: compositional paradigm

Deep model: a sequence of learned non-linear transformations



Why deep learning was not successful?

- no guarantee of the learning success
- non-competitive performance

Why deep learning became popular?

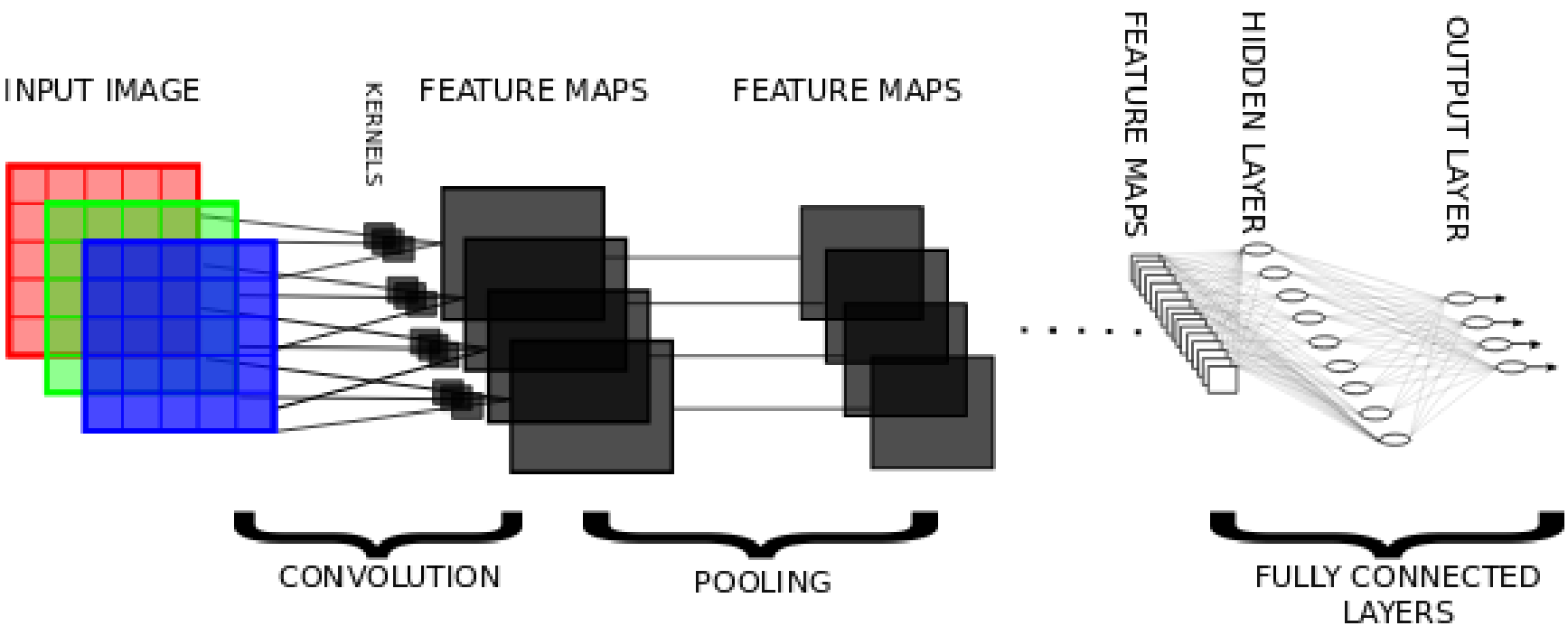
- better modeling and training
- large datasets (n=10^6)
- processing power (TFLOPS)

Suitable for **articulated** data: images, language, speech, bioinformatics...

Convolutional models: architecture

Deep models for image classification typically consist of:

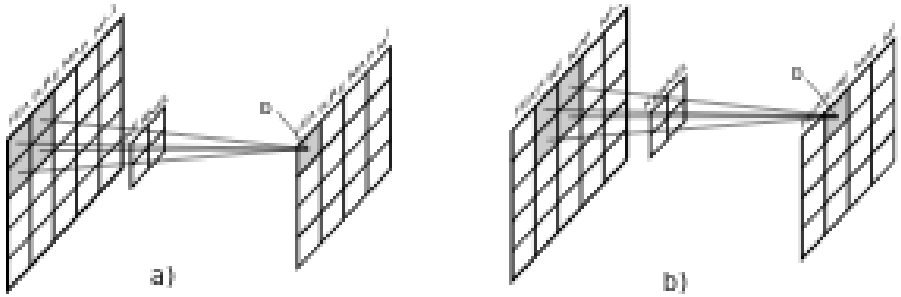
- **convolutions** (linear, recognize object parts)
- poolings (reduce representation dimensionality)
- projections (linear, recognize image as a whole)
- elementwise non-linearities, eg. max(0,x)



[vukotic14ccvw]

Convolutional models: convolutions

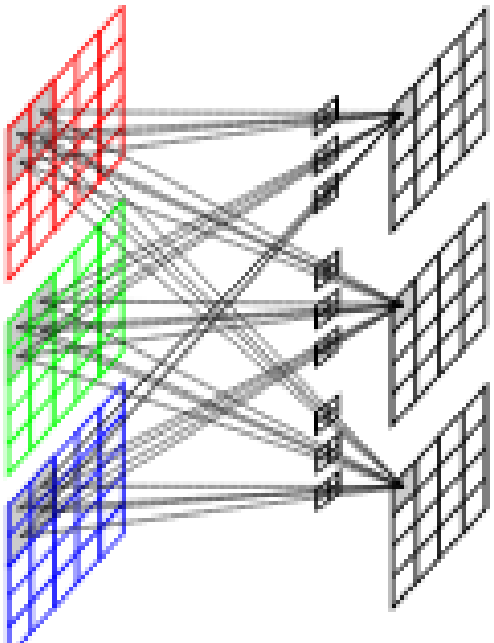
Task: convolve the previous feature map with a kernel

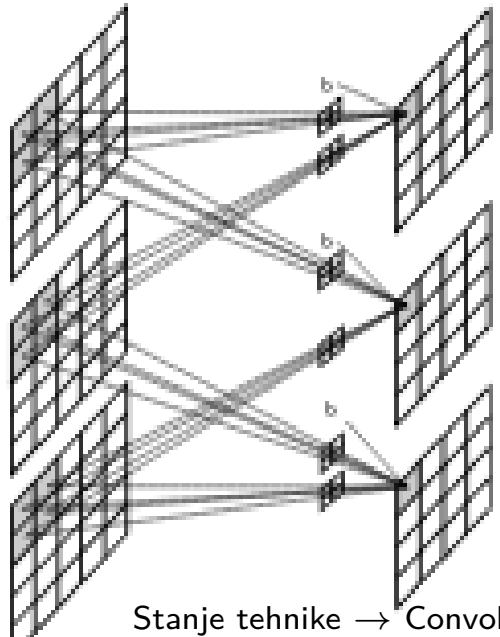


[vukotic14ccvw]

We typically have multiple feature maps on output $\Rightarrow$ multiple kernels

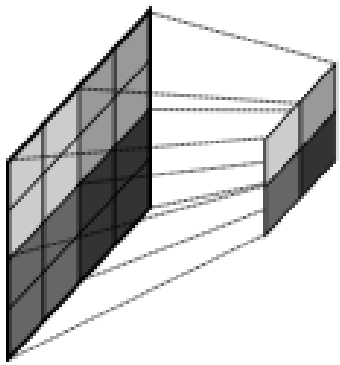
We typically have several feature maps on input $\Rightarrow$ kernels are 3D





Convolutional models: poolings

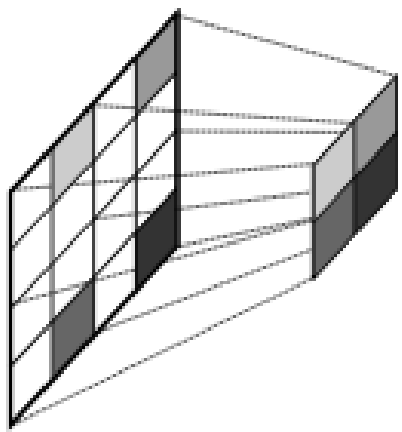
Task: reduce dimensionality to relax memory requirements



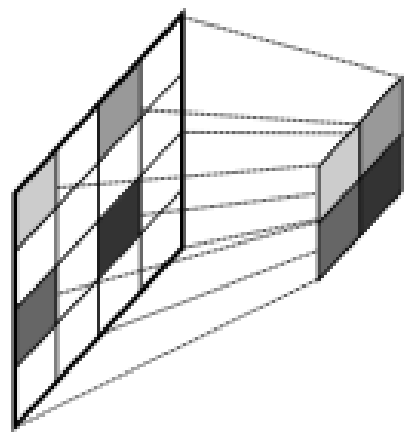
[vukotic14ccvw]

Most often implemented as average pooling or max pooling

It also increases translation invariance:



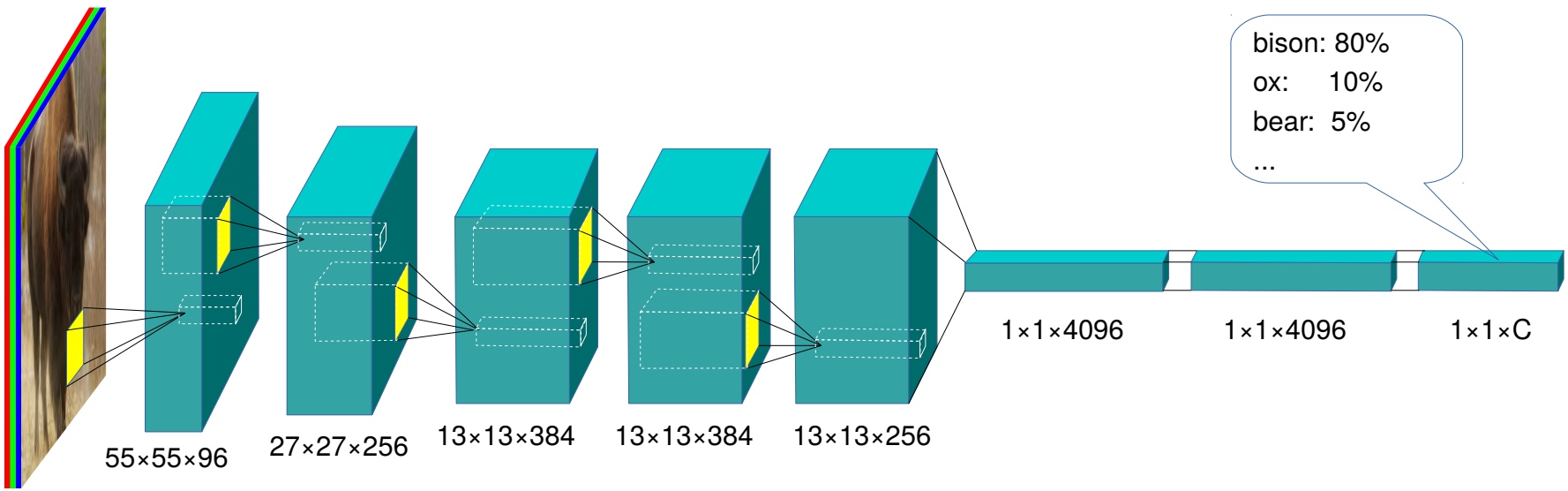
[vukotic14ccvw]



Convolutional models: large-scale

Deep convolutional model for image classification [krizhevsky12nips]

- **input**: image; **output**: distribution over 1000 classes
- **fitness criterion**: average log probability of the correct class
- **structure**: a succession of convolutions and poolings
  - gradual decrease of resolution and increase of the semantic depth
- recent architectures: O(10^2) layers, O(10^6) parameters, O(10^8) bytes, and O(10^9) multiplications for a 224x224 image!



Convolutional models: Two views

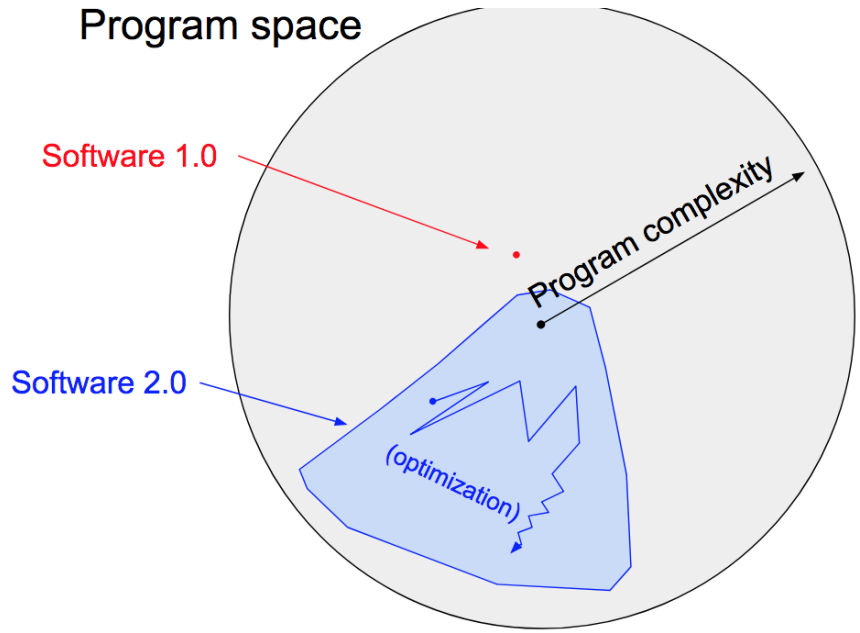
[xkcd 1838]

~~Neural networks~~ →

~~Deep learning~~ →

Differential programming

(software 2.0)



[karpathy17medium]

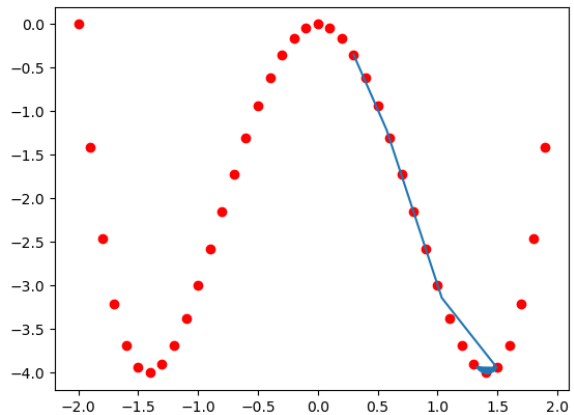
Convolutional models: Diff. programming

```
import torch

step=0.12

x=torch.rand(1, requires_grad=True)
#x=torch.tensor(1.9, requires_grad=True)

for i in range(100):
  y = x**4 - 4*x**2
  y.backward()
  print(x, y, x.grad)
  x.data = x - step*x.grad
  x.grad.zero_()
```

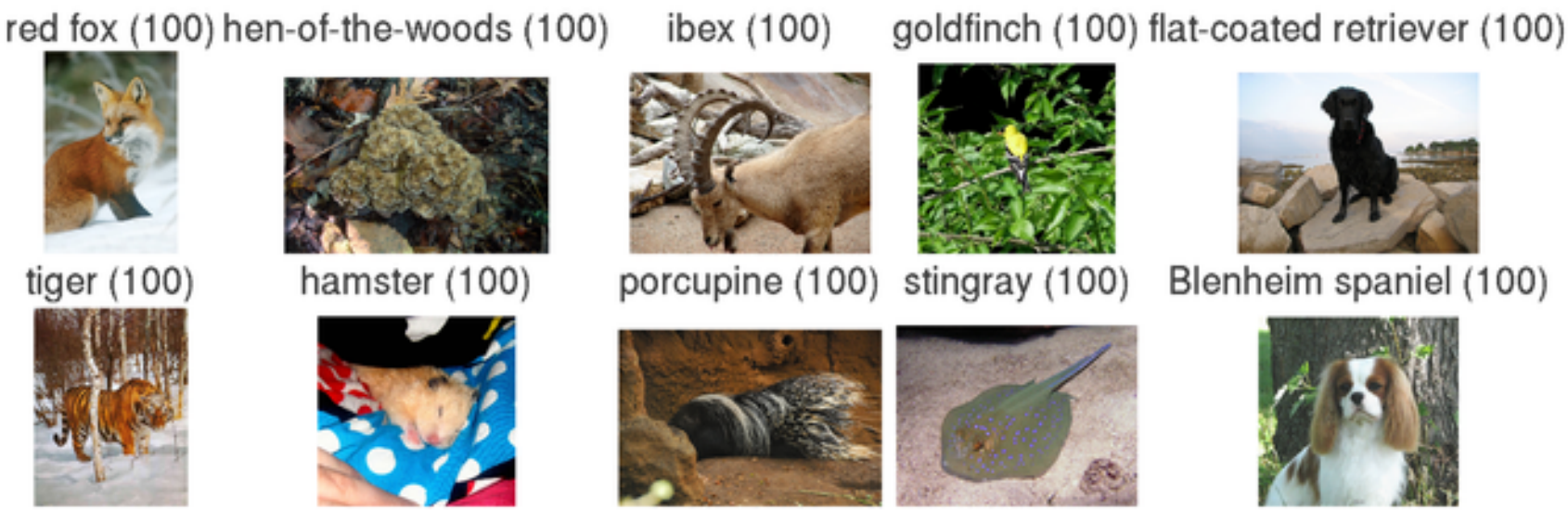


The workhorse algorithm: reverse-mode automatic differentation

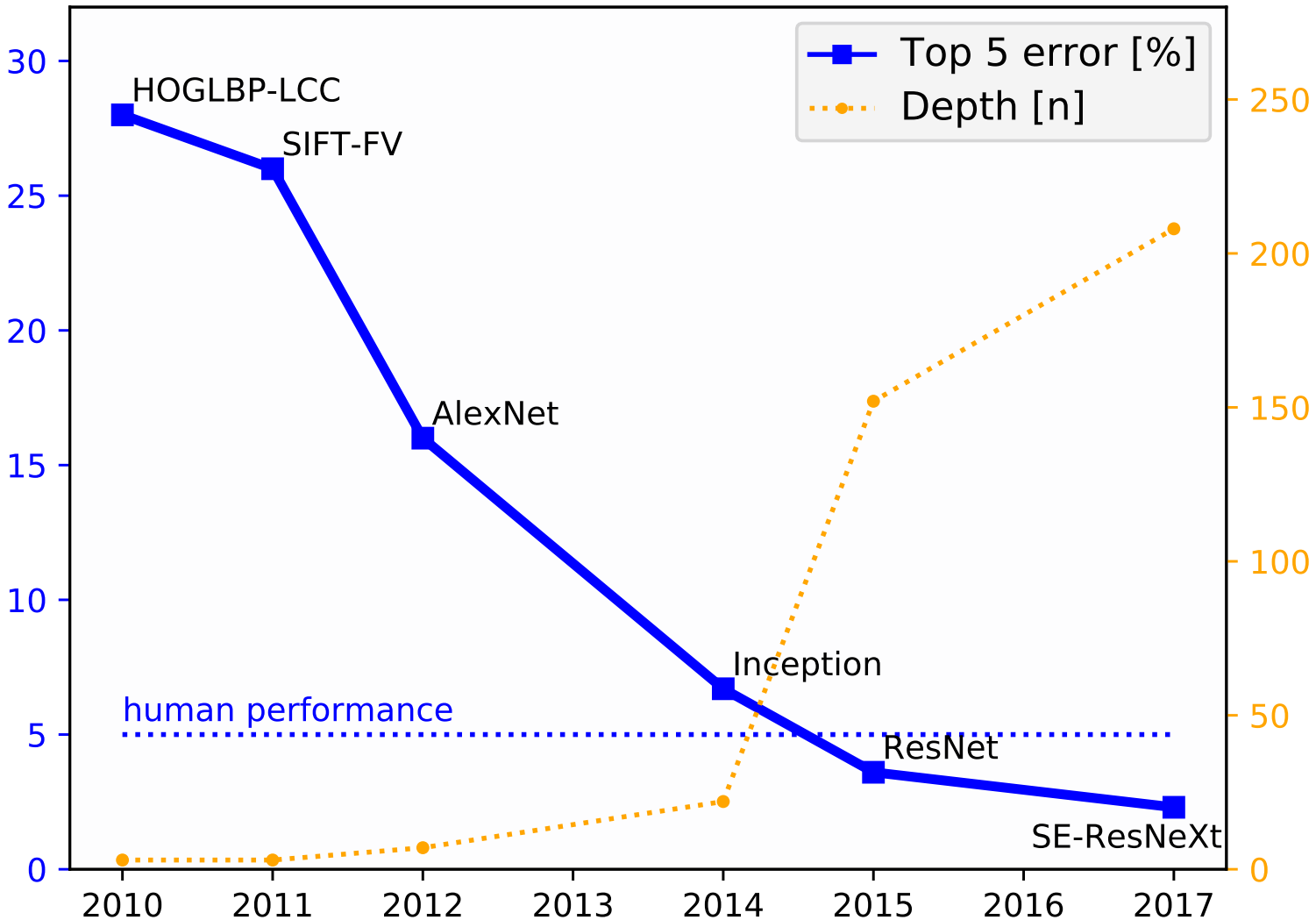
Convolutional models: ImageNet

One of the most popular vision datasets [russakovsky15ijcv]

- annual challenges: classification, localization, detection in video
- we focus on the classification challenge: 10^6 images, 10^3 classes
- fine-grained animals, objects, materials, sports, dishes...
- evaluation metric: top-five prediction error (trained human: 5%)



Convolutional models: ImageNet performance



Convolutional models: ImageNet hard examples

Tasks which are difficult for **humans** [russakovsky15ijcv]:

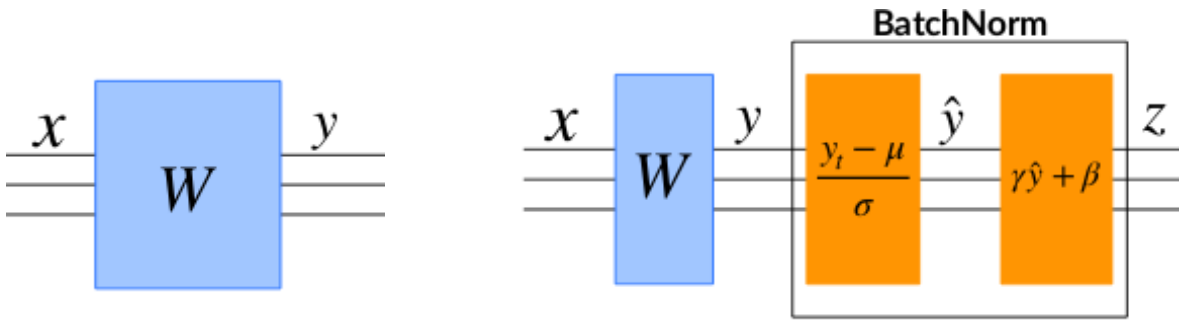
- fine-grained classification (e.g. 120 breeds of dogs!)
- exotic classes (pulley, spotlight, maypole)

Tasks which are difficult for **GoogleNet** (2014, 6.7%):

- little and thin objects, filtered and atypical images
- abstraction (a toy hatchet, images with text)
- large intra-class variance, small between-class variance

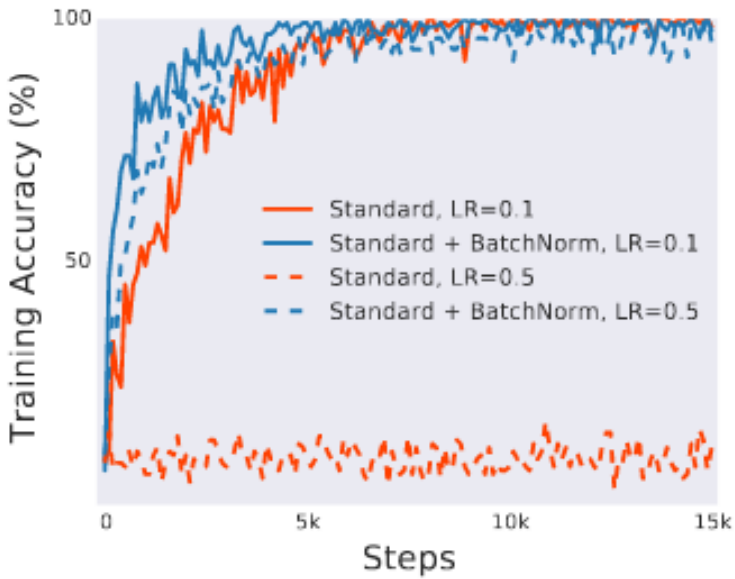
Convolutional models: further improvement

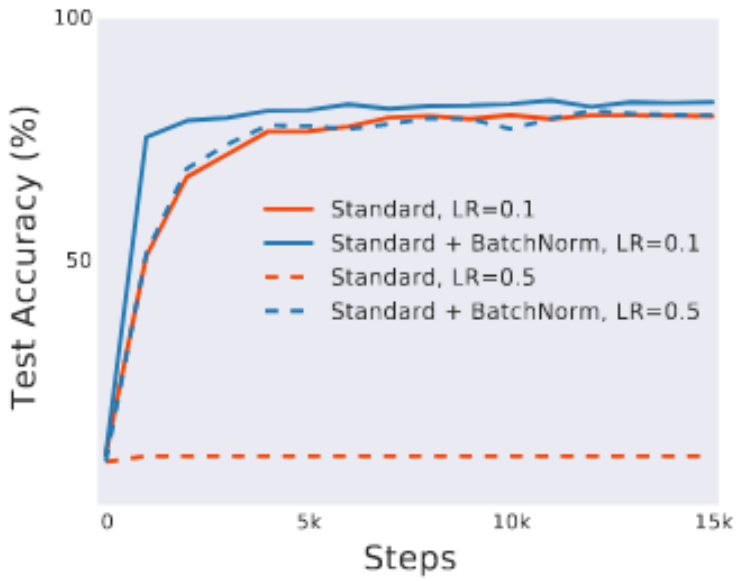
Important idea: normalize activations produced by each layer



vanilla layer layer with batchnorm

Advantages: faster learning, improved generalization

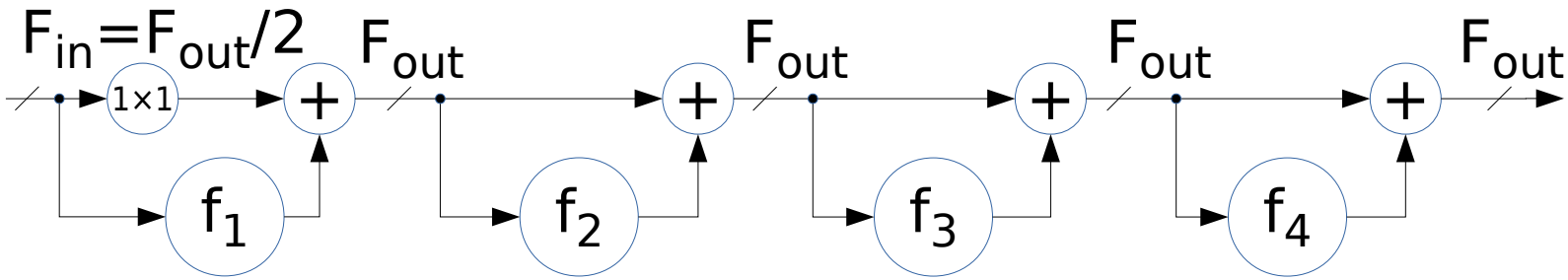


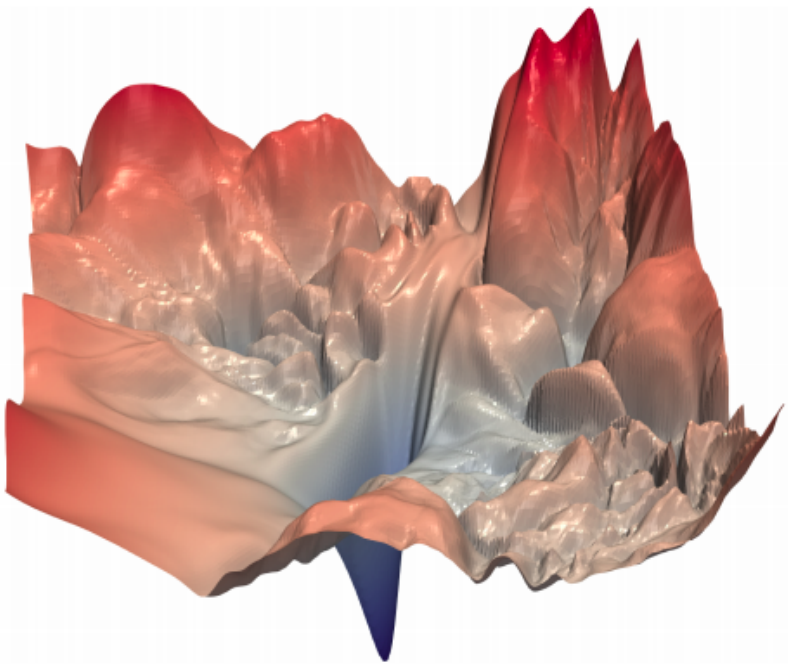


[santurkar18nips]

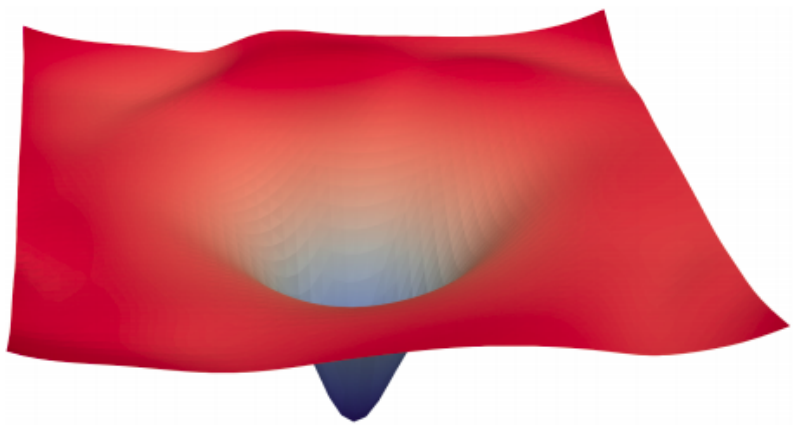
Convolutional models: further improvement (2)

Important idea: enhance the gradient flow towards early layers





(a) without skip connections



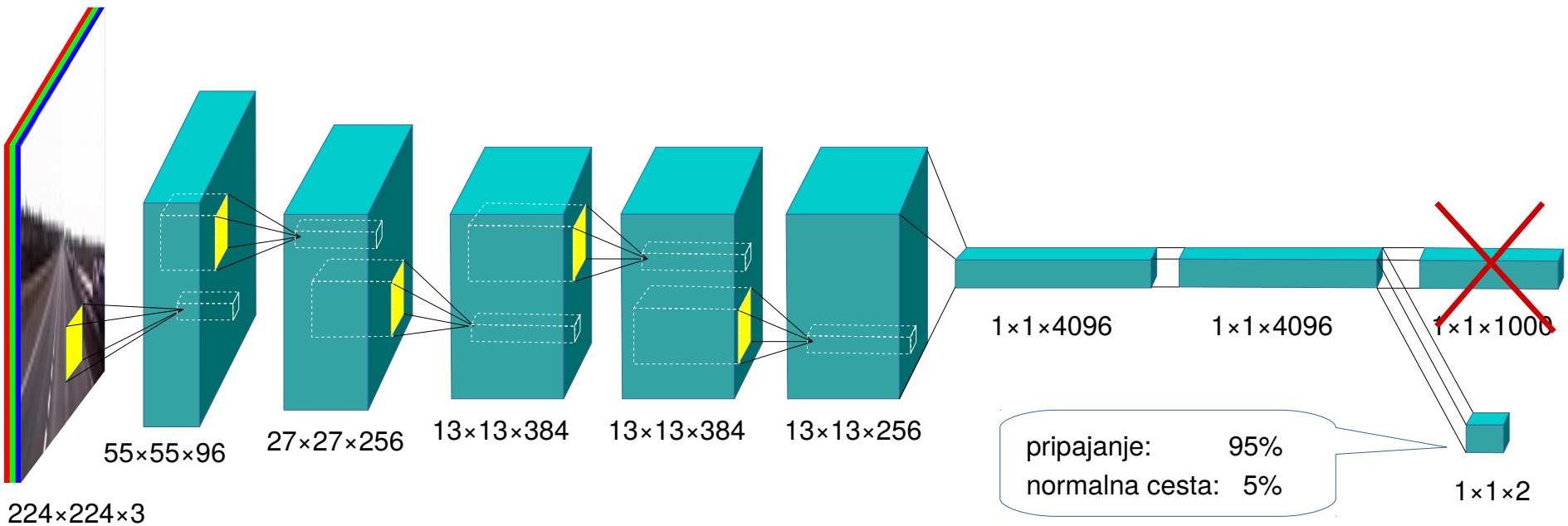
(b) with skip connections

[li18nips]

Convolutional models: knowledge transfer

A deep classification model can be fine-tuned for another (easier) task:

- cut-off the last few layers
- connect the remaining layers with the back end for the new task
- train the resulting model on new images
- inherited layers are well-trained so we can train with less data (few thousands images)



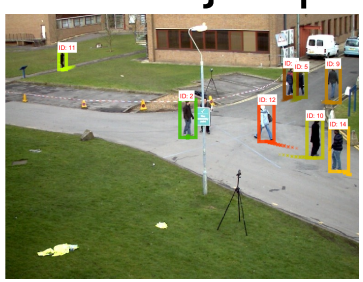
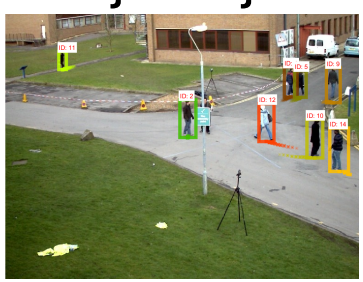
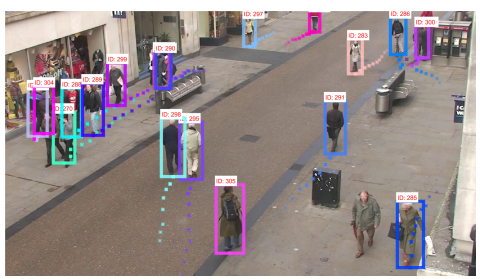
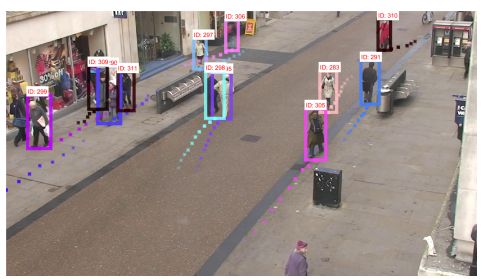
ZADATCI: RASPOZNAVANJE

- Klasifikacija slika i videa



- Objašnjenje predikcija:

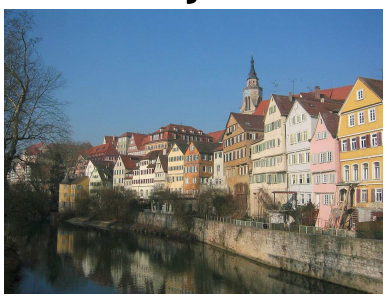
- Detekcija i praćenje objekata:

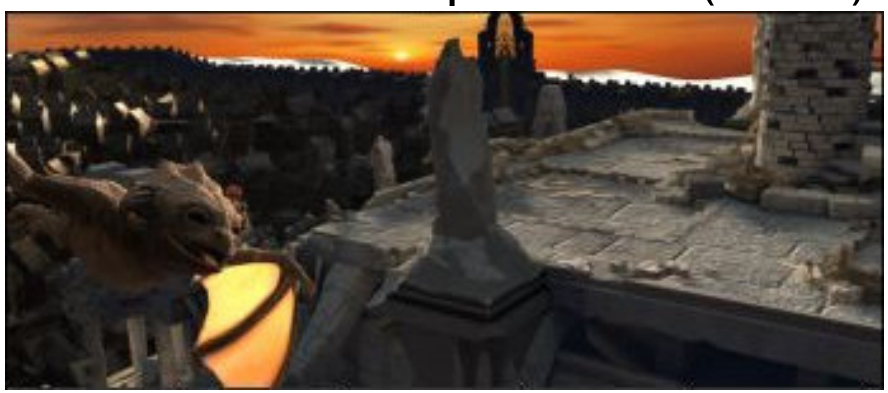
ZADATCI: REKONSTRUKCIJA, ...

- naučena stereoskopska rekonstrukcija [zbontar15cvpr]


- stiliziranje slika prema umjetničkom predlošku [gatys16cvpr]



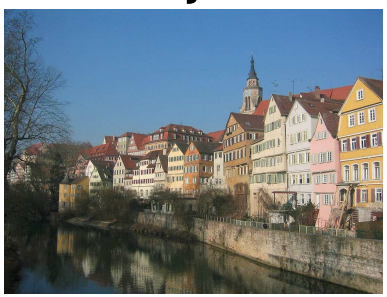
- samonadzirani optički tok (2020)




ZADATCI: REKONSTRUKCIJA, ...

- naučena stereoskopska rekonstrukcija [zbontar15cvpr]

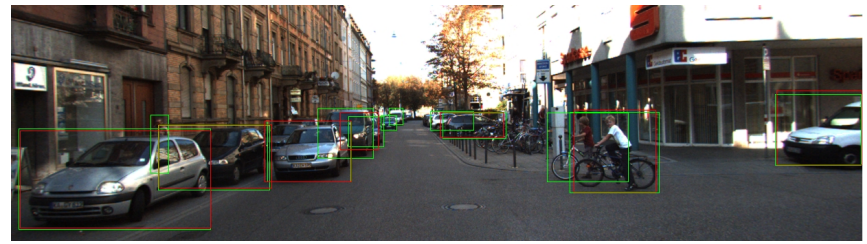

- stiliziranje slika prema umjetničkom predlošku [gatys16cvpr]



- samonadzirani optički tok (2020)


ZADATCI: GUSTA PREDIKCIJA

- Detekcija (lokalizacija) objekata:
  - detektirati prisutnost...
  - ...i odraditi lokaciju



- Semantička segmentacija:
  - klasificirati sve piksele...
  - ...u semantičke razrede

- Stereoskopska rekonstrukcija:
  - odrediti disparitet (dubinu)...
  - ... u svakom pikselu

- optički tok, panoptička segmentacija, semantičko prognoziranje

Hardware: CPU vs GPU

CPU core (AVX) can dispatch up to 2 FMA instructions / cycle

- 2·2·8 operations @3GHz $\rightarrow$ 100 GFLOPS
- if a CPU has 10 cores - that's 1 TFLOPS

Modern GPUs achieve 10+ TFLOPS in matrix multiplication

- in practice, the advantage is $\times 50$ due to better memory bandwidth.

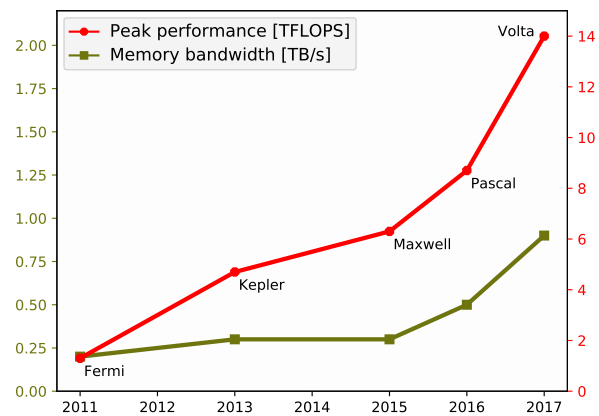
Price of training a simple ImageNet model:

GPU	TFLOPS	price	time
GTX 1070	6.5	4000 kn	52 h
GTX 1080 Ti	11.3	7000 kn	31 h
P100 FP16	25.0	47000 kn	13 h
DGX-1 FP16	170.0	$129000	2 h

Hardware: GPU

Best performance/price is obtained on gaming GPUs:

- NVIDIA Titan X: 250W, 12GB, 11 TFLOPS, 3000 GPU cores
- Radeon Vega FE: 300W, 16GB, 13 TFLOPS, 4000 GPU cores
- NVIDIA Titan V: 250W, 16GB, 110 TFLOPS (?), 12nm, 8cm^2, 21Gt



Actual non-representative measurement:

- GTX1070 (3500 kn) 60$\times$ faster than E3-1220 v3 (1700kn)

Additional challenge: deliver such performance with low power

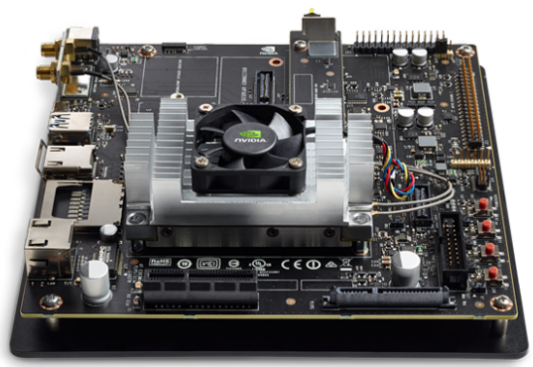
Hardware: embedded

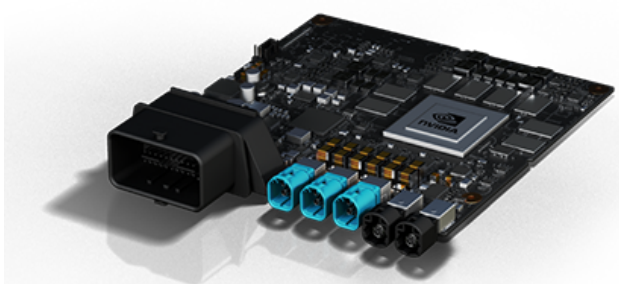
Novel hardware concept: processing units for artificial intelligence

- fast matrix multiplication (10 TFLOPS)
- low power, low precision (8-32 bit)

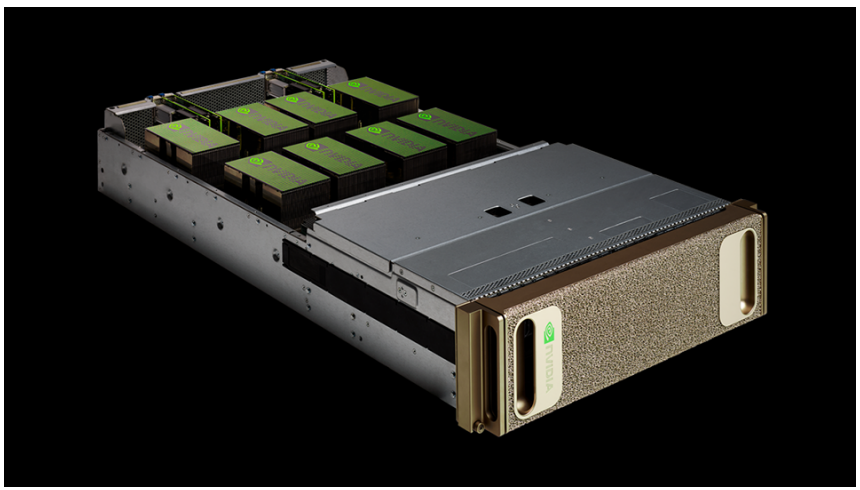
Main players:

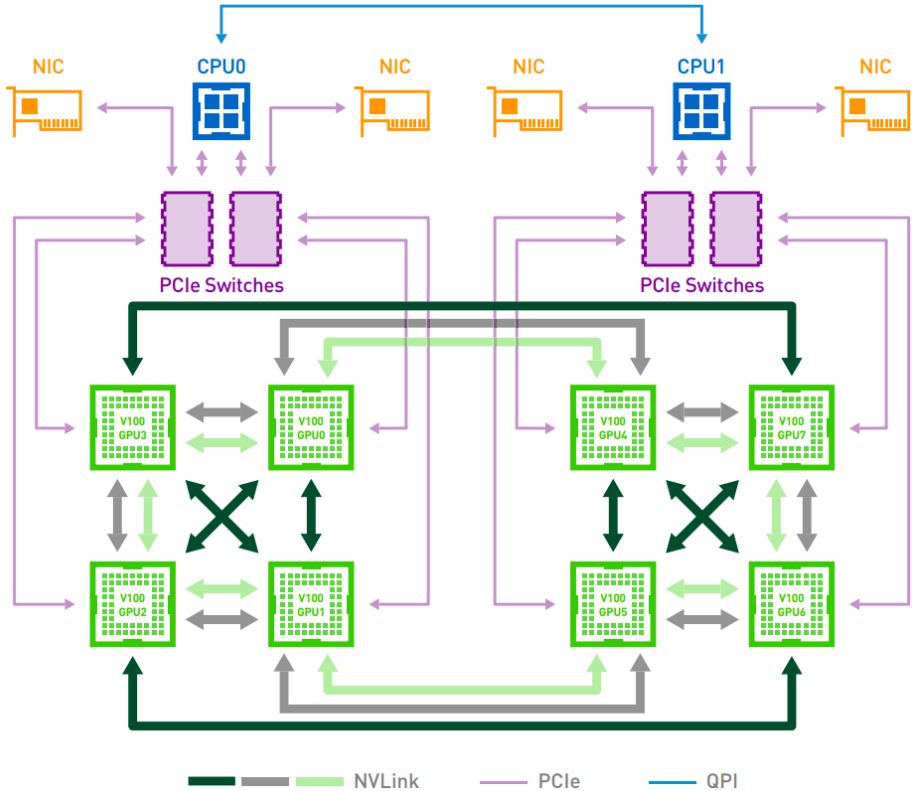
- NVIDIA TX2: 1.5 TOPS, 15W, 8GB RAM
- NVIDIA Xavier: 20 TOPS, 20W, automotive certificate (ISO 26262)
- Google TPU2: 45 TFLOPS
- Microsoft + Intel DPU (Stratix-10): 40 TFLOPS





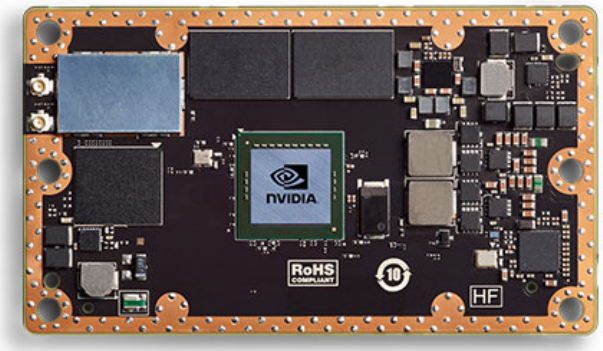
Hardware: datacenter



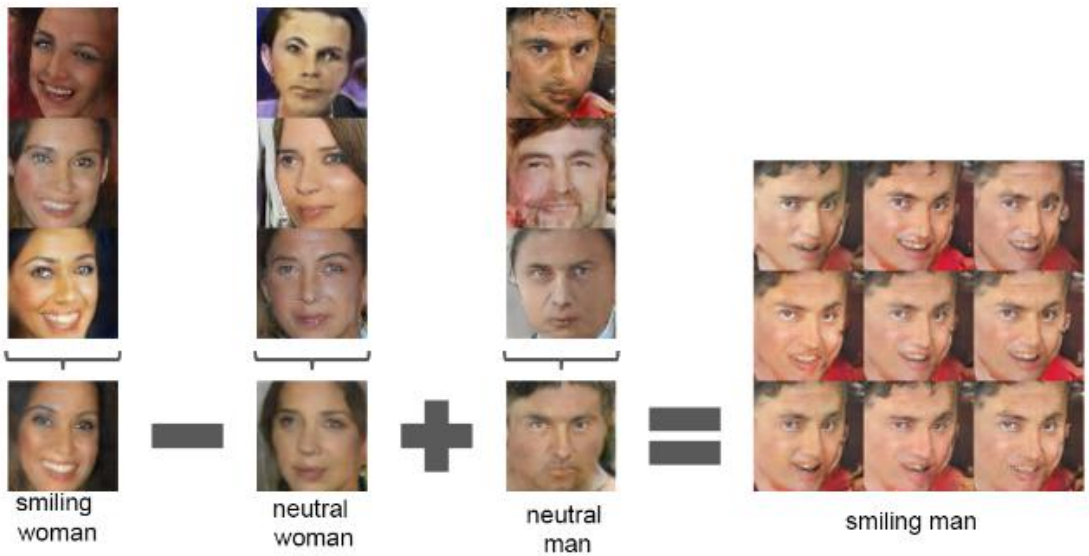


Trendovi

Ugradbene izvedbe (TX2, Xavier)



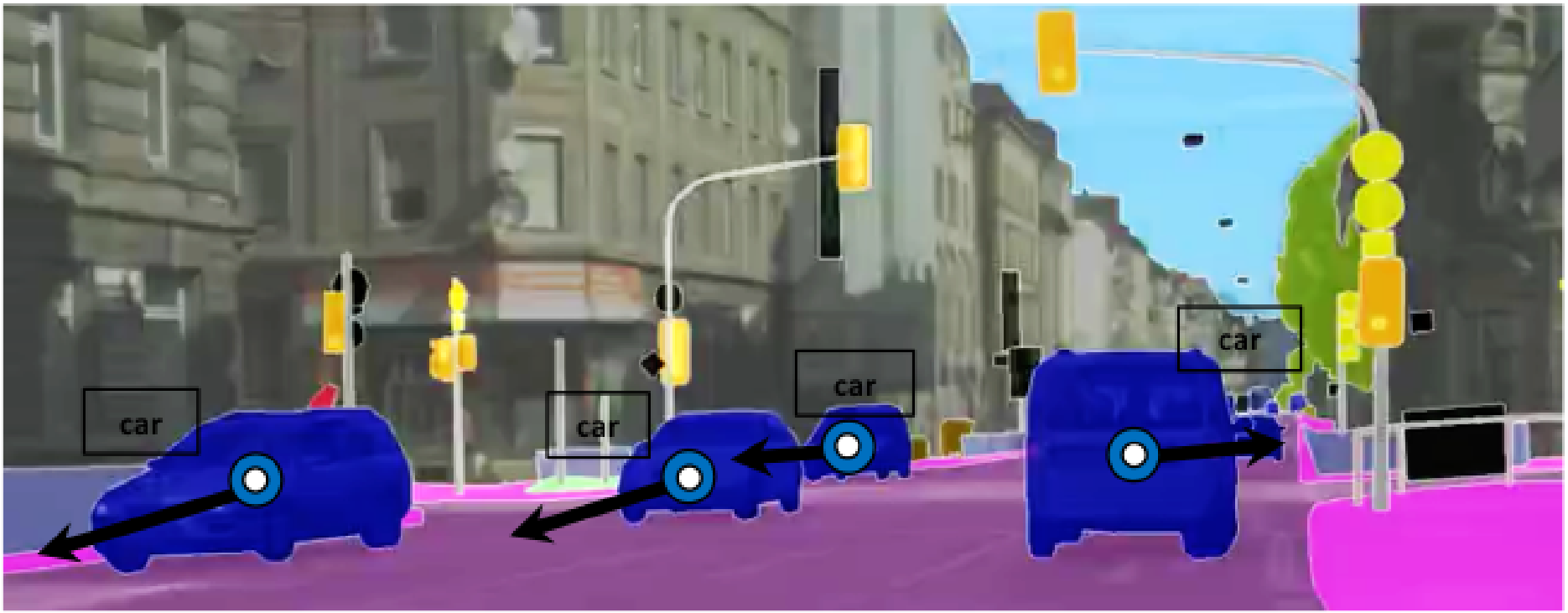
Nenadzirano učenje i generiranje slika:



[radford16iclr]

Trendovi

Predviđanje budućnosti:



[neverova17arxiv]

Trendovi

Prevođenje slika u tekst:

A close up of a person brushing his teeth.

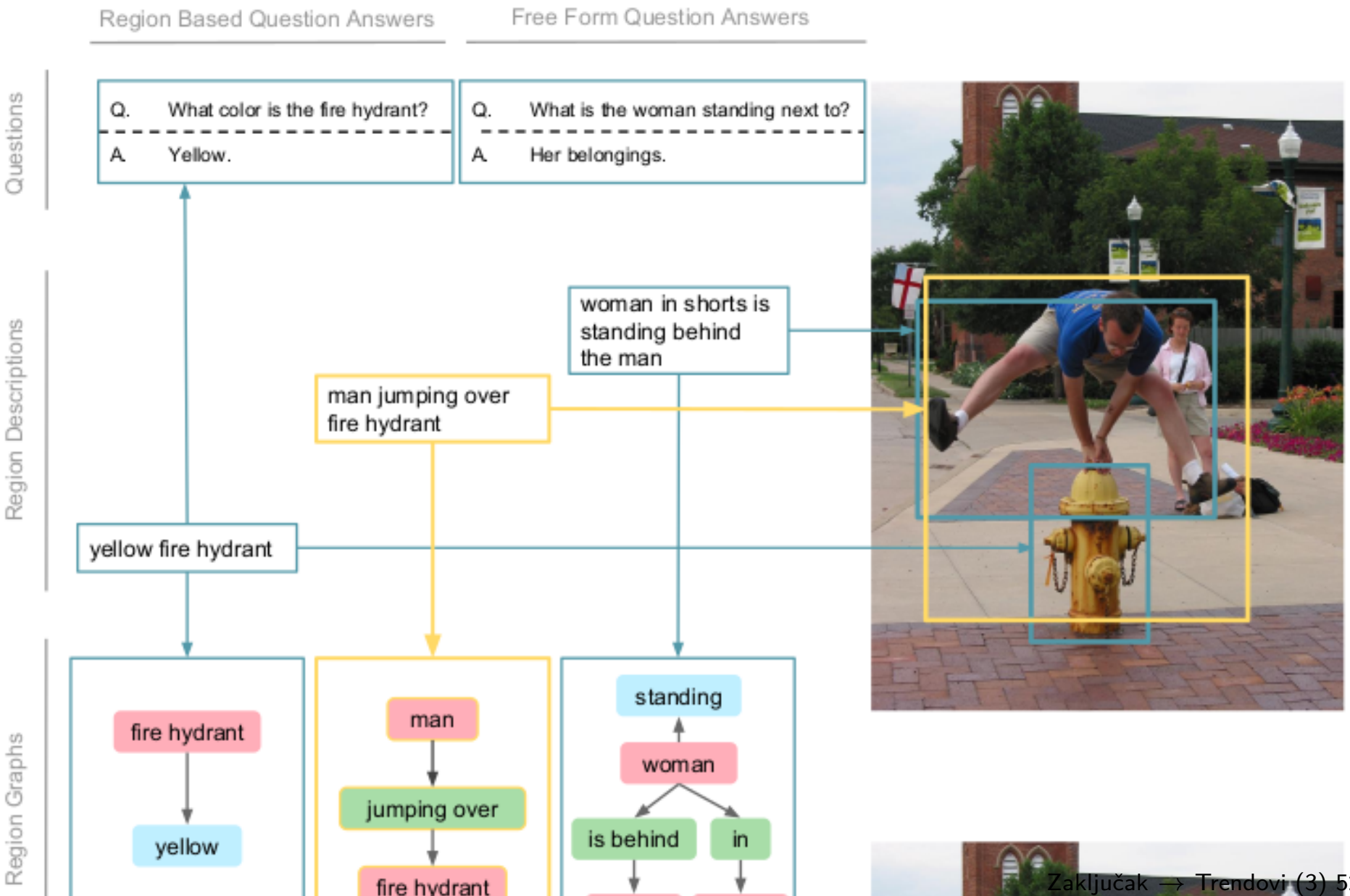
A woman laying on a bed in a bedroom.

A black and white cat is sitting on a chair.

[donahue16arxiv]

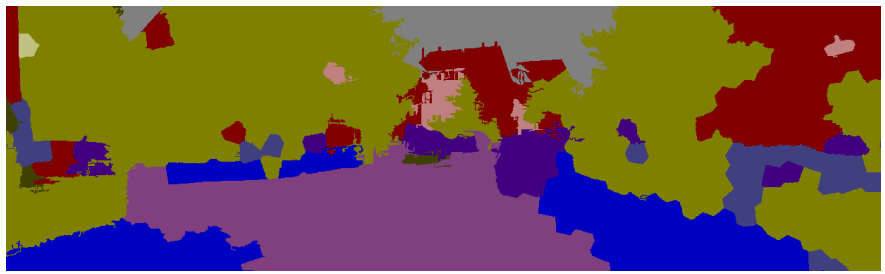
Trendovi

Razumijevanje slike



Conclusion: state of the art

Dramatic improvement of prediction accuracy in the last few years:

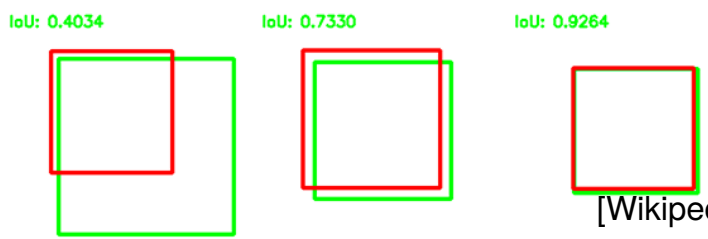


2015 [ros15wacv]

2017 [kreso17cvrsuad]

Cityscapes categories: 91% vs 89.7%

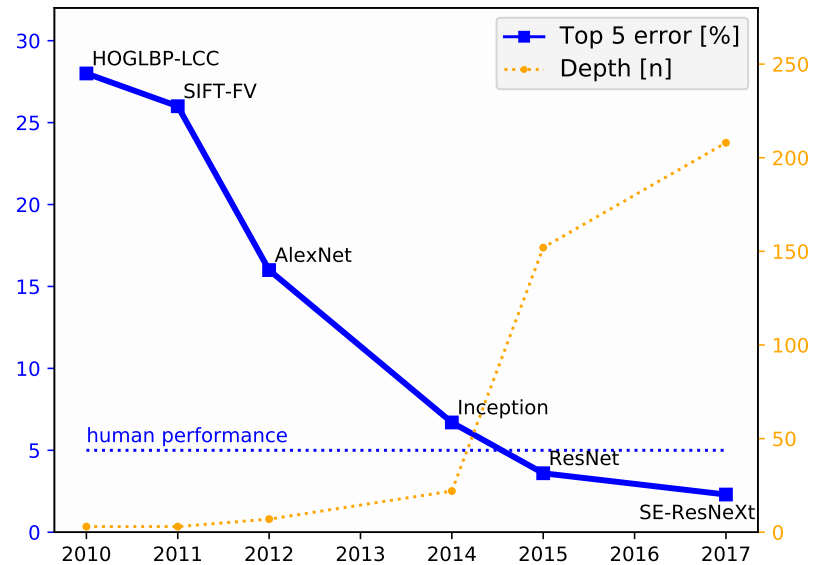
Cityscapes classes: 81.4% vs 78.4 %

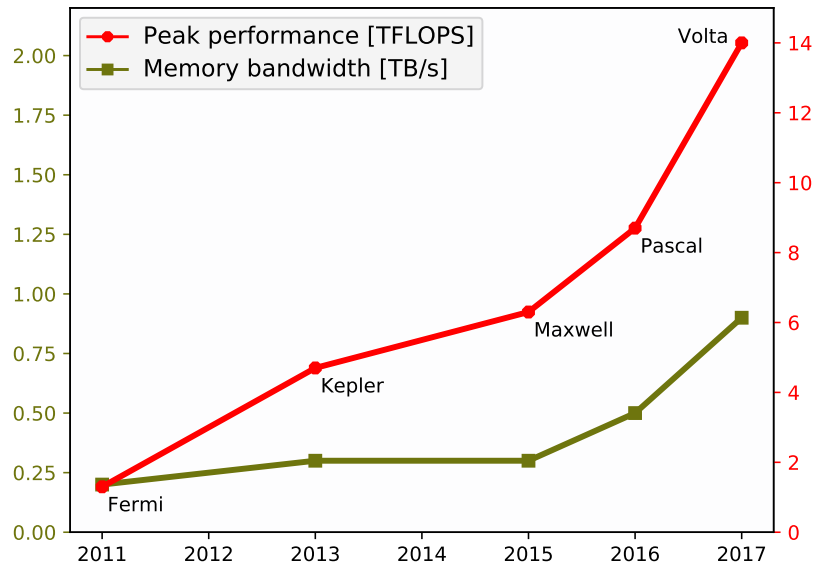


[Wikipedia]

Conclusion: outlook

Performance of computer vision systems will continue to grow





Important research directions:

- relaxing supervision (unsupervised, semi, weakly, multi-domain)
- estimating uncertainties (outlier detection, adversarial examples)
- new layers for visual recognition (transformers)
- low power inference (quantization, pruning, distillation)

ZAHVALA

Ova predavanja proizišla su iz istraživanja koje je financirala Hrvatska zaklada za znanost projektom I-2433-2014 MultiCLoD.

`http://multiclod.zemris.fer.hr`