Convolutional Neural networks

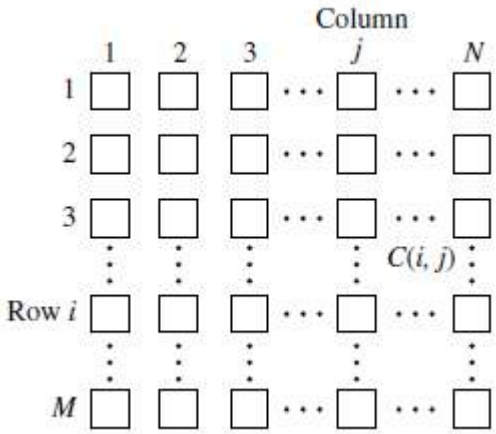
András Horváth

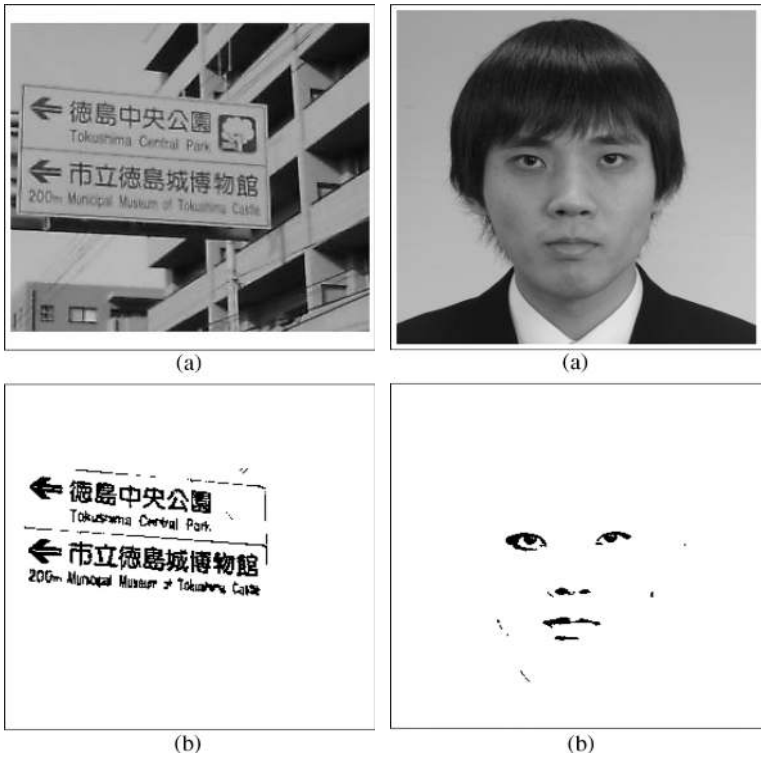
Mire jó egy CNN architektúra?

Turing teljes, de mit old meg hatékonyan?

2D → 2D reprezentáció. Térbeli és téridőbeli lenyomatok („feature”)-ök előállítása

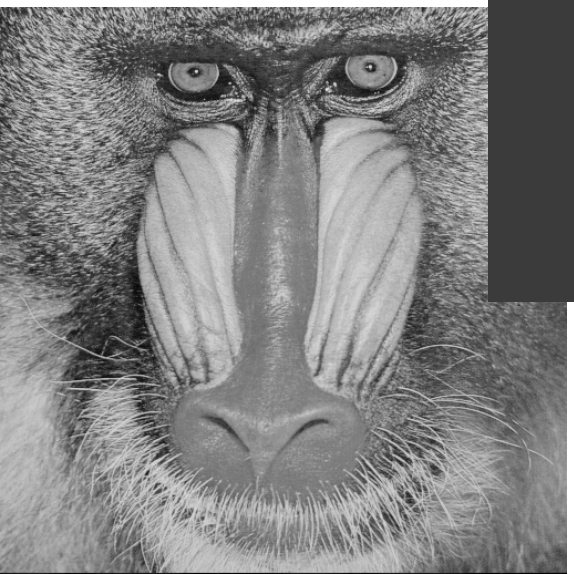
A standarad CNN architechtúra cellák M*N-es reguláris hálózata

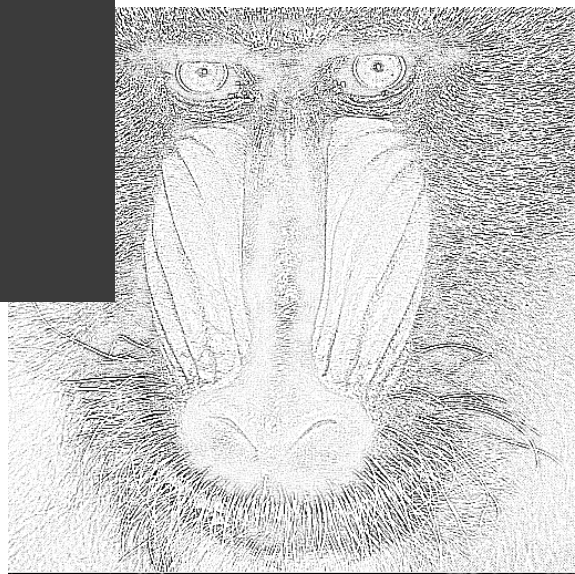




Térbeli lenyomat

Gradient





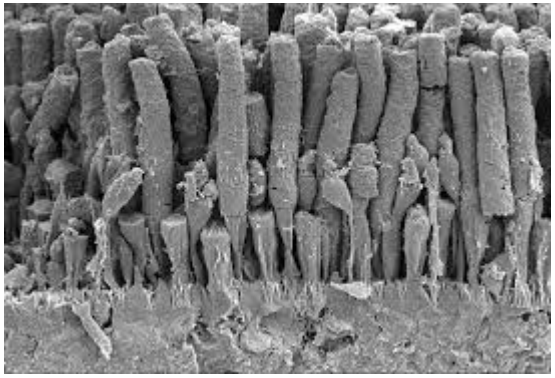
Retina felépítése

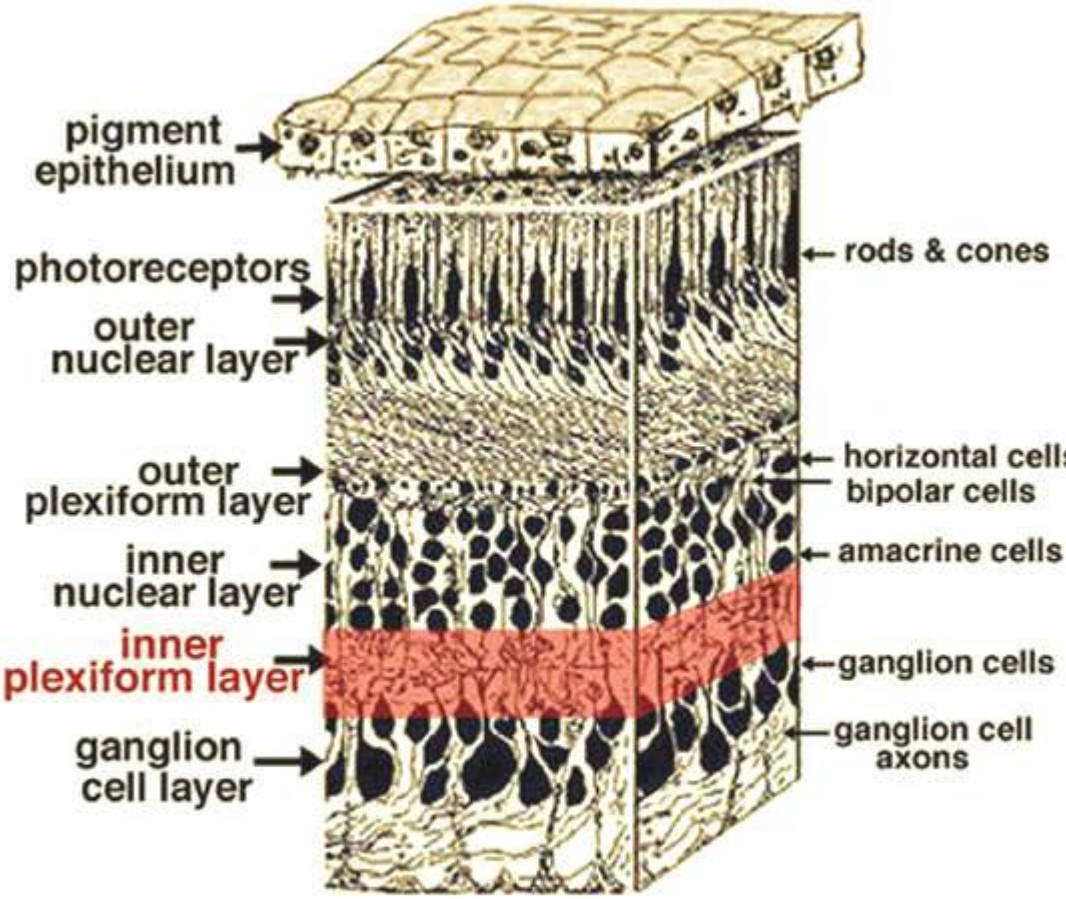
Gerinces retina: Inverz retina (ellentétben a fejlábúakkal, külön fejlődési ág) az érzékelő sejtek a leghátsó rétegben találhatóak

Csapok (cones):
6 millió, Erős megvilágítás mellett
Fovea-ban, színlátás

Pálcikák (rods):
120 millió, gyenge megvilágítás mellett
környéki látás, egy szín

Ganglion, amacrin, bipoláris, horizontális sejtek

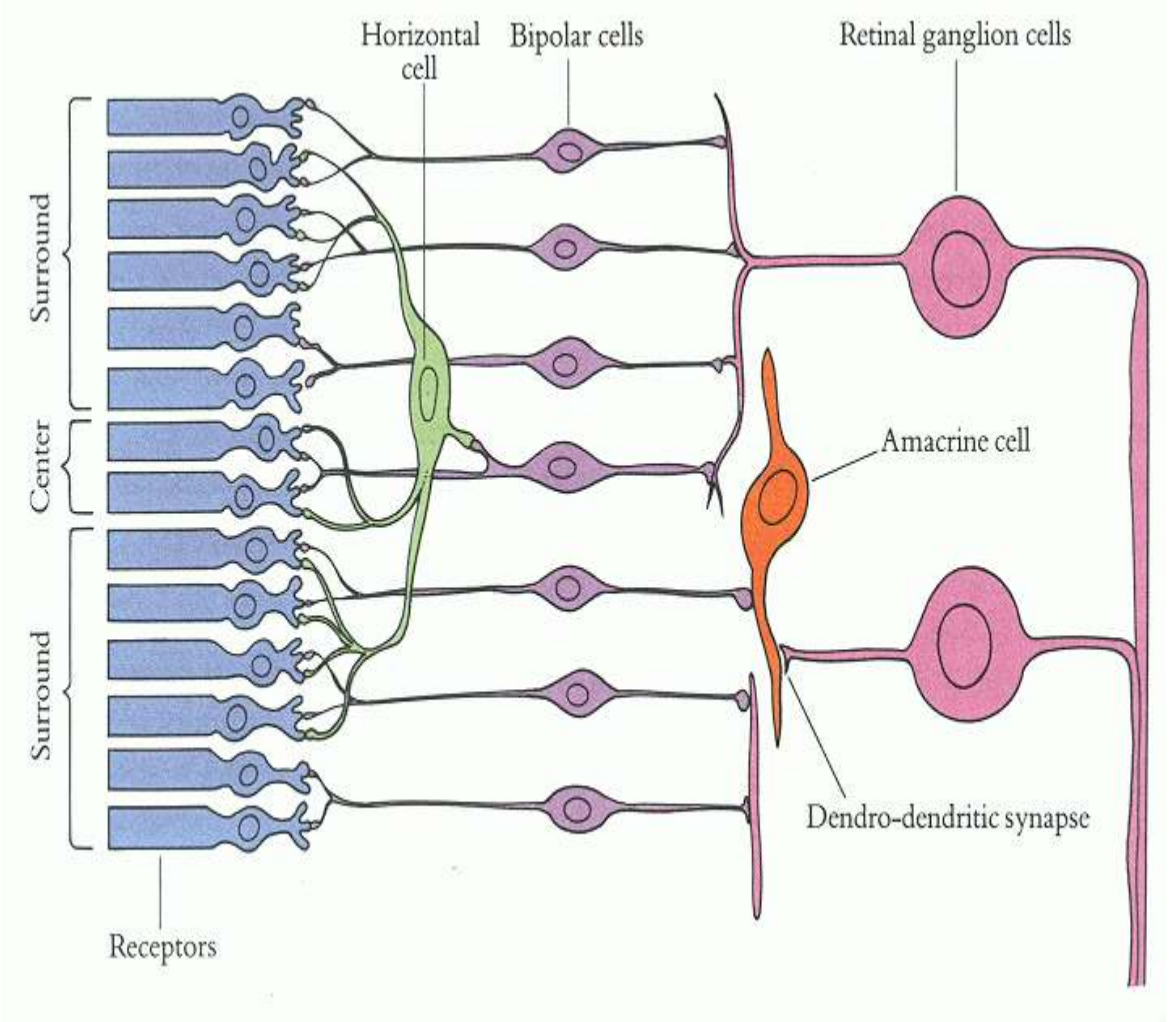




Hierarchikus struktúra a retinában

100x több fotoreceptor, mint ganglion sejt:

Nagymértékű tömörítésre van szükség
Nem pixel szinten látunk:
Területeket változásokat érzékelünk
Sárgafolt
On/Off
Serkentő, gátló régiók



Fides et Ratio

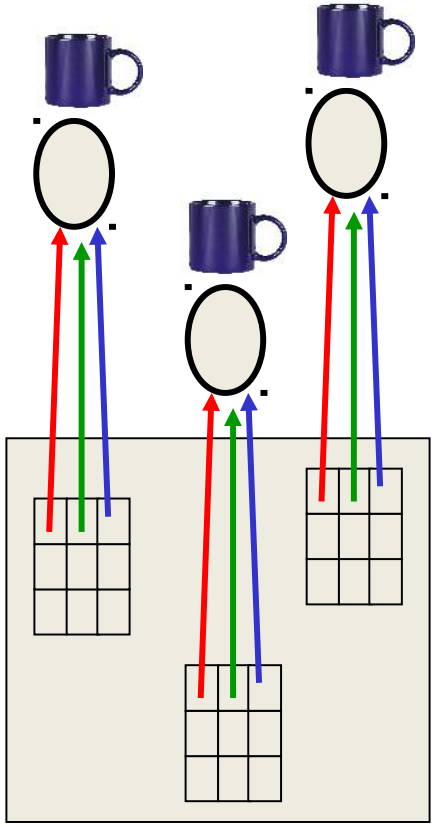
Mit várunk egy felismerő rendszertől? Ismételt lenyomatok (The replicated feature approach)

Nagyon népszerű neurálsi hálózatok esetében

És általában döntési folyamatokban

Ugyanazon lenyomat kinyerő struktúrát használjuk az adaton különböző 'helyzetekben'

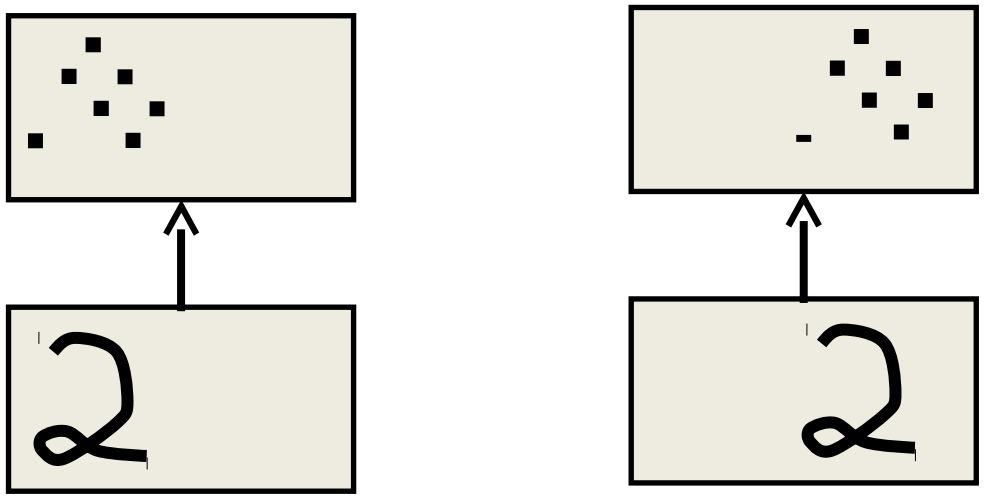
- Különböző pozíciókban azonos választ várnánk
- Azonos választ várnánk különböző méretű és orientációjú objektumokban (nehéz és számításigényes)
- Nagyon leegyszerűsíti a tanítást, sokkal kevesebb szabad paraméter lesz



Mit érünk el a lenyomat kiszámításának topografikus ismétlésével?

Nagyon leegyszerűsíti a tanítást, sokkal kevesebb szabad paraméter lesz

Invariant knowledge:A tanítás esetén ha egy régióban hasznosnak bizonyult egy feature és ezt a rendszer megtanulta,akkor az összes pozicióban hasznos lesz.



Ha a feature-ök detekcióját homogén CNN struktúrával hajtjuk végre, akkor a detekció biztosan invariáns lesz a transzlációra

Deep Learning

Általános feature hierarchiákat tanul meg

'Hagymonáyos tanulás' (Shallow learning) Különböző feature-ök tulajdonságok, melyik hasznosítható?

Kép. Pixel intezitás, élek ,fourier stb - A cél magának a reprezentációnak a megtanulása

Deep learning	vs.	Shallow learning
• Focus on entities and connections		• Focus on unrelated details
• Relates previous and new knowledge		• Information is simply memorized
• Uses reflection to relate theory with experience		• Facts and concepts accepted unreflectively
• Creates understanding, meaning and new ideas		• Aims to pass (or perform) instead of understanding
• Leads to positive emotions and attitudes about learning and self		• Leads to negative emotions and attitudes about learning and self

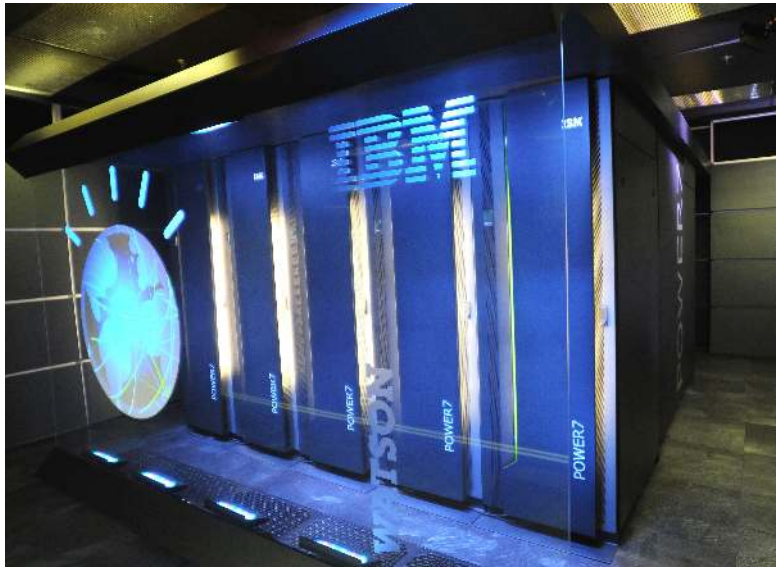
Deep Learning

IBM Watson

A kutatók célja az volt, hogy a természetes szövegértést javítsák a gépek esetében

Neurális hálózatot és deep learning-et használtak egy nagy adathalmazon

A gép megérti a számára feltett kérdéseket és számos esetben jobban teljesít, mint az emberek (Jeopardy)



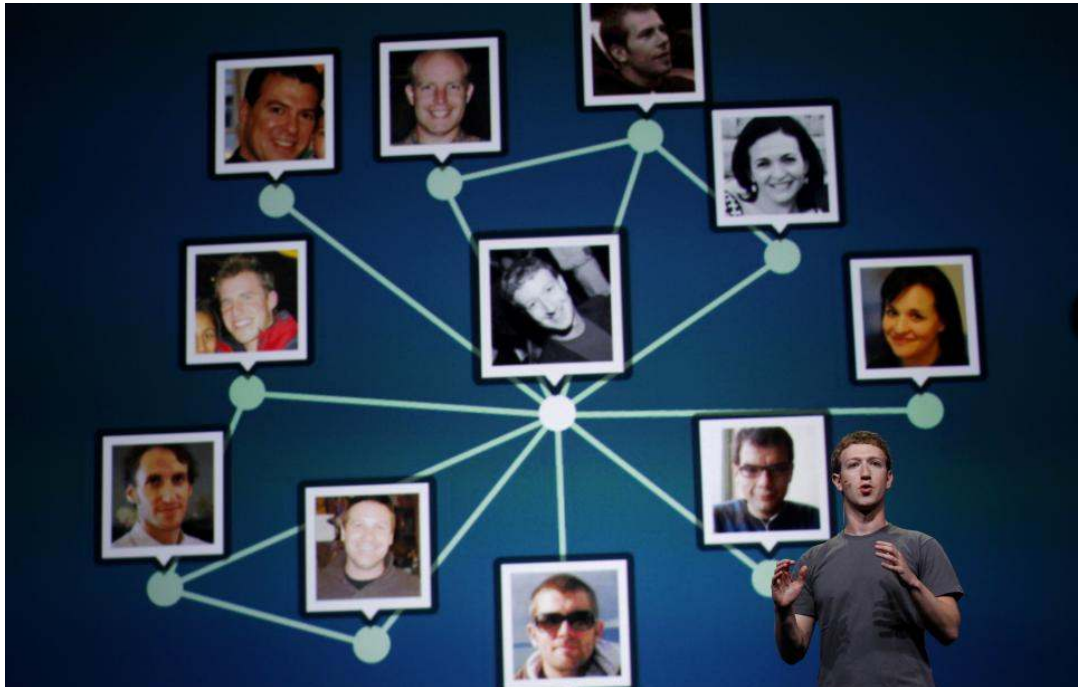
Deep Learning

Deep Face

Facebook arcfelismerő: DeepFace: Closing the Gap to Human-Level Performance in Face Verification

97.35%-so pontosság, egy átlagos ember 98%-os pontossággal dolgozik

Arcfelismerés Shallow learning használatával



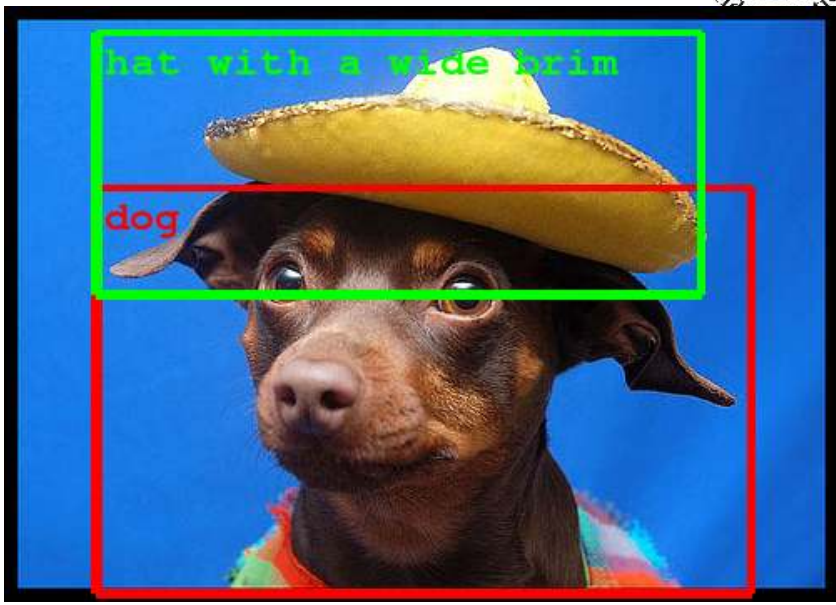
GoogLeNet

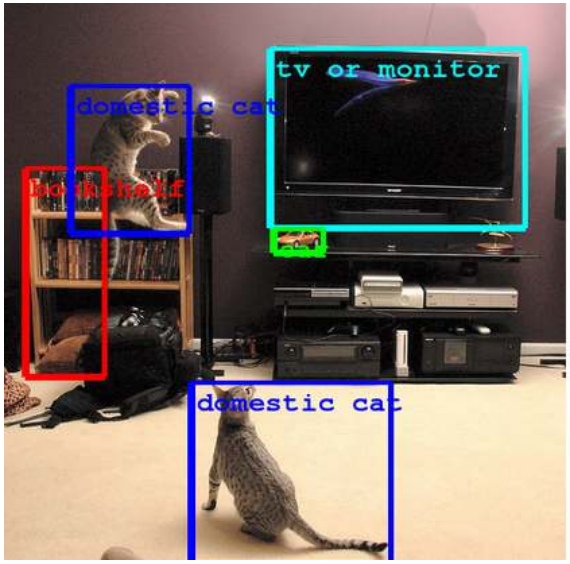
Általános objektum felsimerő algoritmus

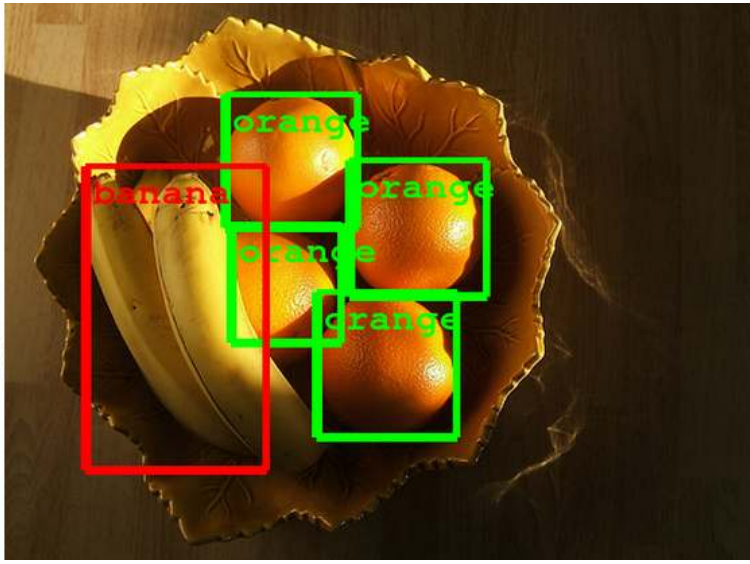
Az ImageNet adathalmazon tesztelve

Több, mint 200 különböző objektum felismerésére tanítva

37%-os pontosság

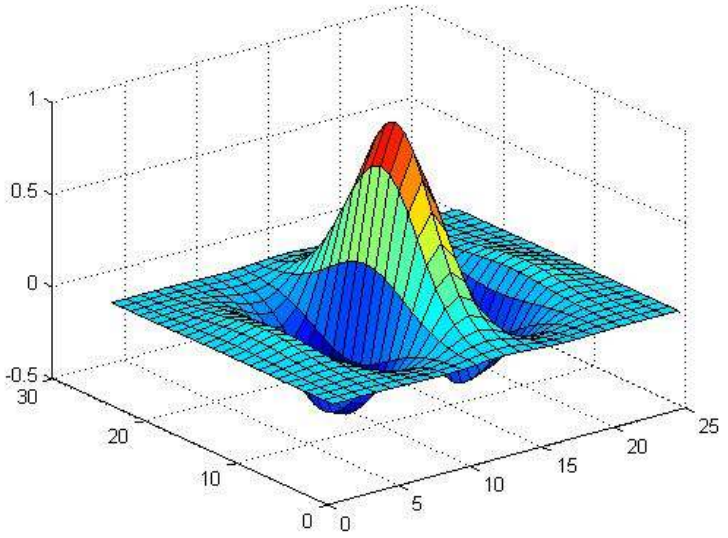






Általános lenyomatok előállítása

Az adat lokális stuktúráit szeretnénk felismerni
(akár térbeli akár időbeli): képek, beszéd esetében



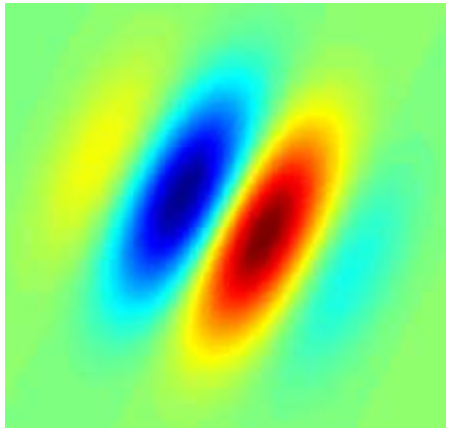
Mi a legáltalánosabb oylan tulajdonság, ami:

- lényeges a detekció szempontjából
- A legtöbb feature kikeverhető belőle
- Invariáns az értékek nagyságára

Ezek általában egyszerű változásdetektorok

Gábor kernel-ek (Egy egyszerű feed-forward template
- csak B template-tel számolható)

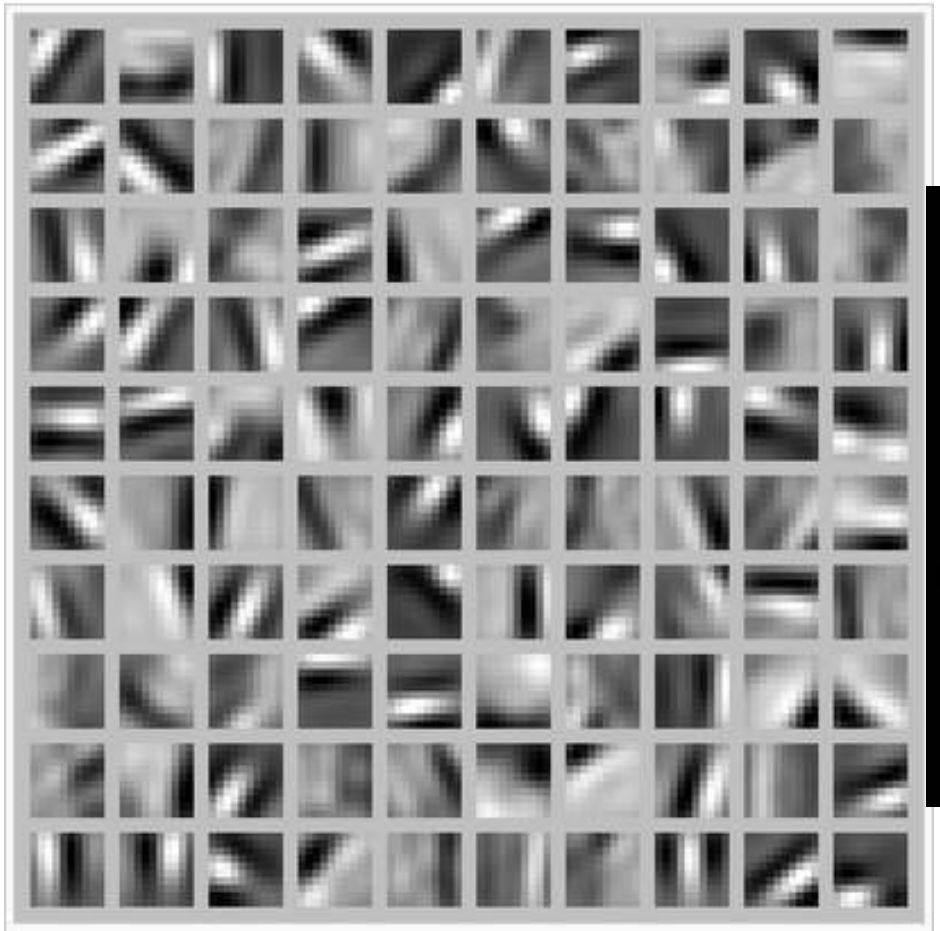
$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \exp\left(i\left(2\pi\frac{x'}{\lambda} + \psi\right)\right)$$

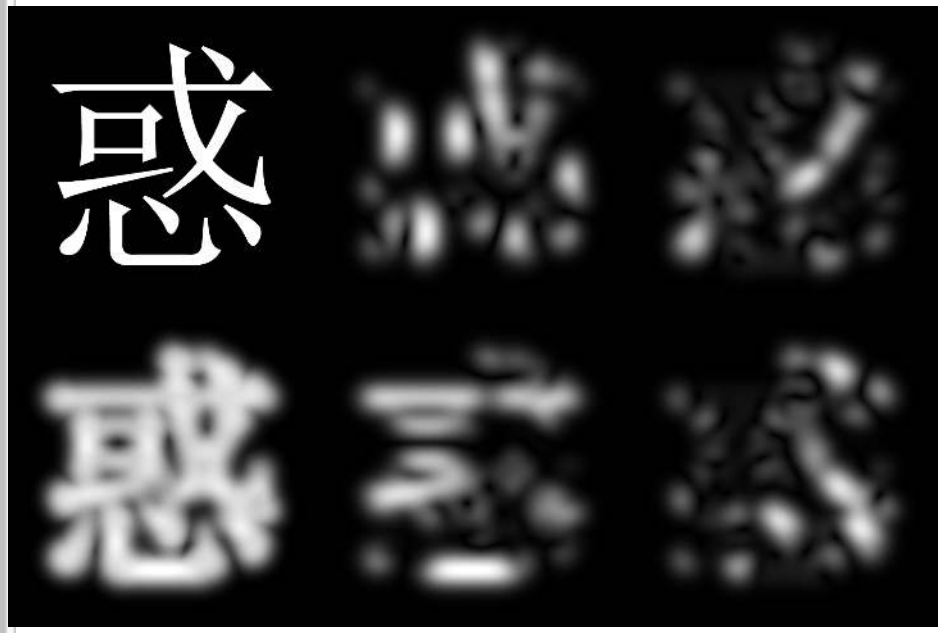


Deep Learning

Általános feature hierarchiákat tanul meg

Képek esetében ezek a következőképpen néznek ki:





Neurális hálók használata detekcióban

Kunihiko Fukushima (1980) – Neo-Cognitron

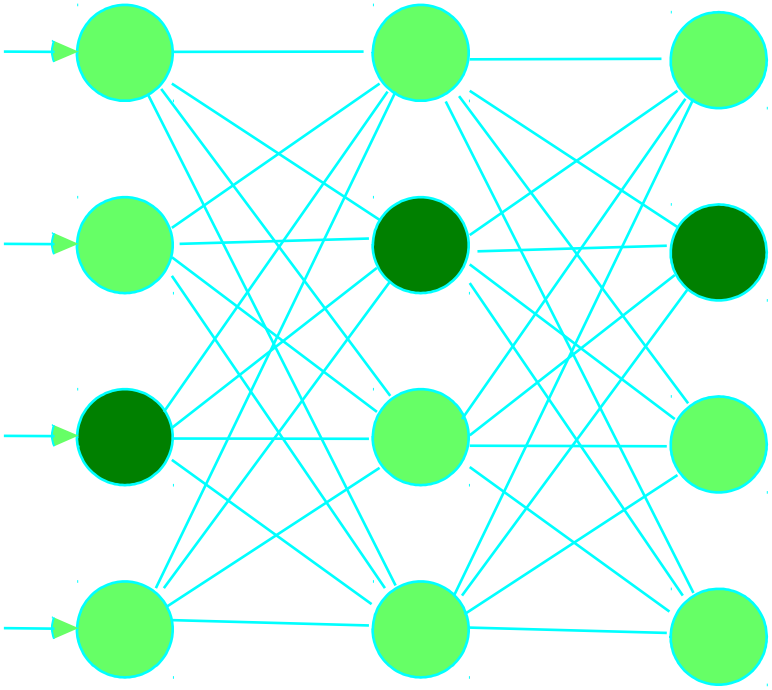
Yann LeCun (1998) – Convolutional Neural Networks

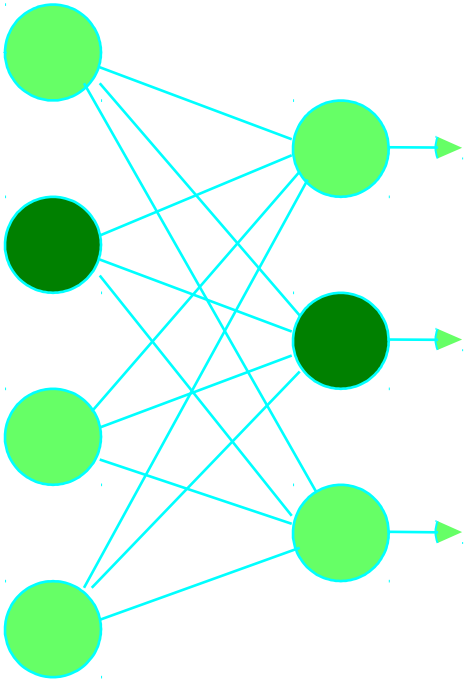
MLP (multilayered perceptron) backpropagation-nel

- Biztató kezdeti eredmények, de megállt jóval az emberi képességek szintje alatt
  - Rendkvül lassú tanítás (gyorsabb felismerés), bonyolult hálózatok
  -

Hogyan lesz döntés a lokális feature-ökből?

Hagyunk-e egy általános struktúrát,ahol sok paraméterünk lesz és nehéz tanítani, vagy élünk-e megkötésekkel az architektúrára nézve?





Mintavételezés - pooling

Hogyan tömörítsük/alakítsuk döntéssé a térbeli lenyomatot?

Mintavételezés pooling

Két doglot szeretnénk elérni:

- Magasabb szinten a feature-ök magasabb szintjén reprezentálni (pl első szint knotraszt/élek), második szint élek egymáshoz viszonyított elhelyezkedése stb...
- Kismértékű transzlációra magasabb szinteken is ugyanazt a választ várjuk

  Pooling – egy területen közelítsük a lenyomatot statisztikai leírókkal (a pixelek átlagával vagy maximumával)
  - Ez nagymértékben lecsökkenti az elemek számát a következő szinten.
- Problem: Nagyon sok pooling után elveszítjük a pontos térbeli reprezentációt

Convolutional Neural networks

C (Convolution) és S rétegek (Pooling) váltakozása (Nemlinearitások alkalmazása) a végén egy Multilayer Perceptron a felismeréshez

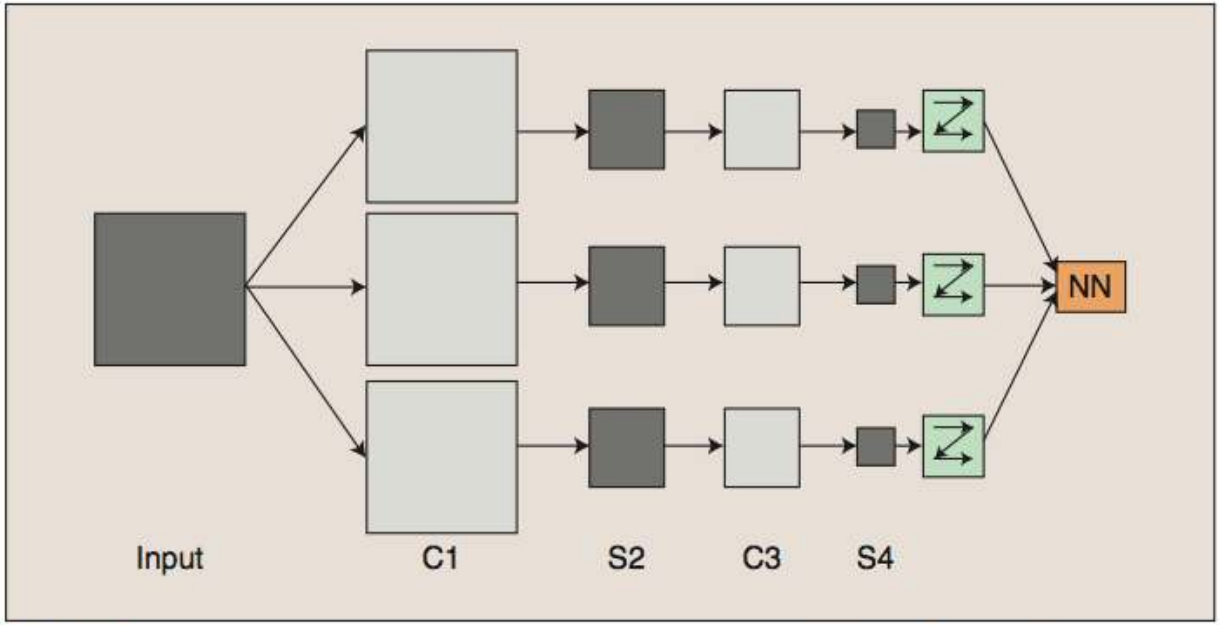
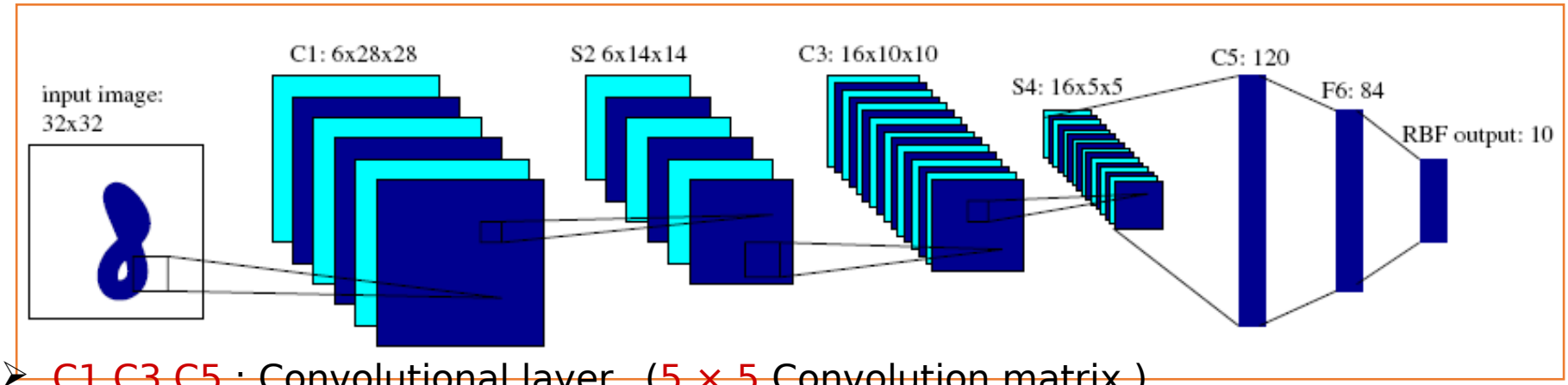


FIGURE 2 Conceptual example of convolutional neural network. The input image is convolved with three trainable filters and biases as in Figure 1 to produce three feature maps at the C1 level. Each group of four pixels in the feature maps are added, weighted, combined with a bias, and passed through a sigmoid function to produce the three feature maps at S2. These are again filtered to produce the C3 level. The hierarchy then produces S4 in a manner analogous to S2. Finally these pixel values are rasterized and presented as a single vector input to the "conventional" neural network at the output.

Convolutional Neural networks

Egy viszonylag sok rétegű hálózat tanítása általános feature-ökkel
„shallow learning” esetén összetett feature-ök ,de 1-2 rétegű hálózat
A későbbi rétegek egyre bonyolultabb feature-öket írnak le:

- Első réteg elék, kontrasztok leírása
- 2nd layer learns higher order features (combinations of first layer features, combinations of edges, etc.)
- A rétegeke felügyeleten tanulással (unsupervised learning) így (az adaathalmaztól függően) általános feature-ök keletkeznek
- Az utolsó réteg tanítása felügyelt, itt megmondjuk, hogy miylen kimenetet szeretnénk egy adott bemenetre



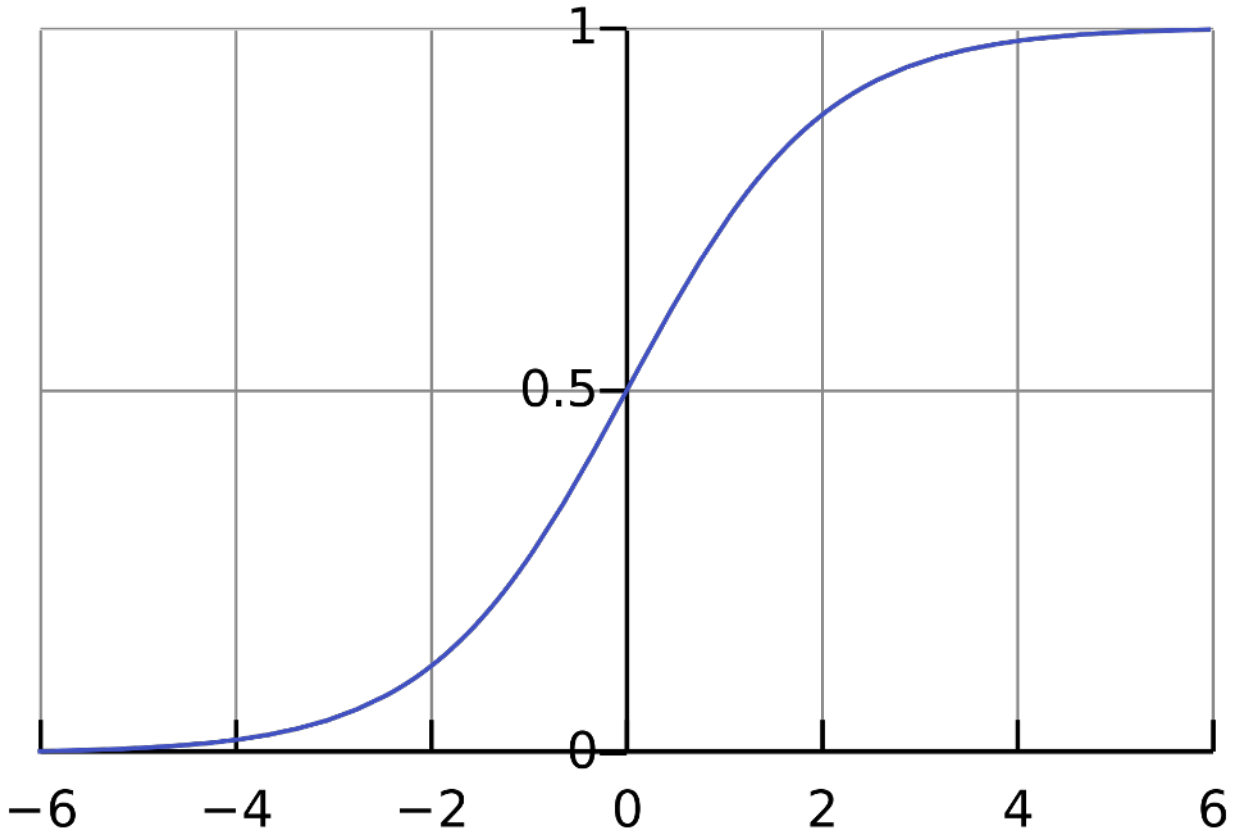
- C1,C3,C5 : Convolutional layer. (5 × 5 Convolution matrix.)
- S2 , S4 : Subsampling layer. (by factor 2)
- F6 : Fully connected layer.

- 187,000 összeköttetés
- 14,000 állítható súly (10X)

A rétegek között használt nemlinearitás

A rétegek közötti értékek korlátosságának megőrzésére egy a standard CNN dinamikához hasonló nemlineraitást használnak

$$f(x) = \frac{L}{1 + \mathrm{e}^{-k(x - x_0)}}$$



Ugyanezt a nemlinearitást használják CNN dinamikákban is (feature kinyerés szempontjából nincs kvalitatív különbség a kettő között)

Hasonlít a biológiai struktúrákhoz

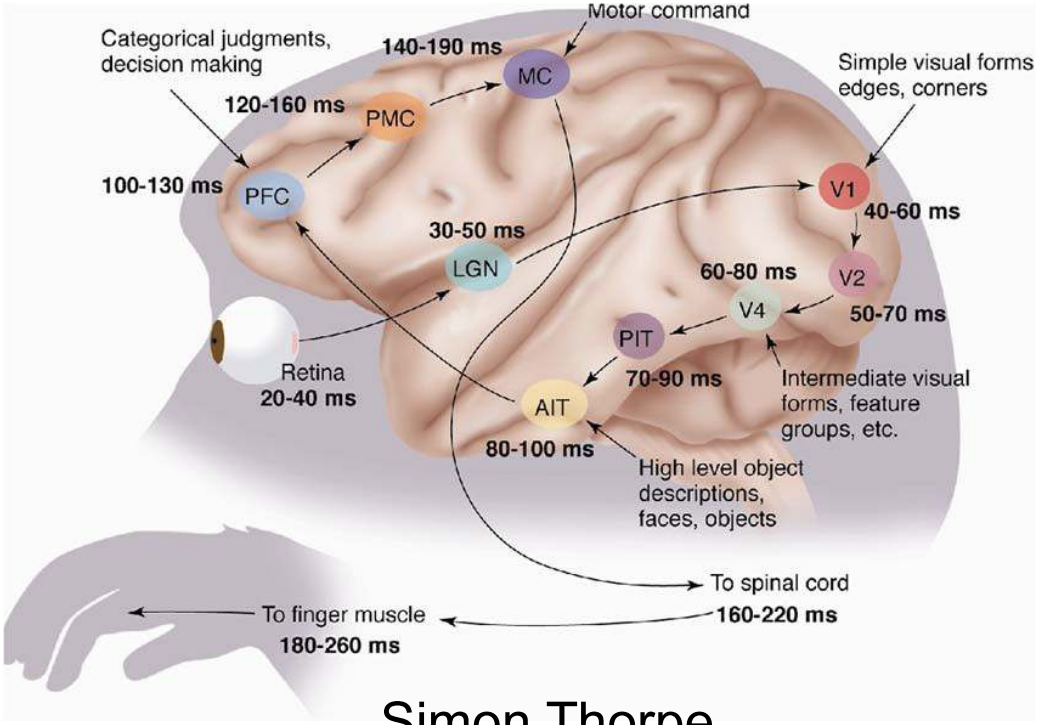
Hierarchikus feléptés az agyban önamguktó lszervező struktúrák

- Topografikus területeken – és feature számítások (retina – V1)
  Mintavételek ezen területekről és komplex döntéshozó rendszerek

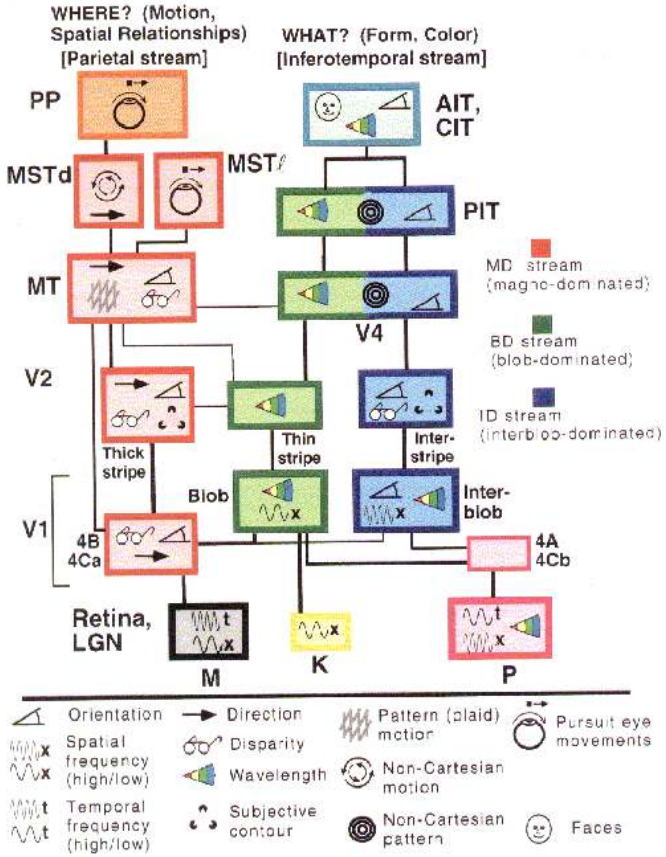
Filozófiai/evolúciós vonatkozás:
Az emberi agy marad legtovább plasztikus a születés után, s rendkívül hosszan megőrzi ezt a képességét a többi főemlőssel összehasonlítva
- ezáltal egy rugalmas, könnyen alkalmazkodó, újrahangolható rendszert kapunk
- Ez a rugalmasság védtelenséggel is jár, s akoari időszakban nagyobb védelemre van szükség

Hasonlít a biológiai struktúrákhoz

The visual cortex is hierarchical – the ventral (recognition) pathway has multiple stages

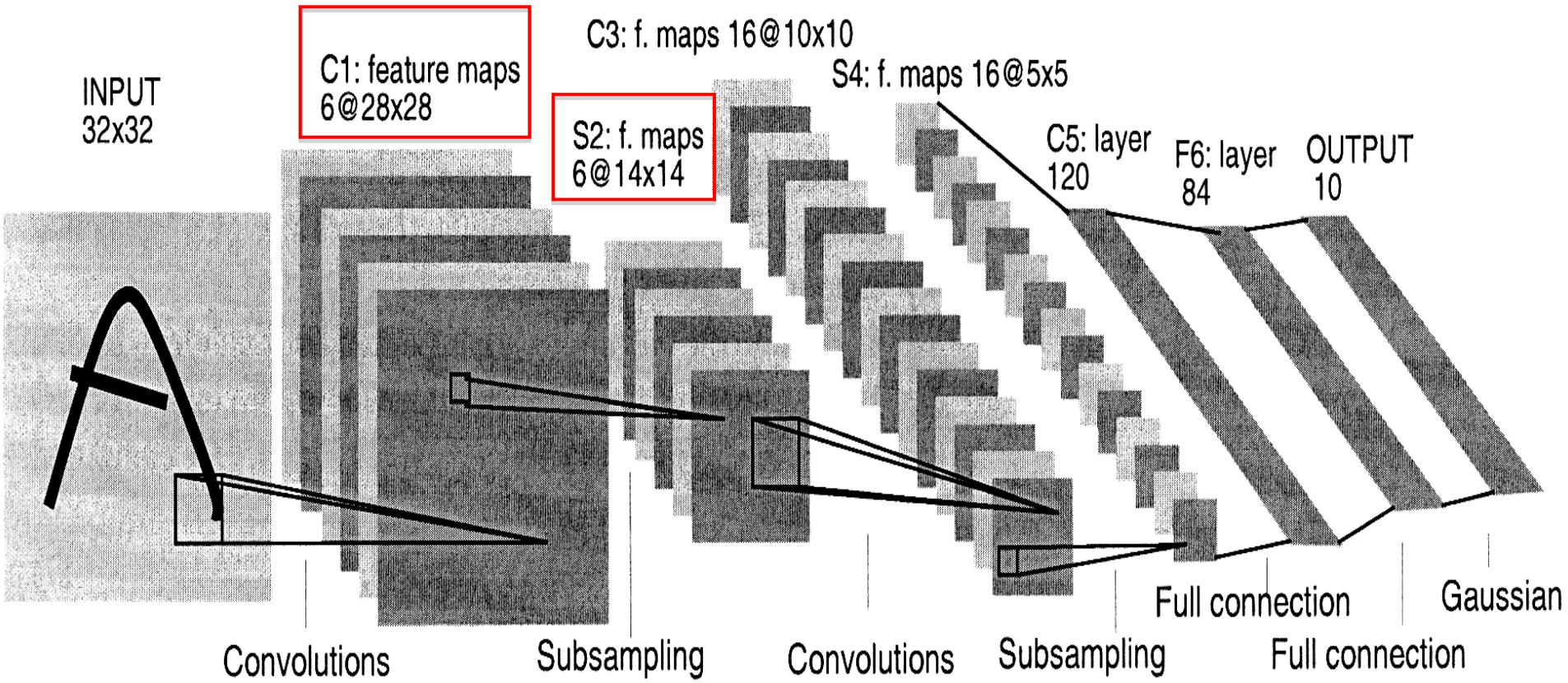


Simon Thorpe



Gallant & Van Essen

Egy gyakorlatban használt ConvNN struktúrája

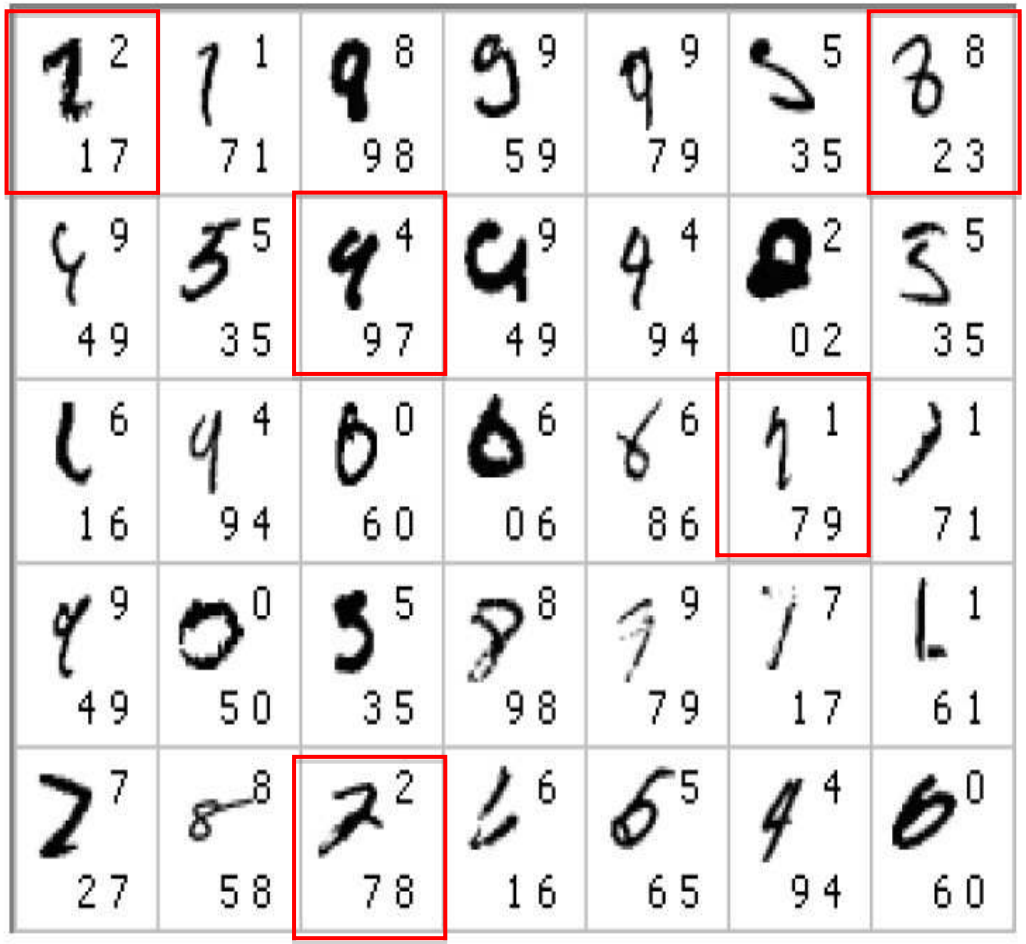


A rendszer által hibásan osztályozott elemek

Összesen 82 hibás detekció

Az emberi hibák száma ezen a halmazon nagyjából 20-30 közötti

A rendszer által hibásan osztályozott elemek

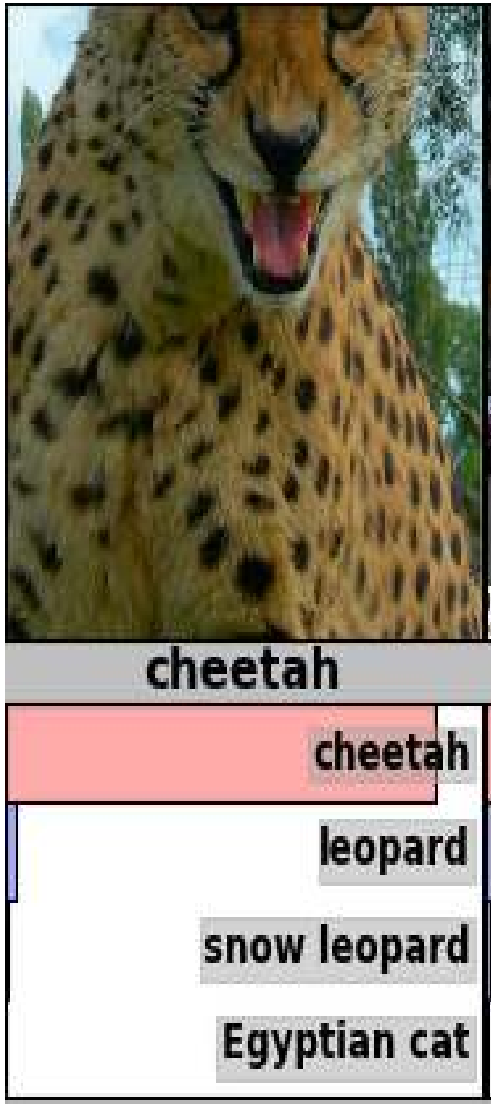


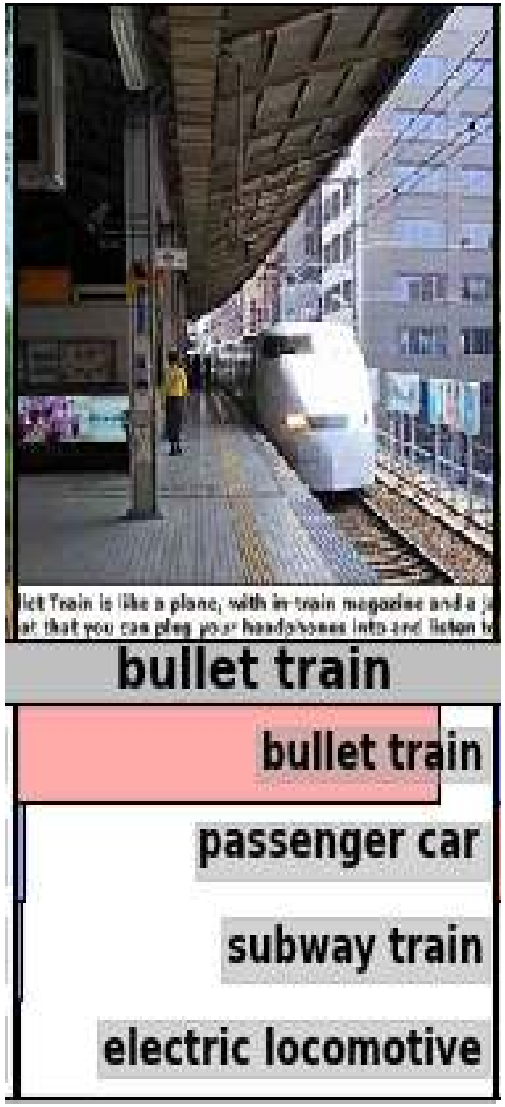
A legfelső sorban található a legnagyobb súlyú válasz
Az alsó sorban a következő két válasz látható

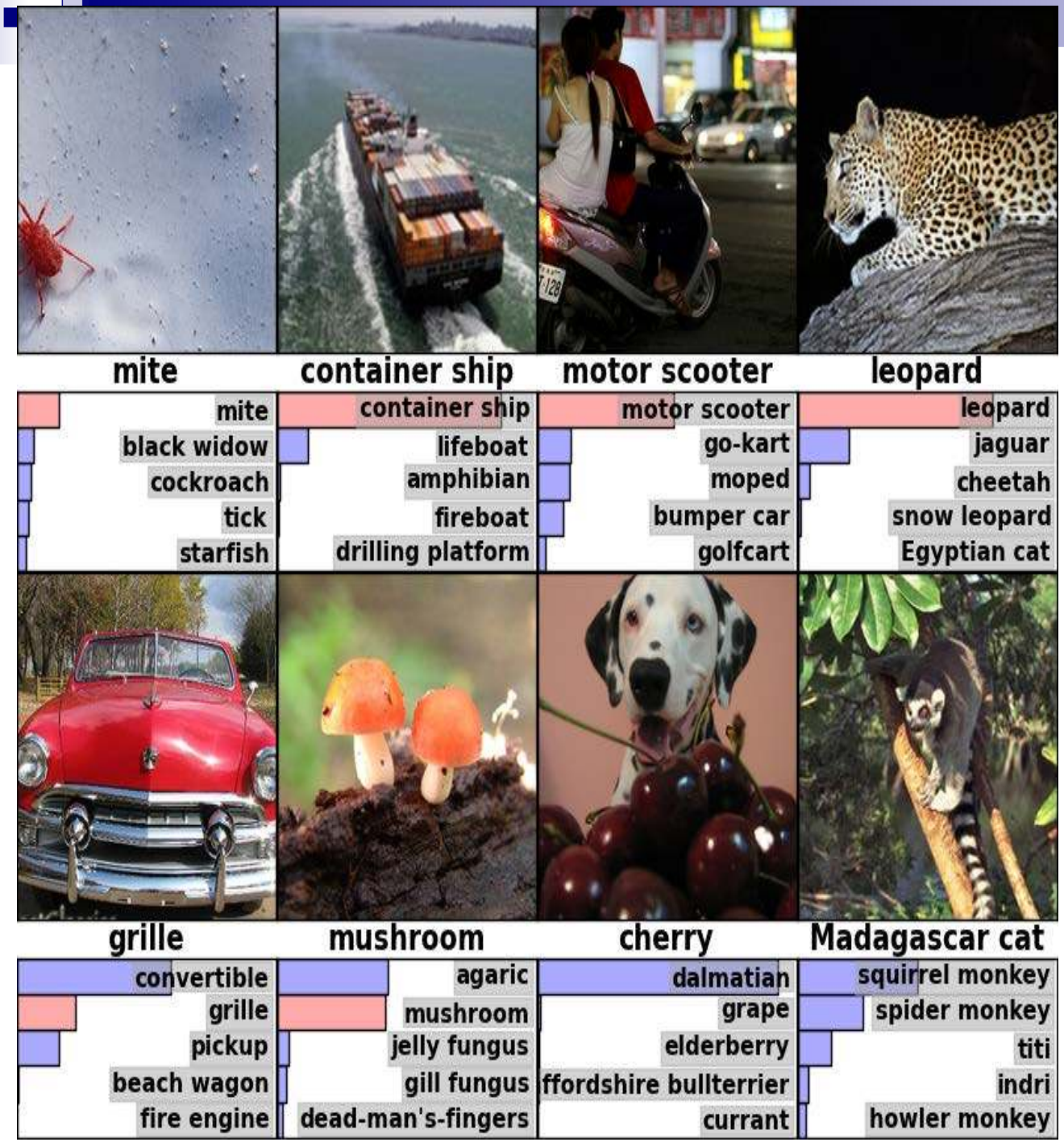
A helyes válasz szinte mindig megtalálható az első három legvalószínűbb válaszban

Jelenleg 20-25 közti hibaszámot érnek el ezen az adathalamzon ~ megegyezi az emberi felismerés potosságával

Néhány példa a GoogLeNet esetében

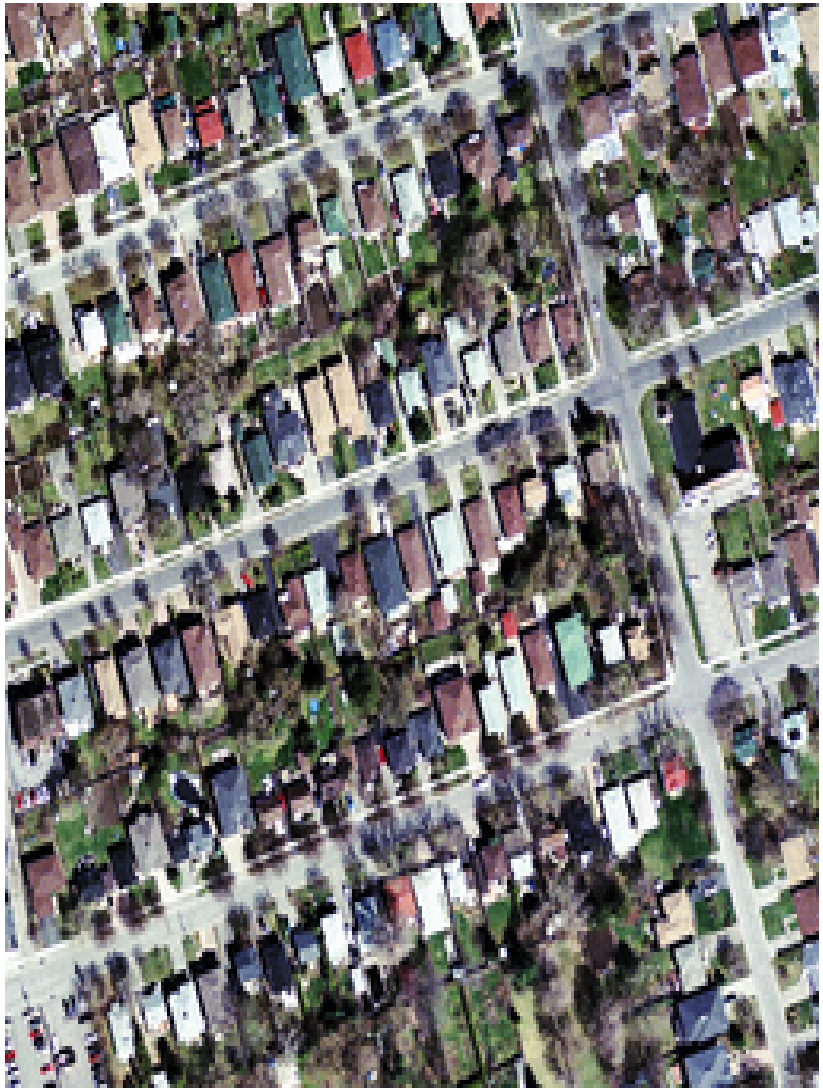






Sok esetben a hibákból látható, hogy a rendszer valóban általános tulajdonságokat tanul meg

Jelenleg a legjobb út azonosító algoritmus



A ConvNN hardware támogatás

A sokmagos architektúrákon, különösen GPU-k esetében ezek a számítséok könynen futtathatók (az architektúra konvolúciók számításaára lett kitalálva SIMD)

Nvidia GTX 580 Graphics Processor Units (több, mint 1000 mag)

- Konvolució, mint mátrix-mátrix szorzás hatékonyan számolható.
- GPUs nagy sávszélesség a memóriához
- A tanítás sokszor meglehetősen lassú → 1 hét is lehet.

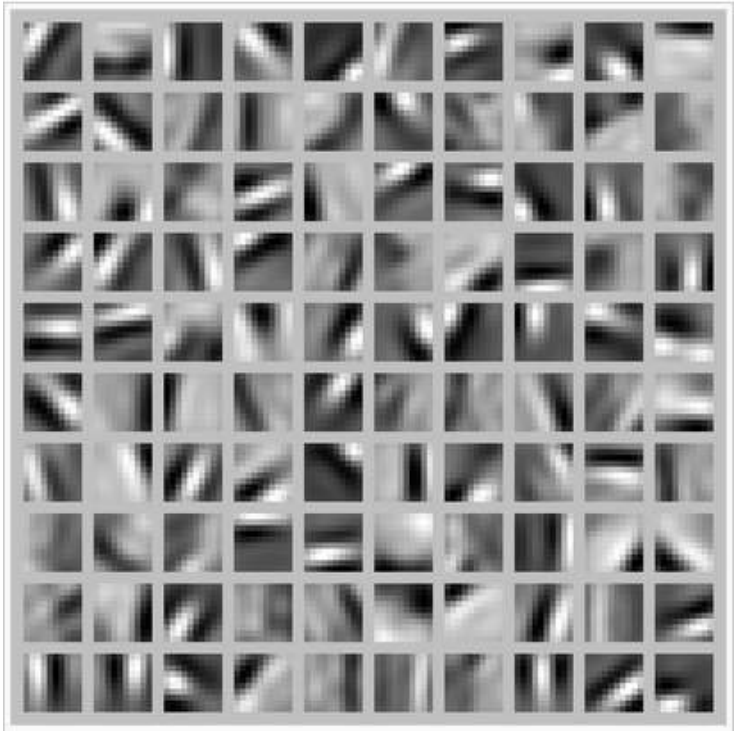
A felismerés 'néhány' onvolucióval kiszámolható → gyakorlatilag valós időben

Fides et Ratio

Sparsity - ritkaság

Egy Sparse reprezentáció segít a tanulásban. Ha már a közbünső rétegekben figyelembe vesszük, hogy 'saprse' reprezentációt szeretnénk – azaz csak néhány feature esetében akarunk nagy választ, akkor jelentősen javíthatjuk a detekciót.

Az itt látható lenyomatok közül ezzel specifikusabbakat választhatunk ki, s érezhető, hogy jobban használható egy oylan minta,ahol 1-2 erős választ kapunk, s a többi csatorna válasza sszinte mind nulla, mint egy olyan, ahol minden csatornán találunk választ.



Ez a 'sparsity' az emberi retina csatornákban is megtalálható