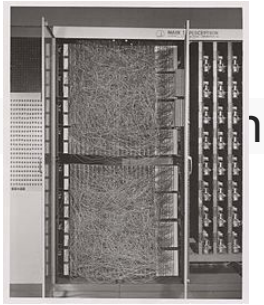
Deep learning

bevezetés

Egy kis történelem - a kezdetek



- 1957 - Frank Rosenblatt: Perceptron
  - A perceptron algoritmus első implementációja a Mark [illegible]n gép
  - 20×20 pixeles képet adó kamerához volt kötve
  - Hatására aktív kutatás folyt róla a 60-as években
- 1960 - Widrow and Hoff: ADALINE
  - Különbség a perceptronhoz képest a tanítás során van (delta szabály)
- 1969: Minsky és Papert: Perceptrons c. könyve
  - A perceptronnak súlyos korlátai vannak a képességeit tekintve, Frank Rosenblatt jóslatai nagymértékben túlzóak
  - Hatására kb 10 évig nem történt kutatás a témakörben (AI Winter)

Egy kis történelem - klasszikus neurális hálózatok

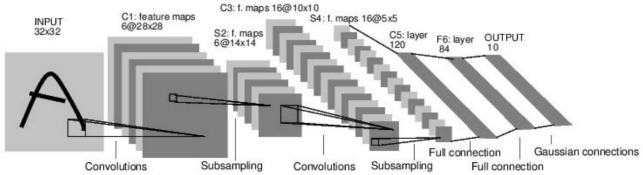
- 1980 - Fukushima: Neocognitron
  - Konvolúciós neurális hálózat elődje
- 1982 - John Hopfield (fizikus): Hopfield hálózat
  - Asszociatív memória, felügyelet nélküli tanítás
  - Energia alapú modellek
- 1985 - ACKLEY, HINTON, SEJNOWSKI: A Learning Algorithm for Boltzmann Machines
  - Valószínűségi eloszlások modellezése neurális hálózattal
  - Energia alapú modell
  - Folytatásaként Restricted Bolzmann Machines, Belief Net és Deep Belief Net, stb...
- 1986 - David E. Rumelhart, Geoffrey E. Hinton and Ronald J. Williams: "Learning representations by back-propagating errors"
  - Hibavisszaterjesztés algoritmusa
  - Újra beindította a kutatást a témakörben

Egy kis történelem - klasszikus neurális hálózatok

- 1989, Yann LeCun et al. (AT&T Bell Labs): ”Backpropagation Applied to Handwritten Zip Code Recognition”
  - Konvolúciós réteg
- 1989 - Alvinn: An autonomous land vehicle in a neural network
- 1989 - Waibel et. al (including Hinton), Phoneme recognition using time-delay neural networks
  - Időfüggő hálózat
  - Visszacsatolt hálózatok megjelenése az alkalmazási területen
- 1991 - Hochreiter (diplomamunka)
  - Sok rejtett rétegű NH-ban eltűnő/felrobbanó gradiensek problémája
- 1993 - Bengio: A Connectionist Approach to Speech Recognition
  - Visszacsatolt neurális hálózatokban hosszútávú memória hiánya

Egy kis történelem - klasszikus neurális hálózatok

- 1995 - Tesauro: TD-Gammon
  - Megerősítéses tanulás
  - Neurális hálózatot használt a hasznosságfüggvény eltárolásához
- 1995 - Vapnik: Support-vector networks
  - Számos problémára sokkal jobb eredményt ad, kevésbé nehézkes a tanítása
  - Hatásásra a neurális hálózatok kutatása (ismét) alábbhagyott
- 1997 - Schmidhuber and Hochreiter: Long Short Term Memory (LSTM)
- 1998 - LeCun, Bottou, Bengio, Haffner: Gradient-based learning applied to document recognition
  - Konvolúciós mély architektúra (LeNet-5)
  - Kézzel írt számjegyek felismerése csekkeken



LeNet-5

Egy kis történelem - deep learning

- 2000-es évek közepén a neurális hálózatok témakörben alig lehetett cikket elfogadtatni
- 2004 - CIFAR (Canadian Institute for Advanced Research) mégis finanszírozta Hinton kutatását a témában, akinek sikerült több más kutatót is meggyőznie
  - Új csomagolás a neurális hálózatoknak: deep learning
- 2006 - Hinton, Osindero, Yee-Whye Teh: A fast learning algorithm for deep belief nets
  - Új áttörés: mély hiedelem hálók rétegenkénti tanítása, majd az így inicializált súlyok további finomítása ellenőrzött tanítással (klasszikus módon), state-of-the-art eredmények (MNIST)
  - (előtanítás a 2002-es contrastive divergence (Hinton) algoritmussal)
  - Újra beindítja a kutatást

Egy kis történelem - deep learning

- 2010, Glorot, Bengio: Understanding the difficulty of training deep feedforward neural networks
  - ReLU, új súlyinicializálási mód
  - Nincs is szükség pretrainingre
- 2011 - Mohamed, A. R., Sainath, T. N., Dahl, G., Ramabhadran, B., Hinton, G. E., & Picheny, M. (2011, May). Deep belief networks using discriminative features for phone recognition.
  - Egy évtizedes rekordot döntöttek meg
  - GPU használata
- 2011 - Google Brain kutatócsoport megalakulása
  - Más cégek is beszállnak a kutatásba
  - Big data fénykorát éli

Egy kis történelem - deep learning - ImageNet

- 2009 - J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database
- Adatkészlet, nagy felbontású képek, 14M kép, 10k+ kategória
  - http://www.image-net.org/about-stats
- ILSVRC - Large Scale Visual Recognition Challenge
  - 1.2M tanító, 200k teszt és validációs kép, 1000 kategória
    - Pl: http://image-net.org/synset?wnid=n11908846
  - 2010 óta minden évben
  - 2011-ben legjobb nem-deep megoldás:
    - 25.8% top-5 hiba

Egy kis történelem - deep learning - ILSVRC

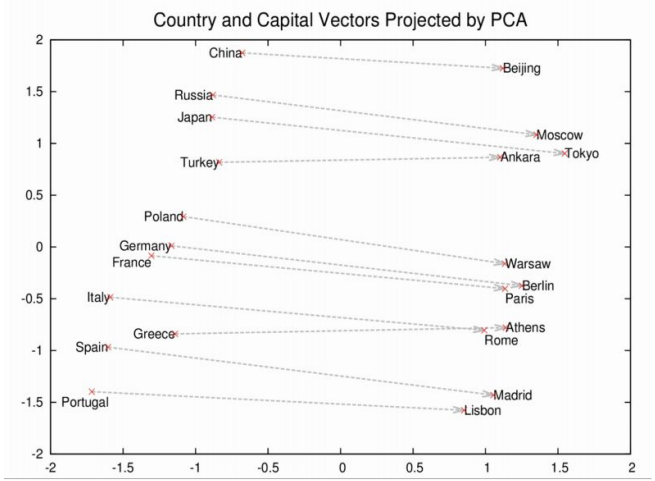
- 2012 - Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton Imagenet classification with deep convolutional neural networks
  - Ez az első tisztán neurális hálózatos megoldás, ami az azévi legjobb eredményt éri el (addig SVM)
  - Innentől kezdve lesz mainstream a deep learning
  - AlexNet: 8 réteg, 35k paraméter
  - 15.4% top-5 hiba
- 2014 - Simonyan and Zisserman, VGGNet
  - 19 réteg, 138M paraméter
  - 7.3% top-5 hiba
- 2014 - Szegedy et al., GoogLeNet
  - 6.9% top-5 hiba
  - 22 réteg, 5M paraméter
- 2015 - He et al., ResNet
  - 3.6% top-5 hiba
  - 152 réteg

Egy kis történelem - deep learning - Google DeepMind

- Deep reinforcement learning
- 2012 - Playing Atari with Deep Reinforcement Learning
- 2014 - január - Google felvásárolja
- 2014 - október - Neural Turing Machines cikk
- 2015 - október - AlphaGo legyőzi az európai Go bajnok Fan Huit
- 2016 - március - AlphaGo legyőzi a Go világbajnok Lee Sedolt

Néhány deep learning eredmény

- Beszédfelismerés (TIMIT adatkészlet)
  - 17.7% hiba
- Nyelvi modellezés
  - Karakter szintű előrejelzés (szöveg generálás)
  - Automatikus fordítórendszer
  - Word2Vec



PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.

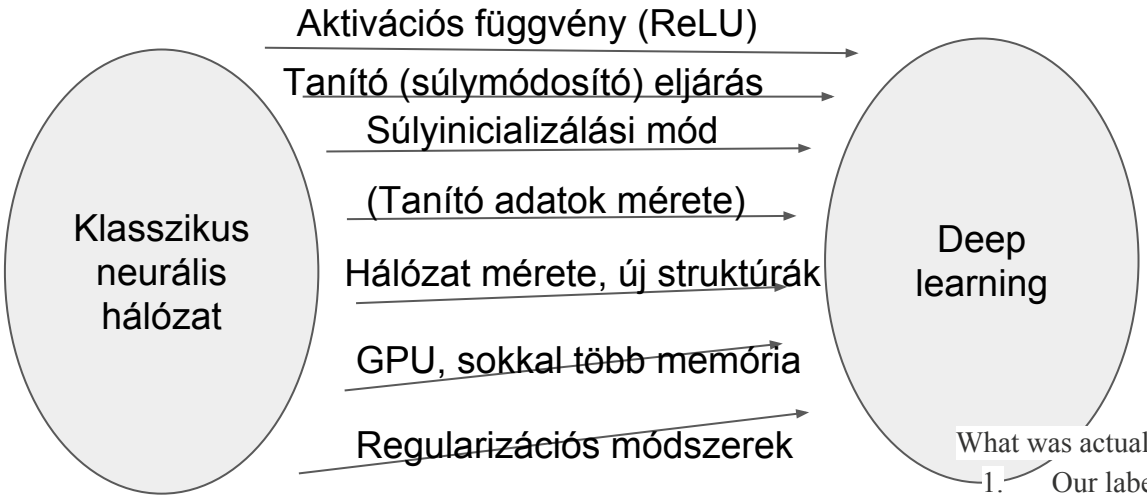
Néhány deep learning eredmény

- Képfelismerés
  - Osztályozás
  - Lokalizáció, szegmentálás
  - Kép leírás generálás

Néhány deep learning eredmény

- DeepDream
- NeuralStyle

Klasszikus neurális hálózat vs. deep learning



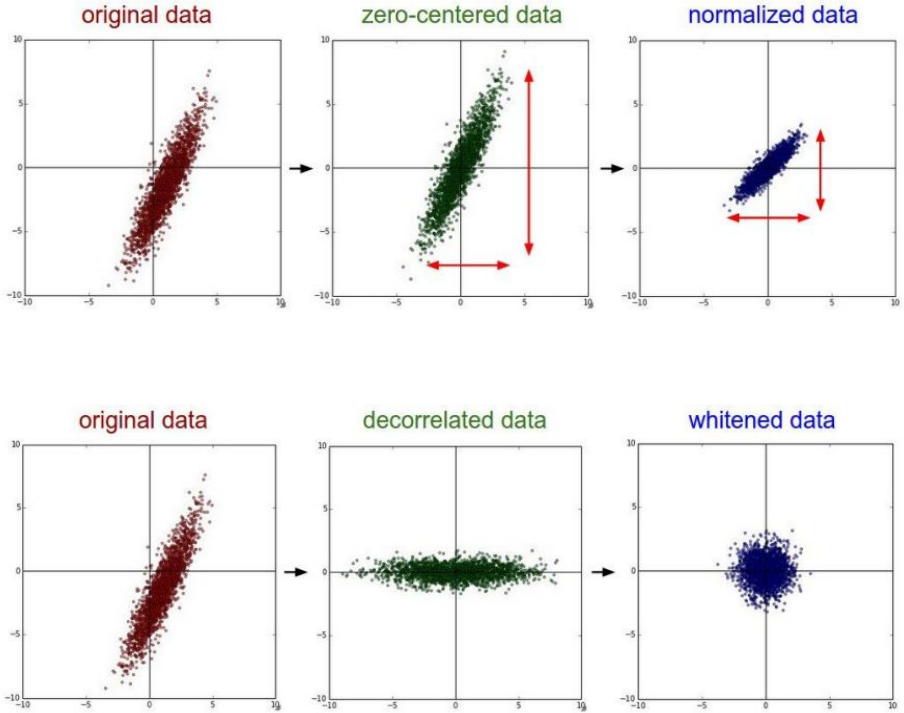
What was actually wrong with backpropagation in 1986?

1. Our labeled datasets were thousands of times too small.
2. Our computers were millions of times too slow.
3. We initialized the weights in a stupid way.
4. We used the wrong type of non-linearity.

Professor Geoff Hinton, “Deep Learning”. May 22, 2015.

Adatok előfeldolgozása

- Normalizálás
  - Nulla várható érték, egységnyi szórás
- PCA
  - Főkomponens analízis
  - Adatok fehérítése (korrelálatlan eloszlás)



Tanítás menete - minibatch (Stochastic Gradient Descent)

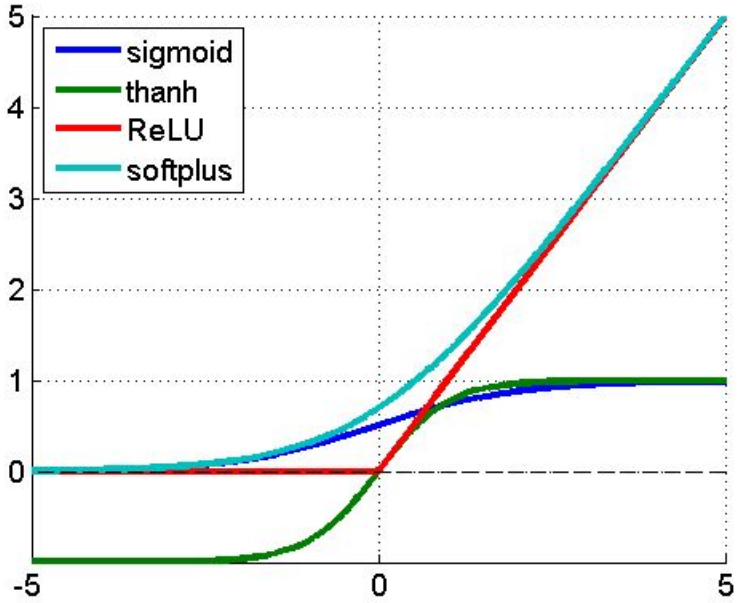
- A tanítás során a teljes mintakészletre vett hiba súlyok szerinti deriváltjával szeretnénk tanítani
- Módosításonként az összes tanító pontra kiszámolni a gradienst túlságosan költséges
- Egy-egy tanítópont alapján számított gradiens nagyon “zajos”, nem becsli jól a tényleges gradienst
  - A zajos gradiens akár teljesen rossz irányba is viheti a tanítást
- Minibatch: vegyünk néhány (~100) mintát, azzal számoljunk egy gradiens becslőt, és azzal tanítsunk
  - A tanítópontok számától független a tanítás hatékonysága

Aktivációs függvények

2009 - Jarrett, K., Kavukcuoglu, K., Ranzato, M. A., & LeCun, Y. What is the best multi-stage architecture for object recognition?

2010 - Nair, V., & Hinton, G. E. Rectified linear units improve restricted boltzmann machines.

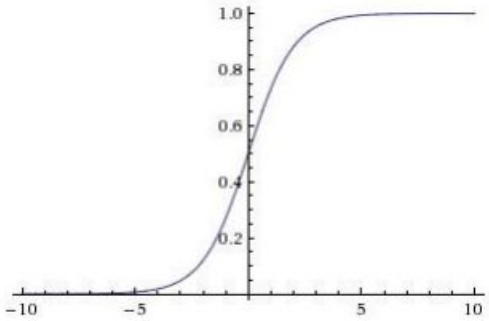
2011 - Glorot, X., Bordes, A., & Bengio, Y. Deep sparse rectifier neural networks.



Aktivációs függvények - szigmoid

- A [0,1] intervallumba transzformálja a kimenetet
- A biológiai neuronhoz hasonlóan telítődően “tüzel”

$$\sigma(x) = 1/(1 + e^{-x})$$

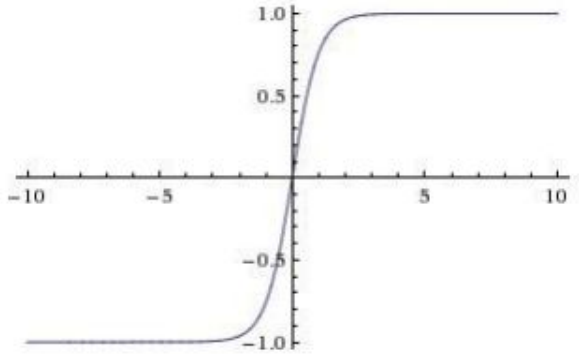


Problémái:

- A telített neuron “megöli” a gradienst
  - x = -10, x = 10, x = 0
- A kimenetei nem a nulla körül vannak szimmetrikusan, kimenete mindig pozitív
  - Súlymódosításnál a következő neuron súlyai mindig vagy csupa pozitív vagy csupa negatív vektorral lesznek módosítva
- exp(-x) függvény számítása költséges

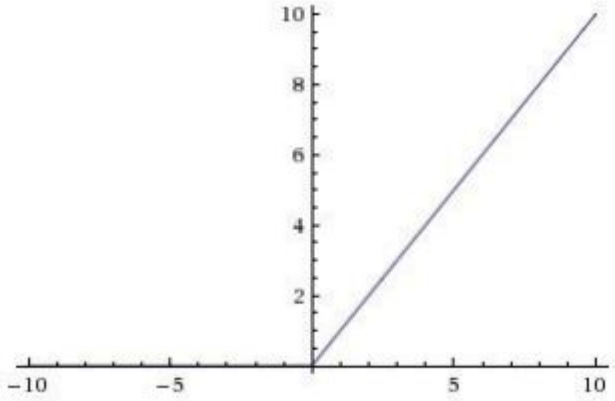
Aktivációs függvények - tanh(x)

- [-1,1] intervallum
  - vannak negatív kimenetek is
- Továbbra is megöli a gradienst
- Továbbra is költséges a számítása



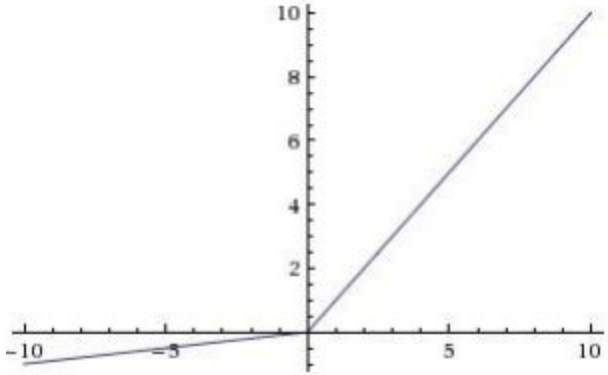
Aktivációs függvények - ReLU

- Rectified Linear Unit
- f(x) = max(0, x)
- Számítása egy művelet
- Gradiens számítása szintén
- Nem telítődik a pozitív tartományban
  - Nem öli meg a gradienst
- Sokkal gyorsabban konvergál, mint a szigmoid vagy a tanh
- x<0-ban 0 a gradiens, ha nem kap pozitív bemenetet, akkor “halott” neuron lesz, soha sem módosul a súlyainak az értéke
  - Inicializáláskor kicsi pozitív bias (0.01)



Aktivációs függvények - Leaky ReLU

- f(x) = max(0.01x, x), f(x) = max(alpha * x, x)
  - alpha tanítható paraméter
- Ugyanolyan jó, mint a ReLU
- Neuron nem tud meghalni
- Számítása kicsit költségesebb



Súlyok inicializálása

- Alap ötlet: kis szórású, 0 várható értékű véletlen számok
  - Kis hálózatokra jól működik, de sok réteg esetén nem lesz azonos a súlyok eloszlása az egyes rétegekben
- Xavier-inicializálás (2010, Glorot, Bengio: Understanding the difficulty of training deep feedforward neural networks)
  - A bemenetek száma is számít: a szórás legyen 1/sqrt(bemenetek száma)
  - Probléma: nem veszi figyelembe a nemlienaritás hatását (főleg ReLU-nál gond)
- (2015, K. He, X. Zhang, S. Ren, and J. Sun, “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification)
  - A súlyokat ReLU esetén 1/sqrt(bemenetek száma / 2) módon inicializáljuk
  - Majdnem egyenletes eloszlást kapunk az összes rétegben

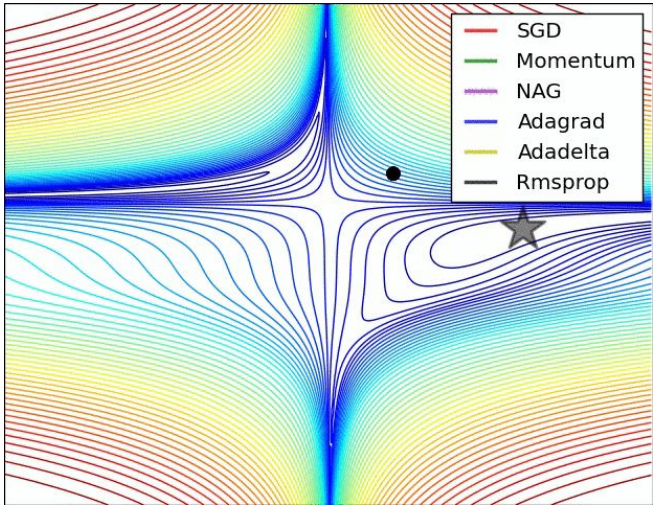
~~Súlyok inicializálása~~ - Batch normalization

- 2015, Sergey Ioffe, Christian Szegedy: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift
- Nem az inicializálás az igazán fontos, a működés közben ne “romoljon” el a súlyok eloszlása
- Használjunk minden összegző után egy olyan réteget, ami normalizálja a nemlinearitások számára a bemeneteket
- Ez egy deriválható függvény
- Minden egyes minibatchre külön számítandó
- Hatására nagyobb bátorsági faktor is használható
- Megkönnyíti a gradiensek visszaterjesztését a hálóban (kevésbé tűnnek el)

$$\widehat{x}^{(k)} = \frac{x^{(k)} - \mathrm{E}[x^{(k)}]}{\sqrt{\mathrm{Var}[x^{(k)}]}}$$

Tanító eljárások

- SGD: legegyszerűbb gradiens módszer
- Momentum módszer
  - v = mu * v - learning_rate * dx
  - x += v
- Egyéb módszerek
- Adam
  - Pillanatnyilag legjobb gradiens alapú módszer
- Másodrendű optimalizációs módszerek
  - Hesse-mátrix négyzetes tárkomplexitás - nem skálázódik jól
  - Kisebb problémákra érdemes kipróbálni
  - BFGS, L-BFGS, LM
  - teljes batch esetekre jók, minibachre kevésbé

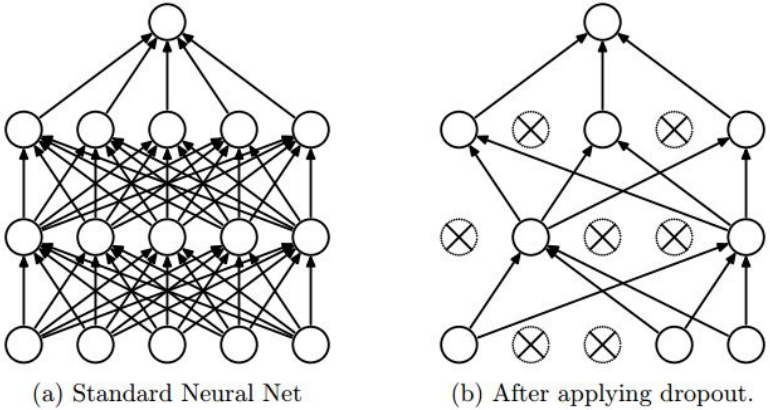


Regularizáció

- Probléma: túltanulás
  - Túl komplex a háló (túl sok paraméter), ÉS túl kevés tanítópont
  - Háló általánosító képessége egy idő után nem javul, sőt el is kezdhet romlani
- Háló méretének csökkentése
  - Nem ez a jó irány, megakadályozhatja a pontosabb illeszkedést
- Legegyszerűbb módszer: korai leállítás
  - Túl agresszív módszer
  - A túltanulást megakadályozza, de feltehetően nem tudjuk kihozni a maximumot a hálóból
- A súlyok L1 vagy L2 normájú regularizációja
  - Működik, de nehéz a paraméterezése
  - jelentős javulást nem várhatunk
- Dropout
  - 2014, Srivastava et al.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting
  - Minden előreterjesztésnél a neuronok kimenetét p valószínűséggel 0-ra állítjuk

Regularizáció - Dropout

- Neurális hálózatot redundáns leképezés megtanulására kényszeríti
- Így gyakorlatilag sok különböző háló átlagát használjuk, és ezeket párhuzamosan tanítjuk
- Teszt közben nyilvánvalóan nem használjuk
- Probléma: más a bemenetek eloszlása
  - Teszt esetben vegyük a bemenet p-szeresét, várható érték így meg fog egyezni a tanítás közbenivel
  - Vagy fordítva: tanítás során vegyük a kimenet p-ed részét, teszt során pedig minden marad a régiben
- Jelentős javulás érhető el, túltanulás nélkül
- Tanítás zajosabb lesz

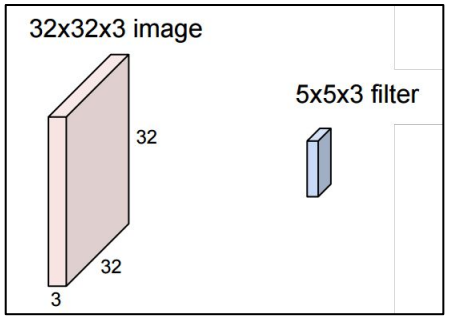


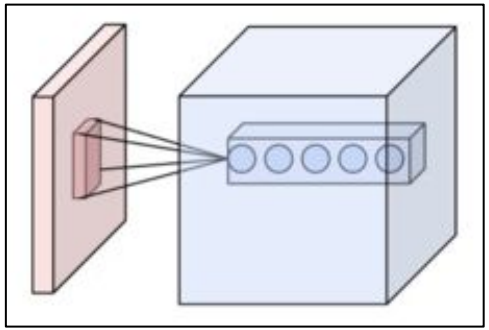
(a) Standard Neural Net

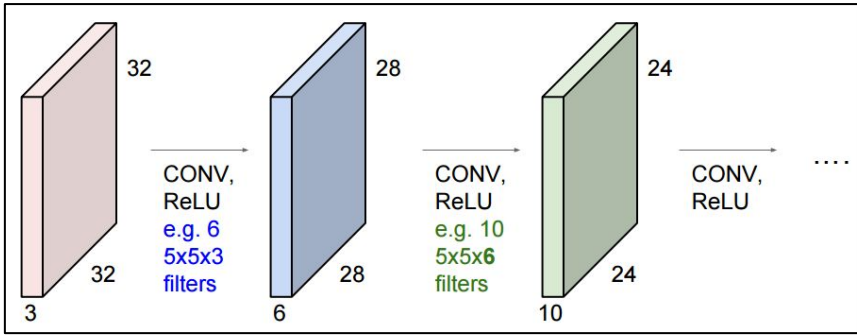
(b) After applying dropout.

Architektúrális változtatások - konvolúciós réteg (CNN)

- Nincs minden mindennel összekötve
- Konvolúciós ablak súlyait tanuljuk
- A súlyok minden bemeneti pixelere megosztott módon hatnak
- Térbeli invariancia a struktúrából fakadóan
- Kevesebb tanult paraméter
  - Gyorsabb tanulás
  - Gradiens kevésbé tud “elromlani”
  - Mélyebb háló architektúra - több nemlinearitás: jobb



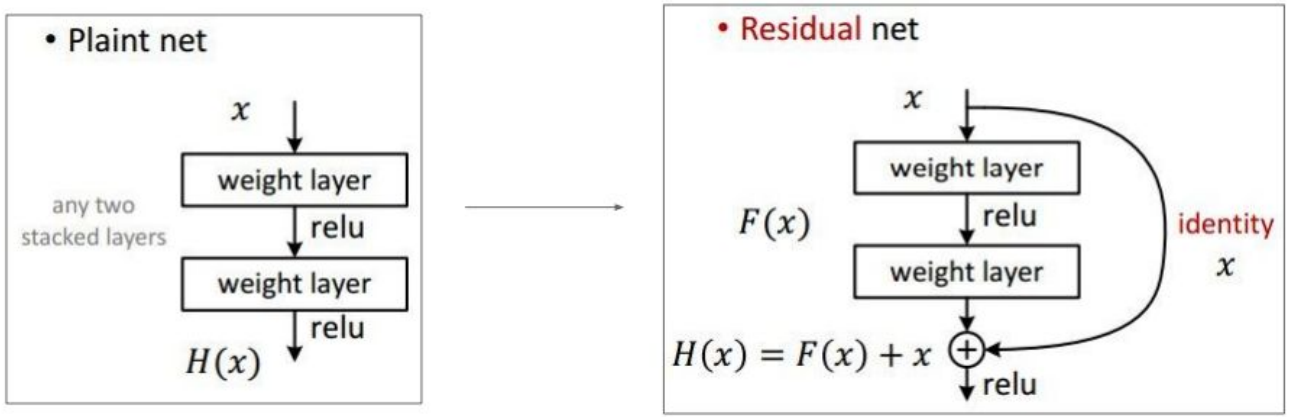




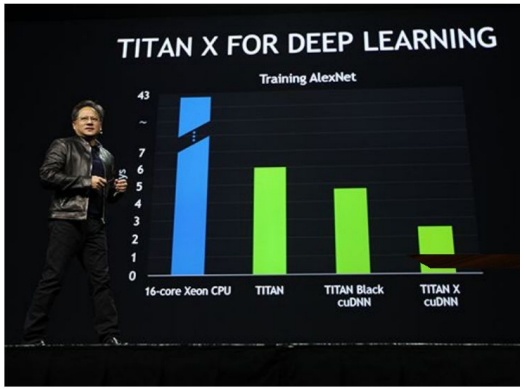
Architektúrális változtatások - ResNet

ResNet (residual network) - 2015 dec 10:
He et al., Deep Residual Learning for Image Recognition

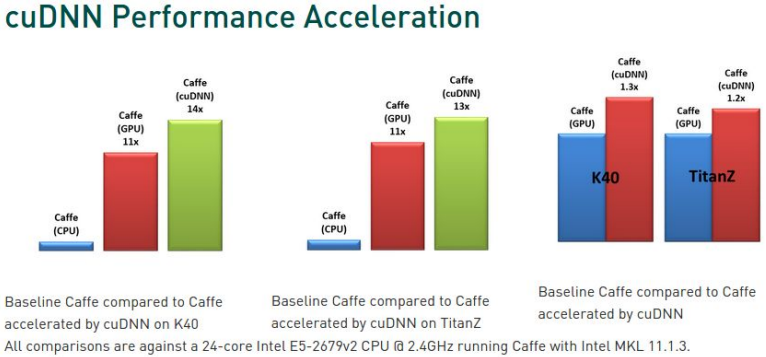
- Gradiensnek van direkt útja vissza a legelső rétegig



Hardver - GPU



- Moore “törvény”
- Mátrixműveletek jól párhuzamosíthatóak GPU-n
- NVIDIA aktívan támogatja a deep learninget saját hatékony implementációkkal (cuDNN)
- 32 bit elég (single prec.)
- 16 bit? (következő architektúra)
- Courbariaux and Bengio, February 9 2016:
  - Train with 1-bit activations and weights!
  - All activations and weights are +1 or -1
  - Fast multiplication with bitwise XNOR
  - (Gradients use higher precision)



Hivatkozások

A prezentáció a következő helyekről vett tartalmakkal készült:

- A 'Brief' History of Neural Nets and Deep Learning
  - http://www.andreykurenkov.com/writing/a-brief-history-of-neural-nets-and-deep-learning/
  - http://www.andreykurenkov.com/writing/a-brief-history-of-neural-nets-and-deep-learning-part-2/
  - http://www.andreykurenkov.com/writing/a-brief-history-of-neural-nets-and-deep-learning-part-3/
  - http://www.andreykurenkov.com/writing/a-brief-history-of-neural-nets-and-deep-learning-part-4/
- CS231n: Convolutional Neural Networks for Visual Recognition, Stanford University
  - http://cs231n.stanford.edu/syllabus.html