Quick Introduction to Data Analysis and Machine Learning

Michal Barla, Márius Šajgalík
PeWe.Data intro

Strojové učenie – machine learning

- Použitie algoritmov riešiacich isté typy úloh nájdením/naučením sa súvislostí/závislostí v dátach
  - Tam kde návrh explicitného pravidlového systému nie je prakticky možný

Základné delenie

- S učiteľom – supervised
  - Máme označkované dáta
  - Úloha je nájsť mapovanie vstupu na výstup
  - Klasifikácia, regresia
- Bez učiteľa – unsupervised
  - Úloha je nájsť štruktúru v dátach
  - Zhlukovanie, feature learning

Učenie s učiteľom

- Snažíme sa opísať jednotlivé pozorovania cez nejaké príznaky, vlastnosti, features
- A učiacim algoritmom nájsť takú kombináciu týchto vlastností, ktorá dáva najmenšiu chybu
  - Nájsť váhy
- Algoritmov a prístupov je veľké množstvo
  - Klasifikácia? Binárna? Multi-class?
  - Regresia? Logistická? Lineárna?

Učenie s učiteľom

- **Snažíme sa opísať jednotlivé pozorovania cez nejaké príznaky, vlastnosti, features**
- Učiacim algoritmom nájsť takú kombináciu týchto vlastností, ktorá dáva najmenšiu chybu
  - Nájsť váhy
- Algoritmov a prístupov je veľké množstvo
  - Klasifikácia? Binárna? Multi-class?
  - Regresia? Logistická? Lineárna?

Učenie bez učiteľa

- Nepotrebujem označené dáta
  - Takže ich môžem mať oveľa oveľa viac
- Hľadám v nich súvislosti, štruktúru
- Môžem sa “naučiť” príznaky, ktoré sú vhodné pre učenie s učiteľom
  - Takto sa učia aj malé deti

Príklad: word feature vectors

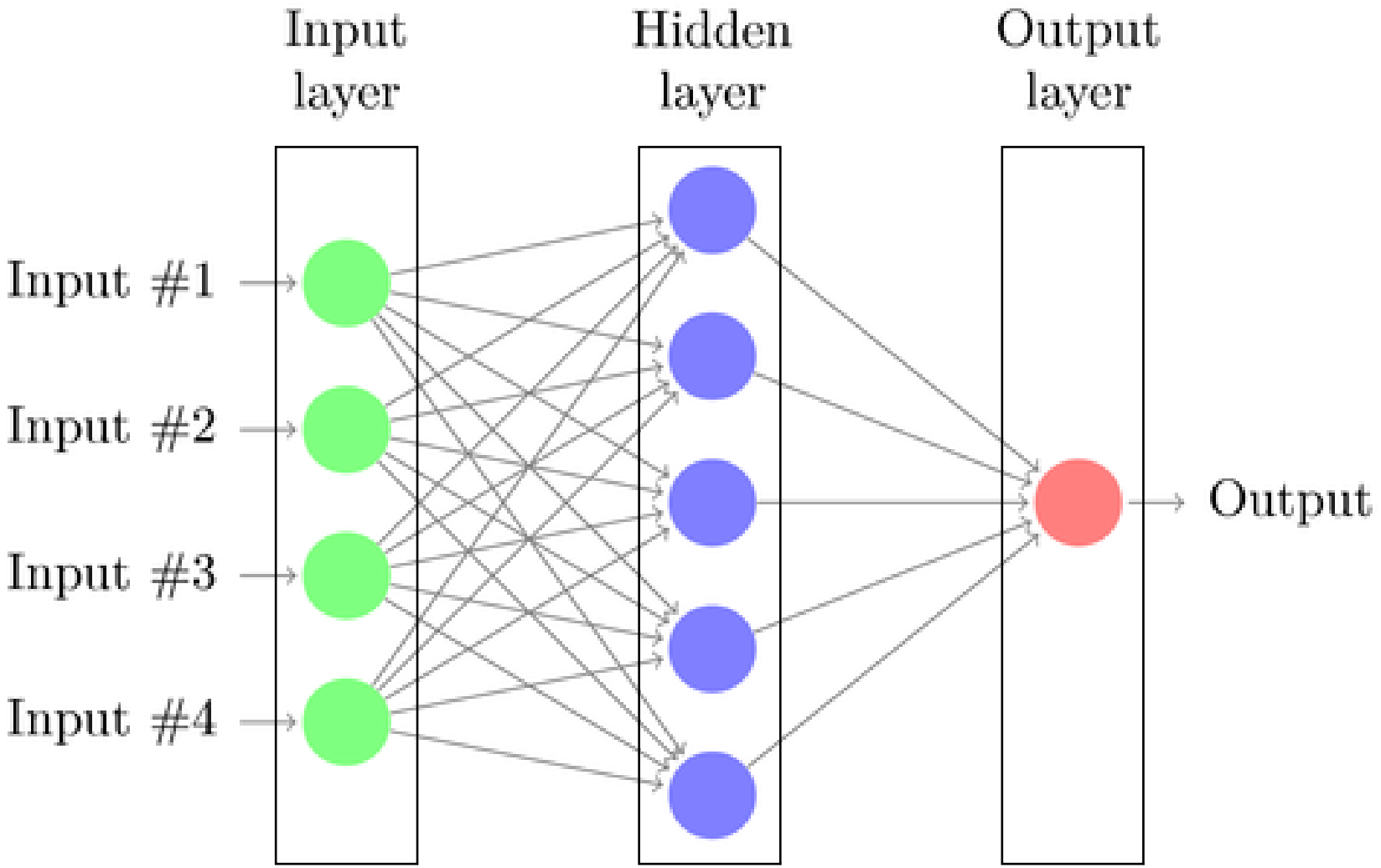
- Použitie RNN na naučenie sa modelu jazyka
  - Ktorý je lepší ako prístupy založené na n-gramoch
  - Unsupervised!
- Naučené vektorové reprezentácie kódujú mnoho **sémantických** and **syntaktických** vzťahov medzi slovami[1]
  - vector(“Kráľ”) ≈ vector(“Králi”)
  - vector(”Kráľ”) - vector(”Muž”) + vector(”Žena”) ≈ vector(“Kráľovná”)

1 - Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. "Linguistic Regularities in Continuous Space Word Representations." *HLT-NAACL*. 2013.

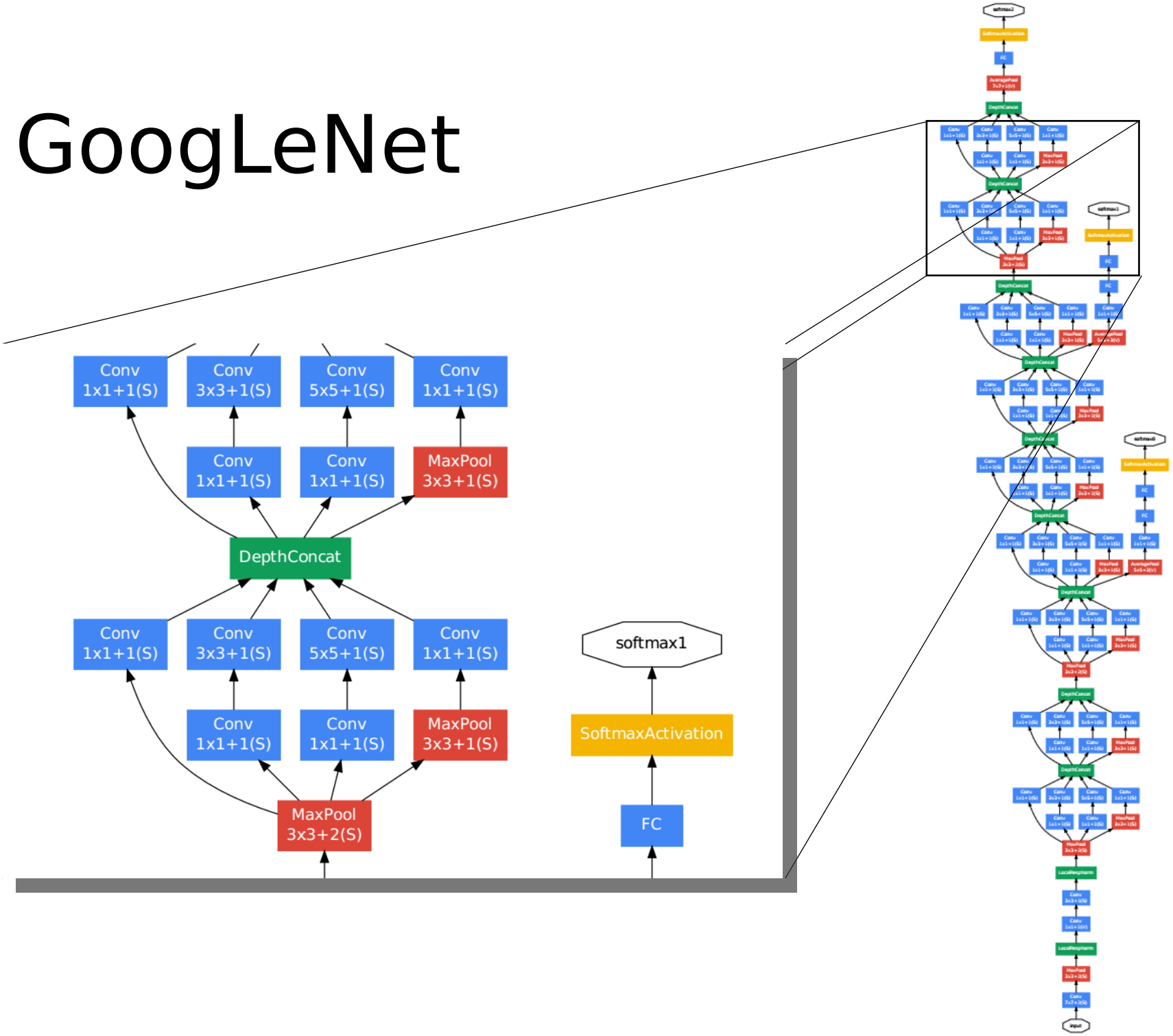
Unsupervised feature learning and deep learning

- Rovnaké tkanivo v ľudskom mozgu vie spracovať zrak, zvuk, dotyk
  - Mohol by existovať jeden učiaci algoritmus umelých neurónových sietí, ktorý sa dokáže naučiť rozlišujúce vlastnosti vašich dát
  - Objavili sa nové algoritmy, ktoré umožňujú neurónovým sietiam škálovať

Tradičná umelá neurónová sieť

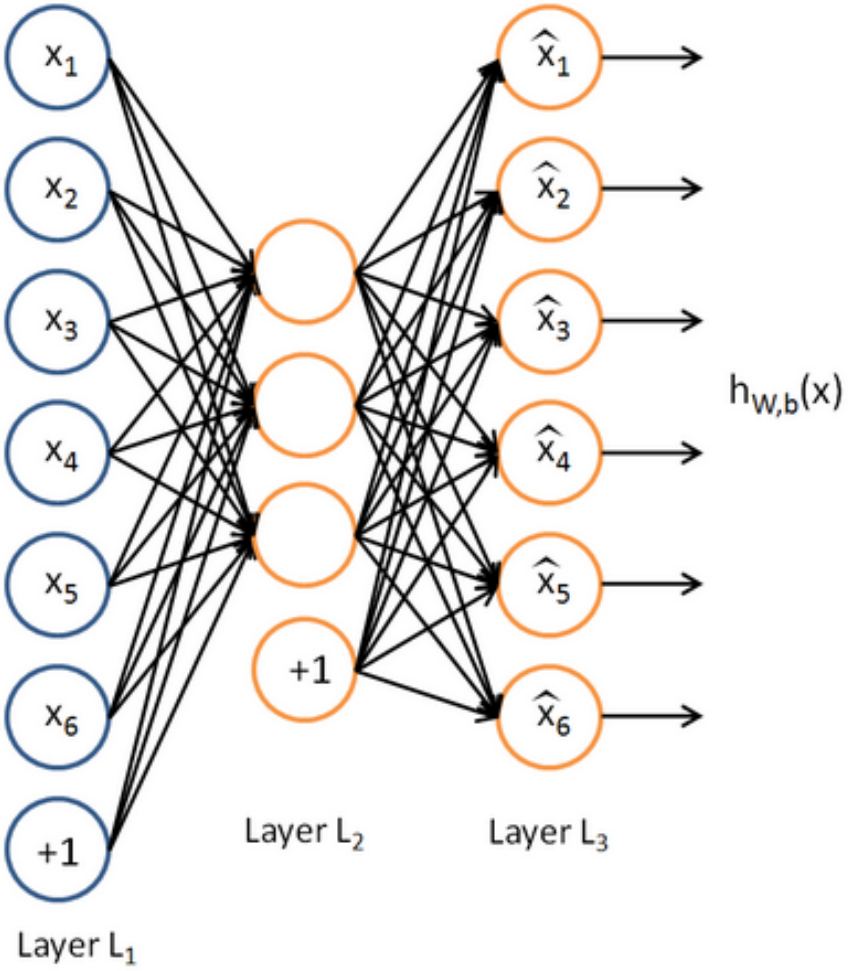


GoogLeNet


Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
MaxPool 3x3+2(S)
softmax1
SoftmaxActivation
FC

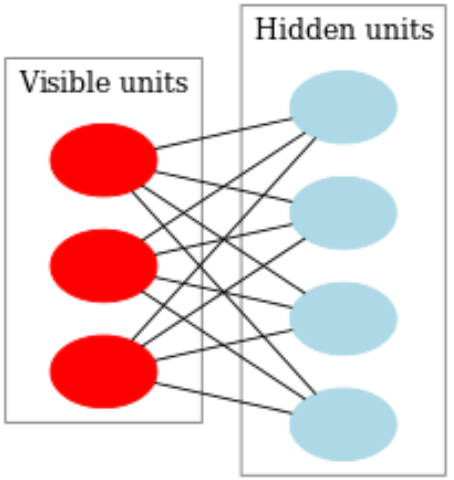
Autoencoder

- Naučme sa modelovať vstup



Restricted Boltzmann machine

- Viac pravdepodobnostne založené
- Contrastive divergence algorithm



- Vypočítaj pravdepodobnosť (sprav vzorku) skrytých neurónov
- Sprav vzorku viditeľných neurónov a z toho opäť vzorku skrytých

$$\Delta w_{i,j} = \epsilon(vh^{\mathsf{T}} - v'h'^{\mathsf{T}})$$

Deep belief net

- Navrstvené RBM
- Predtrénovanie po vrstvách
- Jemné dotrénovanie

Stacked autoencoder

- Navrstvené autoencodere
- Predtrénovanie po vrstvách
- Jemné dotrénovanie

Sparsity

- Obmedzíme priemernú aktiváciu neurónov
- Prinútime neurónku, aby využívala menej neurónov

Dropout/dropconnect

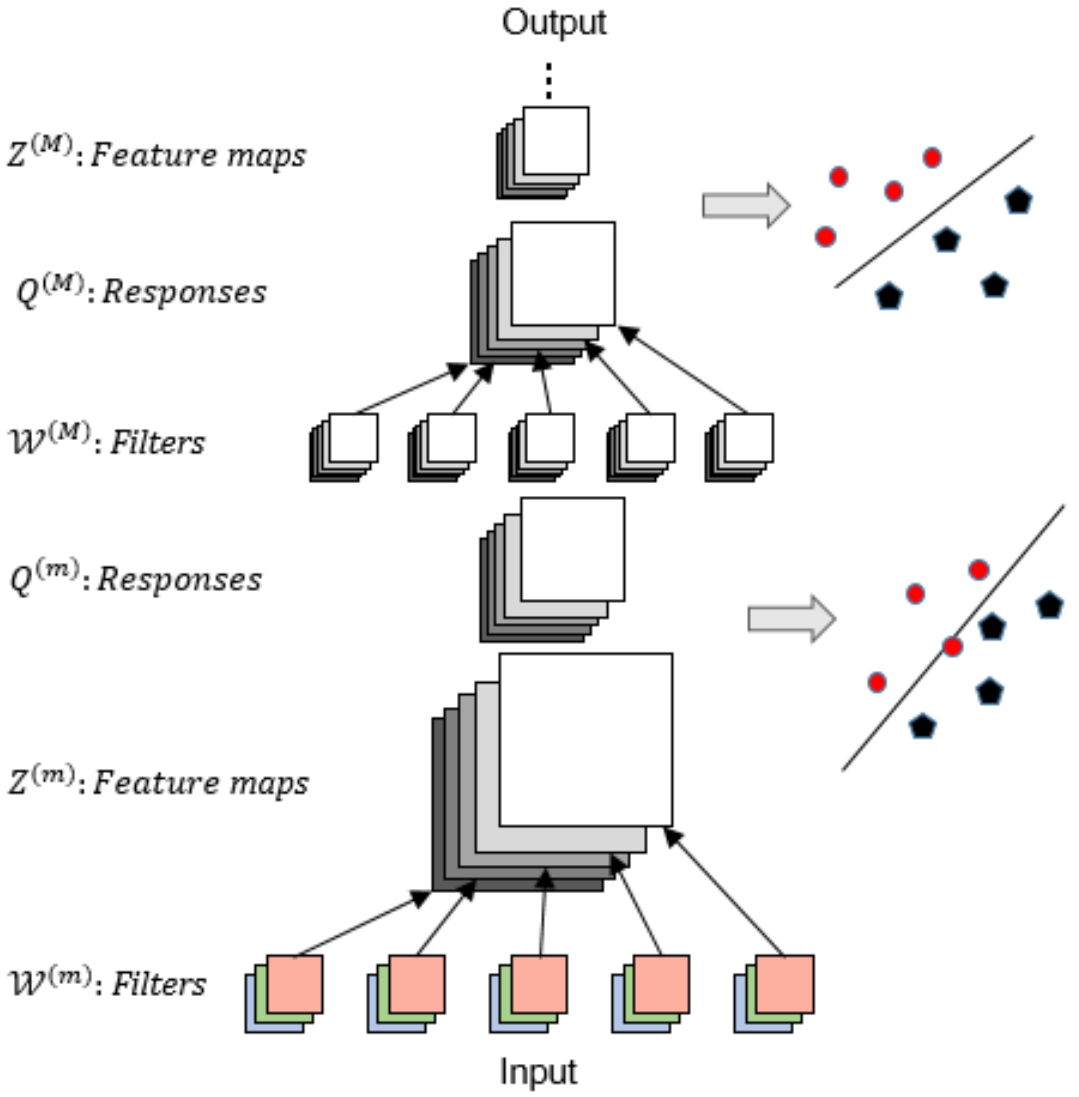
- Zavedieme pravdepodobnosť, že neurón vypadne
- Rozšírenie dropconnect – niekedy vypadne spojenie (váha)

Návrh architektúry

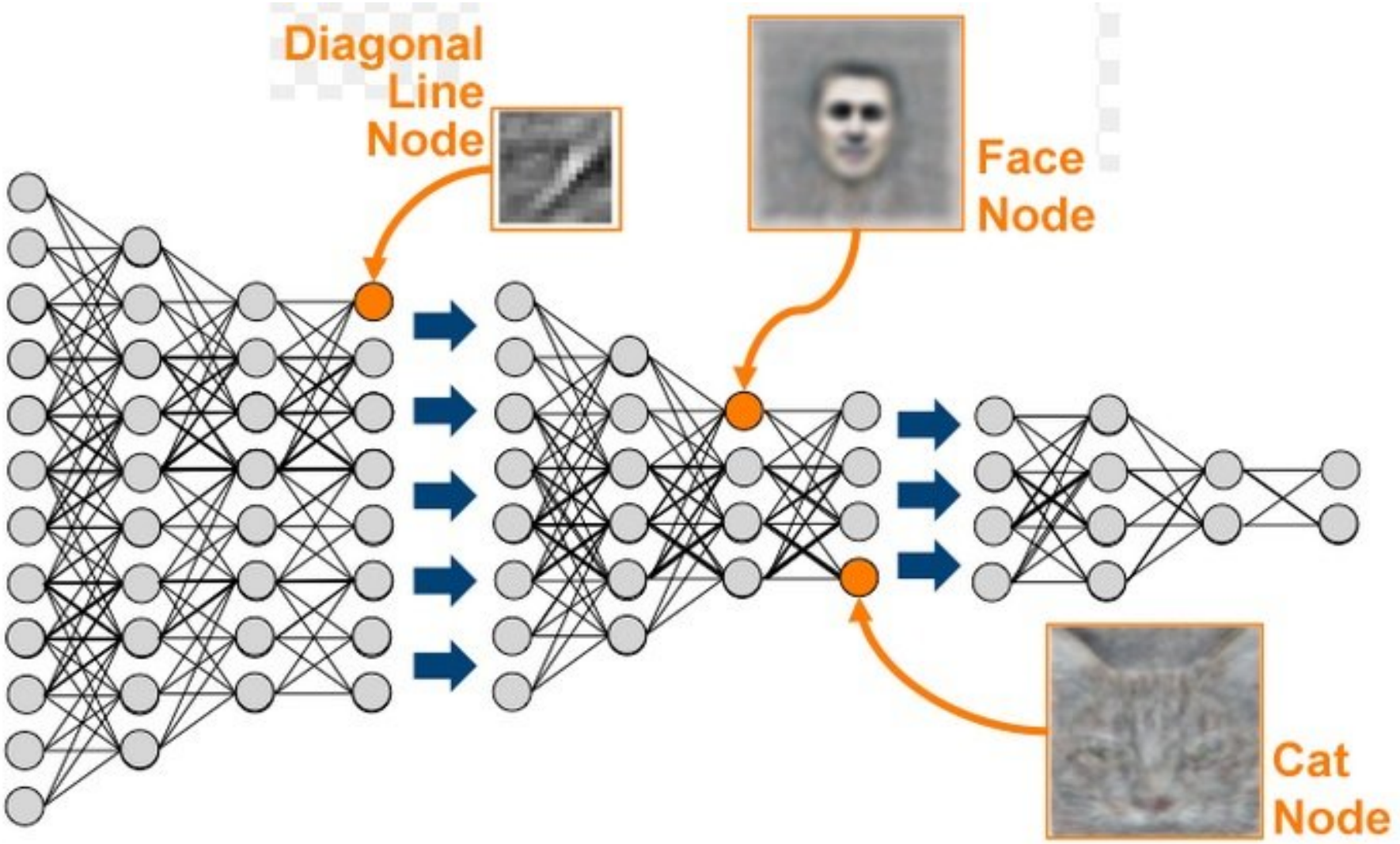
- V súčasnosti sa ukazuje, že je veľmi dôležitý výber architektúry
- Neurónku treba vedieť poskladať – **otvorený problém**
- Nie všetky (populárne) neurónky sú hlboké!

- word2vec – bez skrytej vrstvy, správny výber aktivačnej funkcie

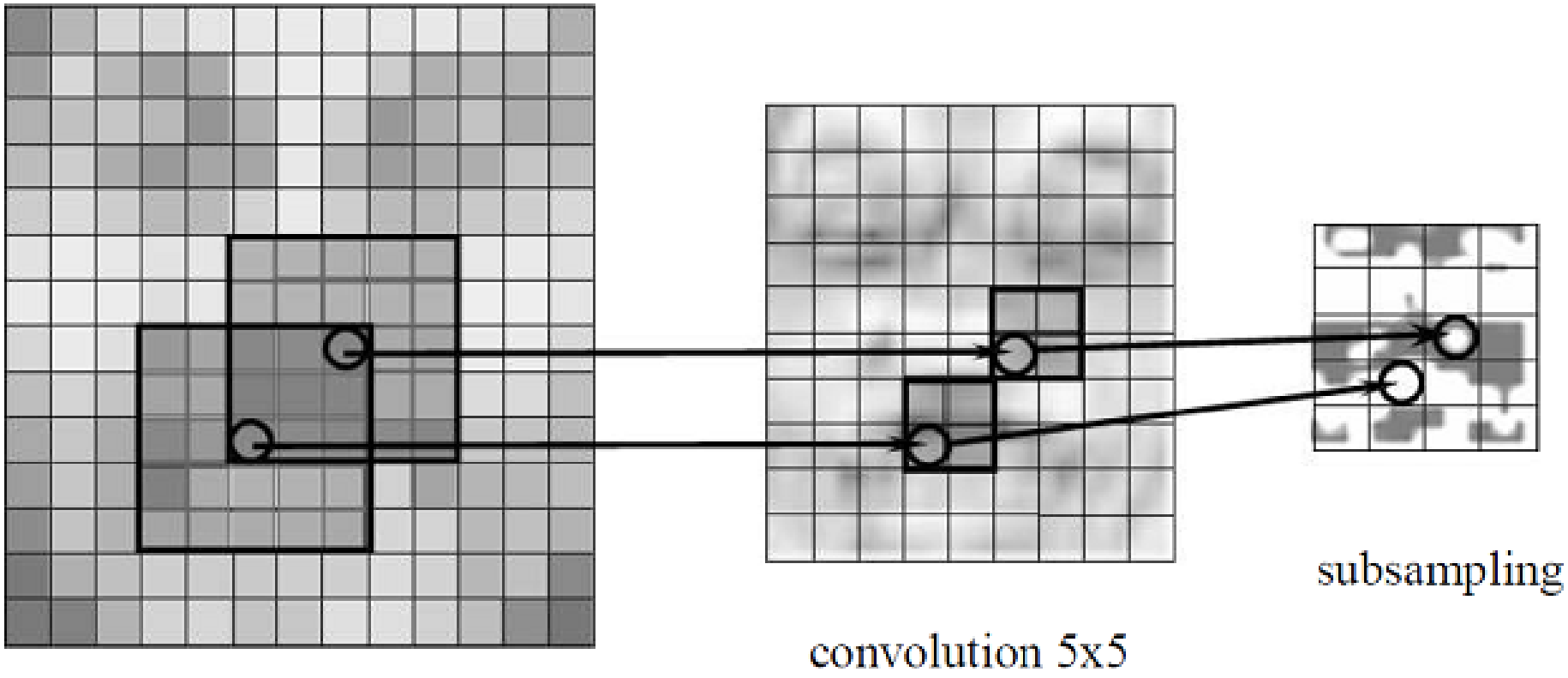
Discriminative feedback neurons



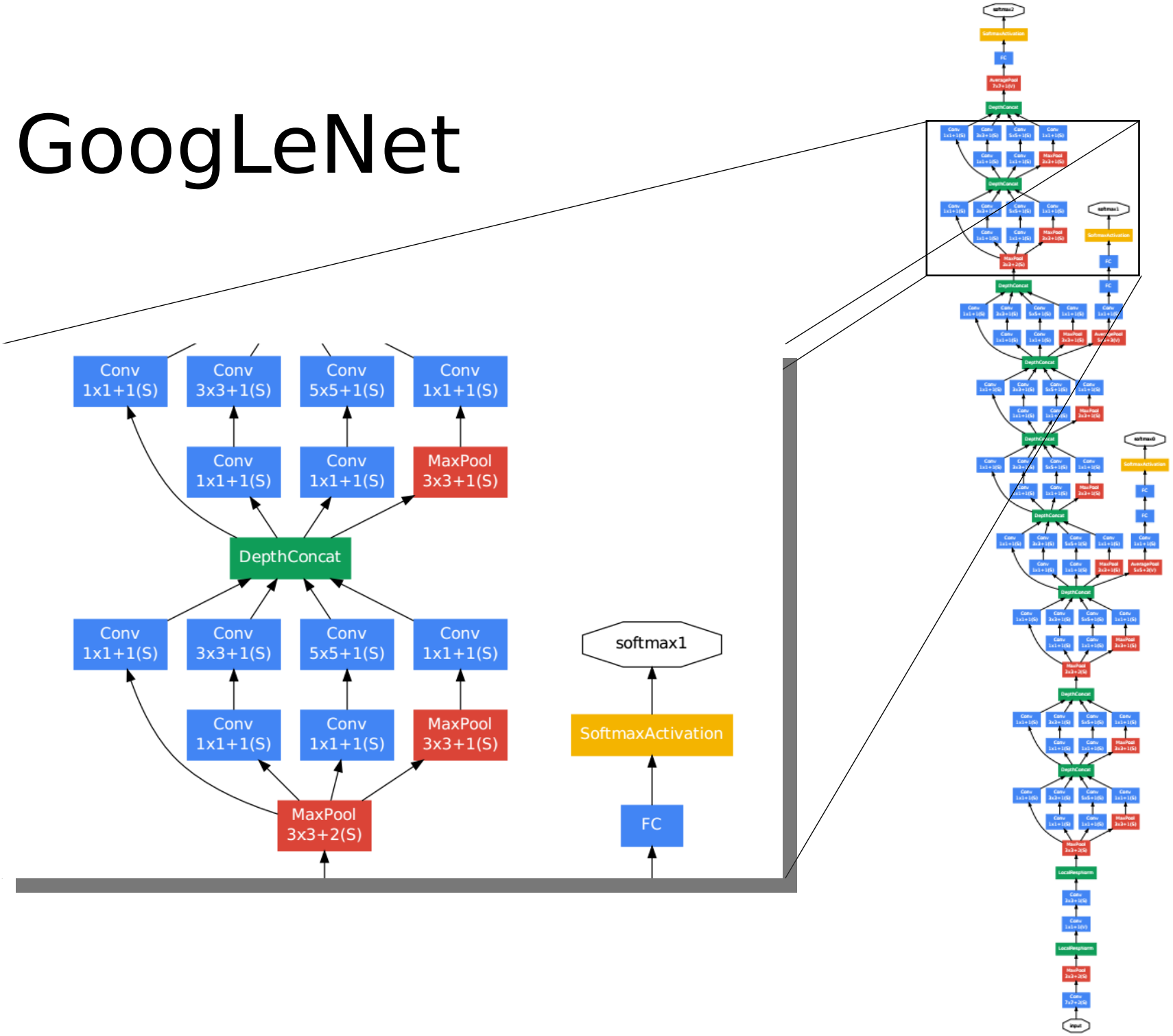
Konvolučné hlboké siete



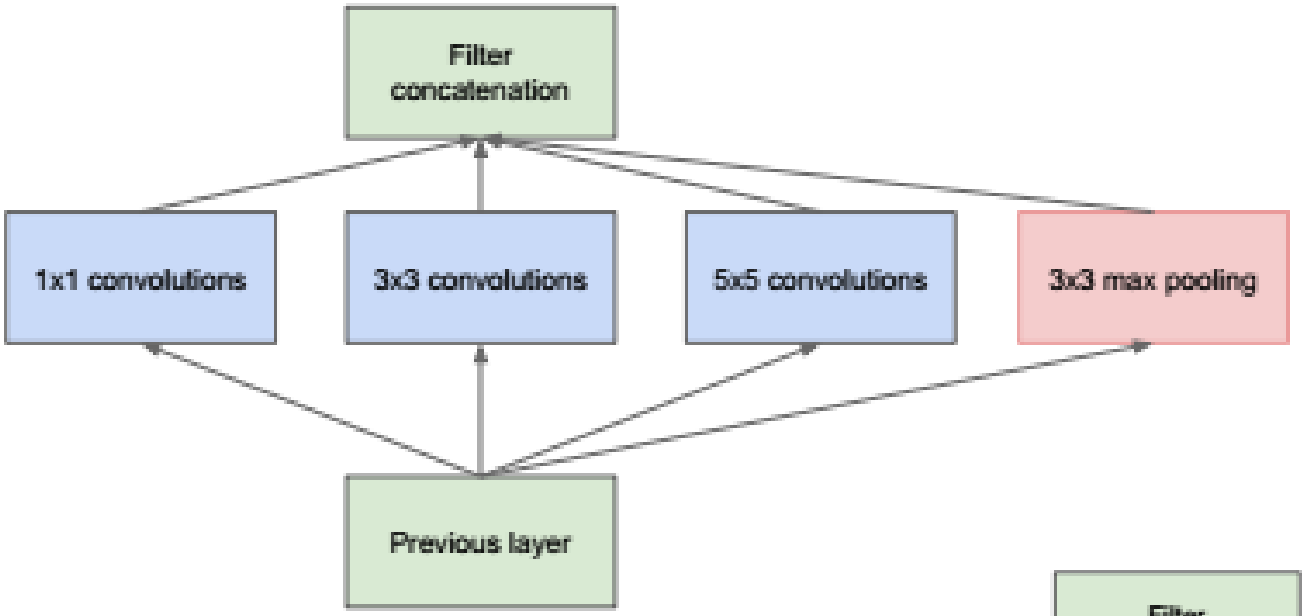
Konvolúcia

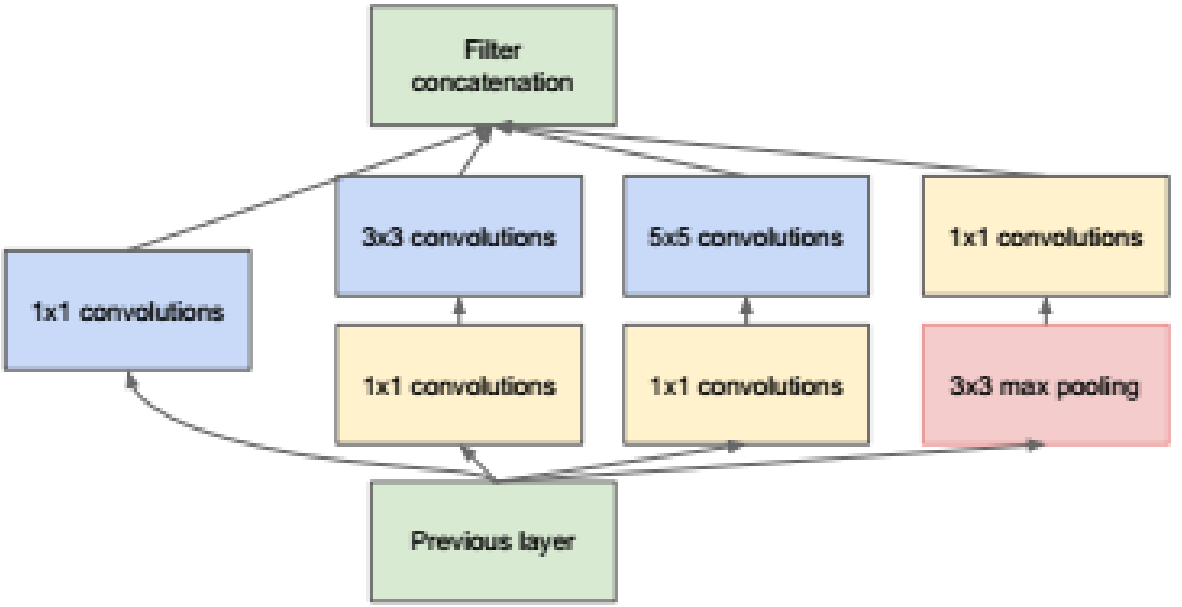


GoogLeNet


Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
DepthConcat
Conv 1x1+1(S)
Conv 3x3+1(S)
Conv 5x5+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
Conv 1x1+1(S)
MaxPool 3x3+1(S)
MaxPool 3x3+2(S)
softmax1
SoftmaxActivation
FC

Inception module





Nástroje a zdroje

- R
- Rapid Miner
- Theano
- Octave, SciPy, NumPy

- ML @ Coursera (Andrew Ng)
- Johns Hopkins University @ Coursera