Dreidimensionale Objektklassifizierung mithilfe der Convolutional Neuronal Networks

Master Grundseminar WiSe 2014/2015

Victoria Bibaeva

Inhalte

- Einführung
- Convolutional Neuronal Networks (Faltungsnetzwerke)
- Aktueller Forschungsstand
- Ziele: Projekt 1, Masterarbeit
- Literaturliste

Einführung

Motivation

[http://www.roboterwelt.de/companies/irobot/]

- Zukunft: Roboter soll den Menschen ersetzen
  - Aufgaben übernehmen
  - Arbeit erleichtern
  - z.B. morgens Kaffee holen
- Dafür müssen die Roboter die Umgebung erkennen und entsprechend handeln
- Erkennen = sehen + lokalisieren + klassifizieren
- Service-/Assistenzrobotik

Problemstellung

- Roboter soll eine Tasse Kaffee machen
- Roboter sucht dafür die Tasse
  - Objekterkennung oder
  - Objektklassifizierung
- Weitere Schritte:
  - Nach der Tasse greifen
  - Zur Kaffeemaschine fahren
  - usw.

[http://winfuture.de/news,83329.html]

Anwendungsgebiete

- Im Projekt:
  - Interaktion von Robotern mit Umgebung
- Auch anwendbar für:
  - Hilfe für pflegebedürftige Menschen
  - Haushaltshilfe
  - Navigation mit dem Auto, Einparkhilfe
  - Kameraüberwachung
  - Bilddatenbanken
  - Markerless Motion Capture ☺

[http://www.bielefeld-marketing.de/de/service/bibewegt/meldung.html?idpm=2011-12-09-09.46.27.430989]

Herausforderungen

- Klassifizierung ist schwieriger als Erkennung
  - Vielfalt der Objekte
  - Diverse Aussichtspunkte, eventuell Verdeckung
  - Einfluss von Umgebung (Licht, Farbe, Hintergrund)
  - Ähnlichkeit zweier Klassen
  - Objekte einer Klasse sehen unterschiedlich aus
- Haushaltsrobotik ist schwieriger als Industrierobotik
  - Umgebung und Position der Objekte sind nicht konstant

[5]

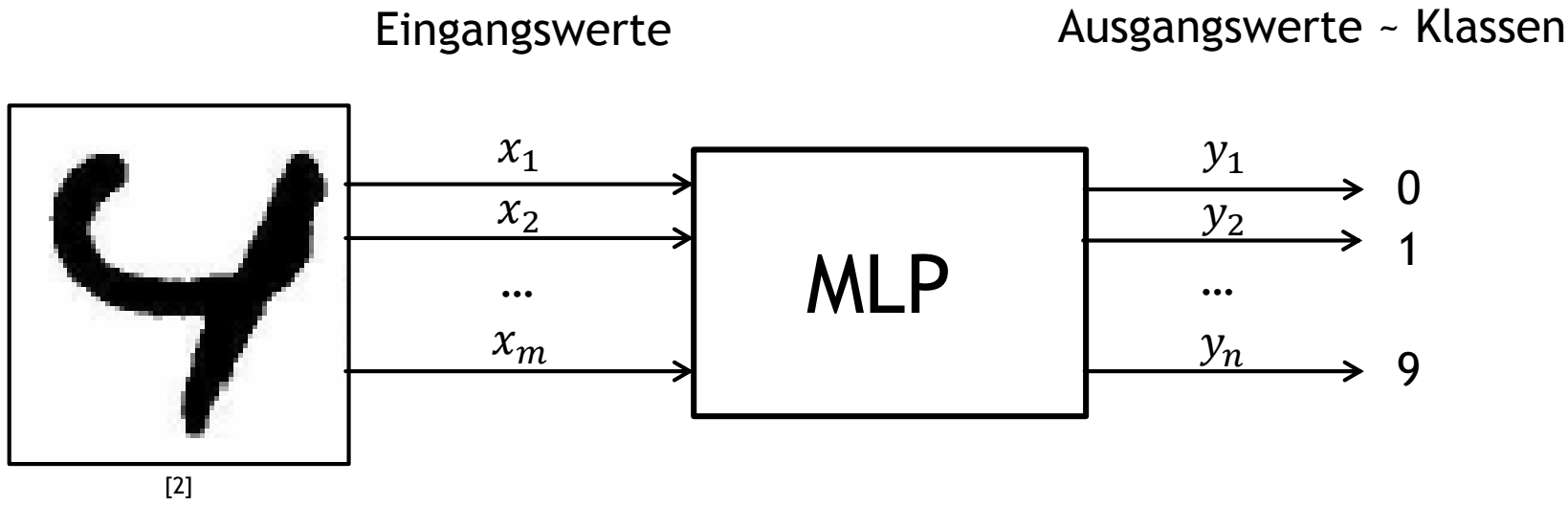
Convolutional Neuronal Networks

Was ist CNN?

- Dt. „Faltungsnetzwerk“
- Eine Variante von MLP mit einer besonderen Architektur
- Inspiriert von Sehrinde der Katzen
- Vorgestellt von Yann LeCun et al. in 1989 [2]
- Liefern sehr gute Ergebnisse z.B. bei:
  - handschriftlicher Zeichenklassifizierung
  - Gesichtserkennung
  - Vierbeinererkennung
- Robust: unempfindlich gegen Rotation, Translation, Skalierung, usw.

Zur Erinnerung: MLP

- Ein vereinfachtes künstliches neuronales Netz mit mehreren Schichten

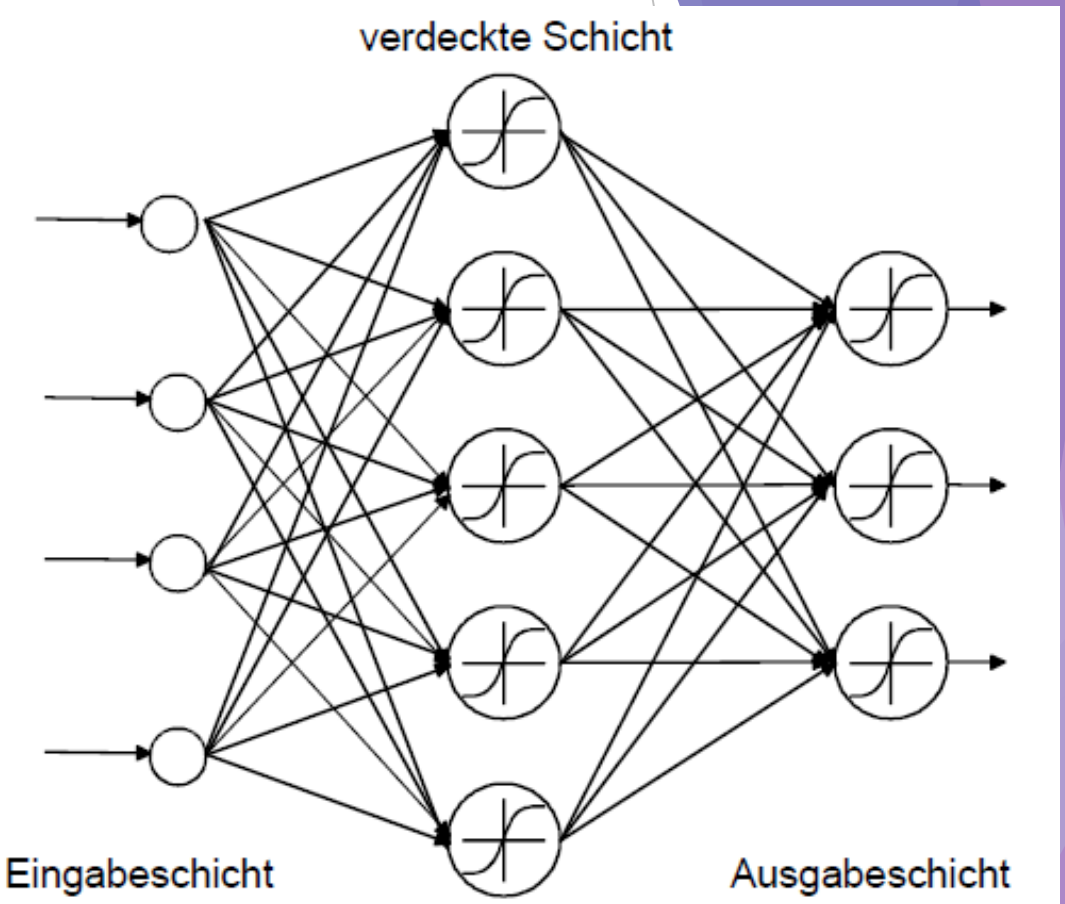


[2]

- Jeder Ausgangswert ist von allen Eingangswerten abhängig

Innere Struktur von MLP

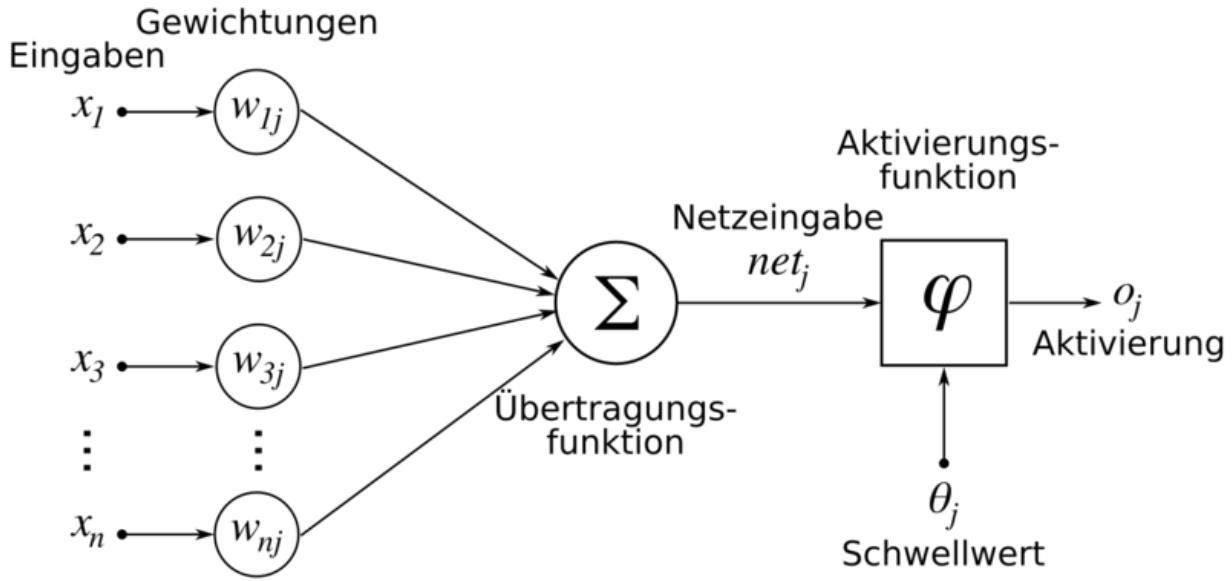
- Feed forward
- Full Connection
- Überwachtes Lernen
- Back Propagation
- MLP mit einem Hidden Layer kann jede Funktion approximieren (Cybenko Theorem, 1989)



A. Meisel. Vorlesungsskript „Robot Vision“ 2012

Neuron

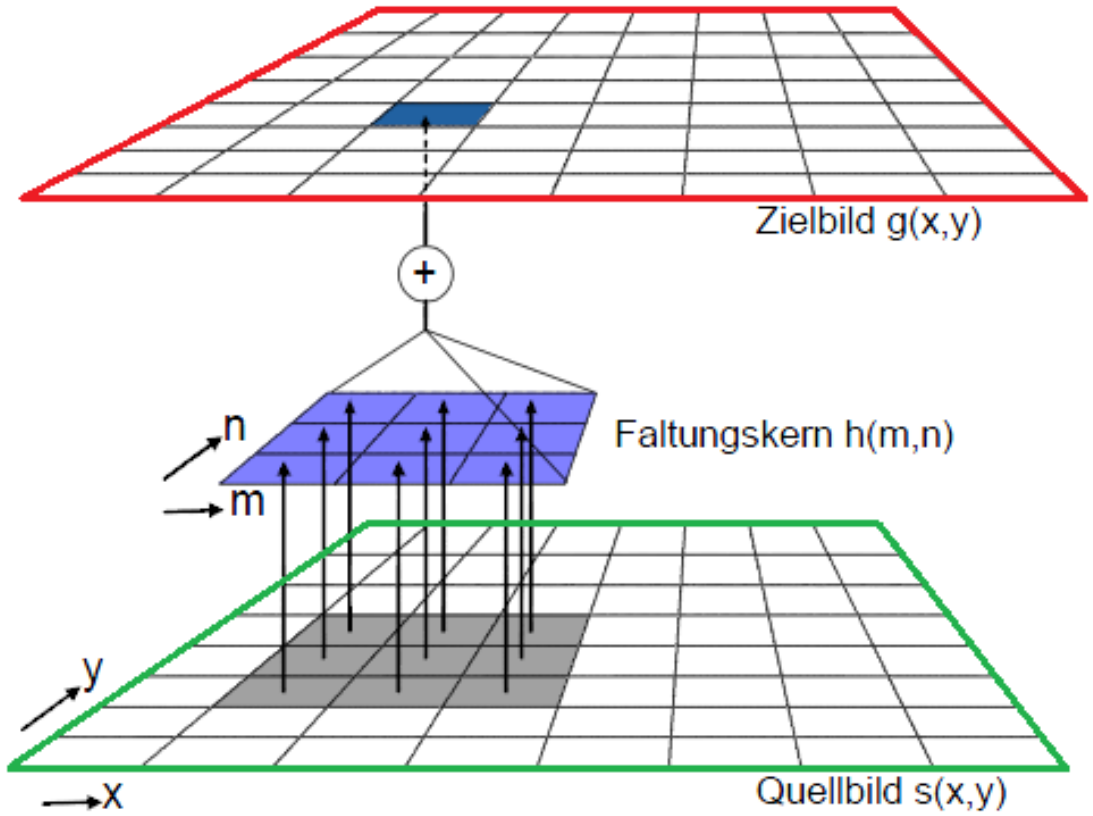
- Jede Schicht besteht aus Neuronen:



[http://de.wikipedia.org/wiki/Datei:ArtificialNeuronModel_deutsch.png]

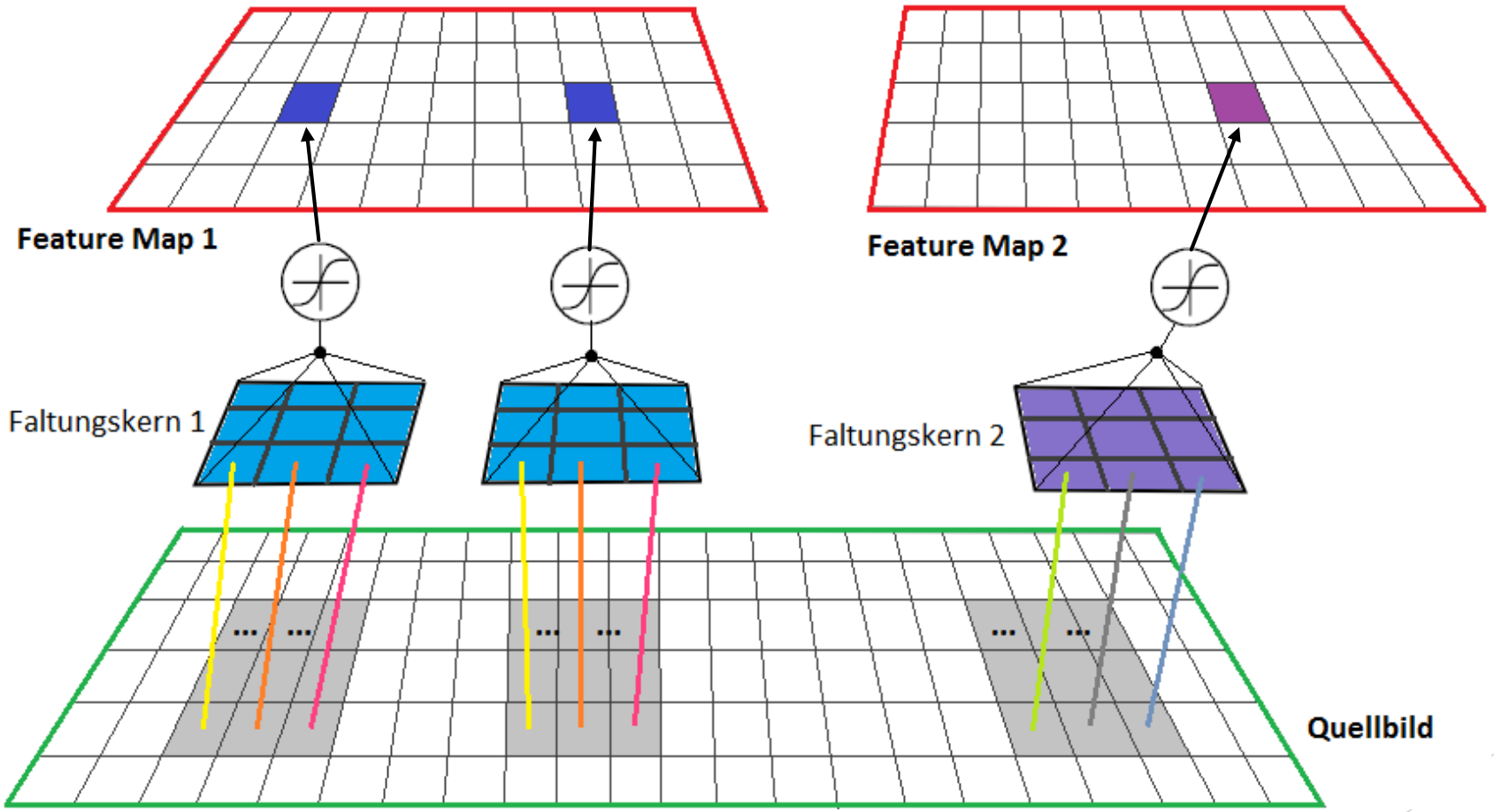
Faltung

- Jeder Zielbildgrauwert wird aus dem entsprechenden Quellbildausschnitt berechnet
- Anwendungsbeispiele:
  - Bildglättung
  - Bildschärfung
  - Kantenfilter

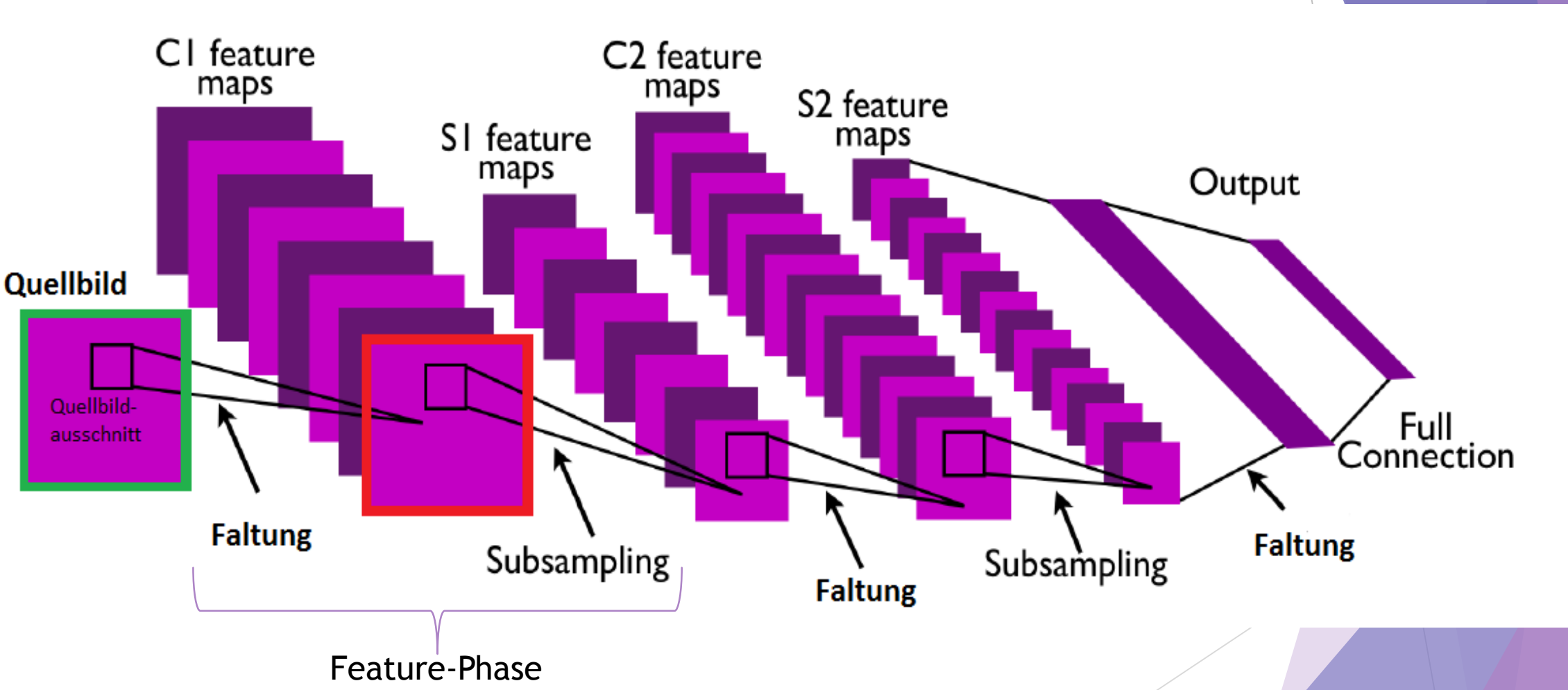


A. Meisel. Vorlesungsskript „Robot Vision“ 2012

Faltung im Neuronalen Netz



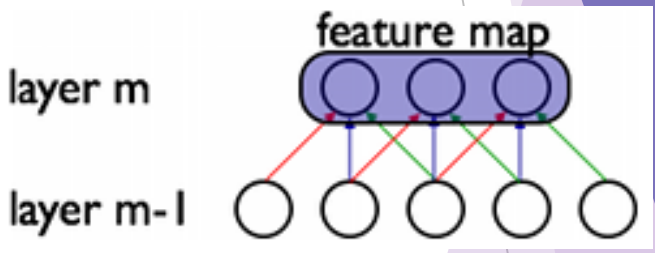
Faltungsnetzarchitektur



Layer der Feature-Phasen (1/3)

Filter Bank Layer:

- Das Netzwerk lernt eine Menge von n Faltungskernen
- Diese werden auf die überlappenden Quellbildausschnitte angewendet
- Dadurch entstehen n Feature Maps
- Ein Feature Map besteht aus Neuronen, die:
  - dieselbe Parameter (Gewichte und Bias) haben
  - dazu dienen, auf ein und dasselbe Feature (dt. Merkmal) zu reagieren
  - und zwar unabhängig von der Position im Quellbild!
- n Feature Maps bilden den Filter Bank Layer



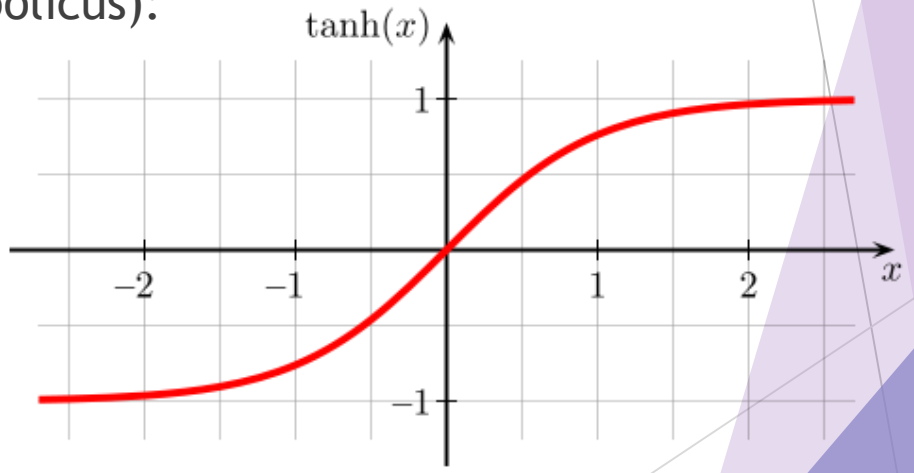
[http://deeplearning.net/tutorial/lenet.html]

Layer der Feature-Phasen (2/3)

Non-Linearity Layer

- Hier wird eine nichtlineare Aktivierungsfunktion auf die Ausgabe des Filter Bank Layers punktweise angewendet
- Typischerweise Sigmoidfunktion (Tangens Hyperbolicus):

$$\tanh(t) = \frac{e^{t} - e^{-t}}{e^{t} + e^{-t}}$$



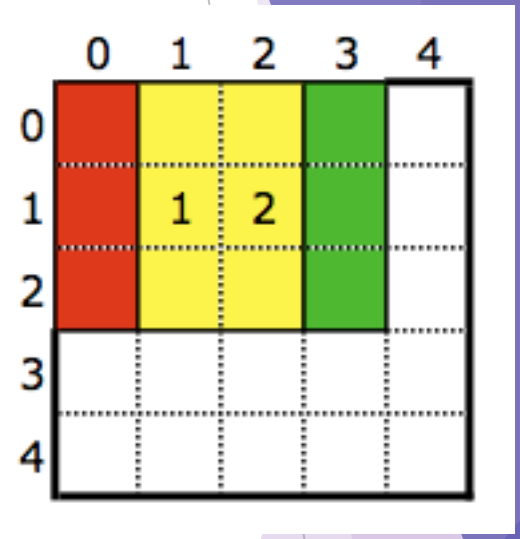
[http://de.wikipedia.org/wiki/Datei:Hyperbolic_Tangent.svg]

- Dies führt zu schnellerem Training

Layer der Feature-Phasen (3/3)

Feature Pooling Layer

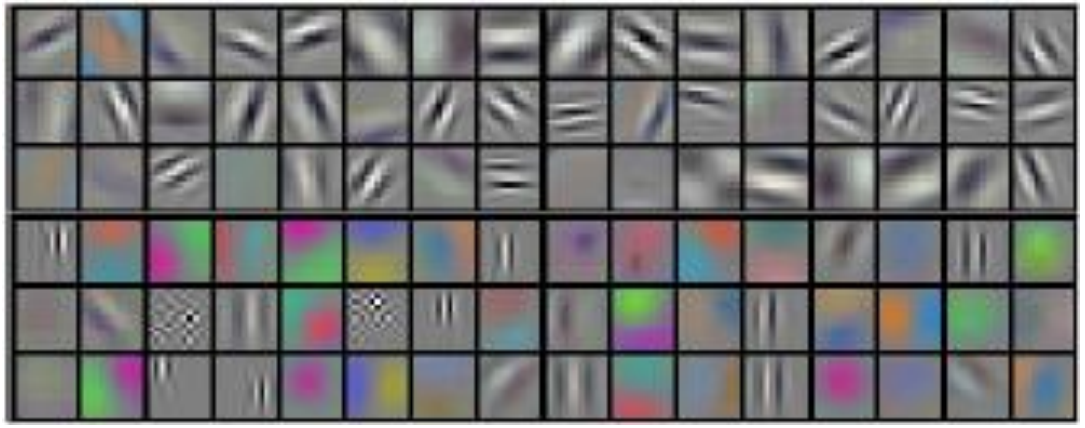
- Entspricht dem „*Subsampling*“
- Jedes Feature Map wird nun separat betrachtet
- Man berechnet typischerweise den mittleren oder maximalen Wert der Nachbarschaft
- mit dem Schritt von 1 Pixel bis zur Nachbarschaftsgröße
- Das Ergebnis = ein neues Feature Map
  - mit einer kleineren Auflösung
  - Robust zur kleinen Lageverschiebungen des Features



[http://stackoverflow.com/questions/5923696/efficient-2d-mean-filter-implementation-that-minimises-redundant-memory-loads]

Beispiel (1/2)

- Gelernte Faltungskerne der ersten Feature-Phase:

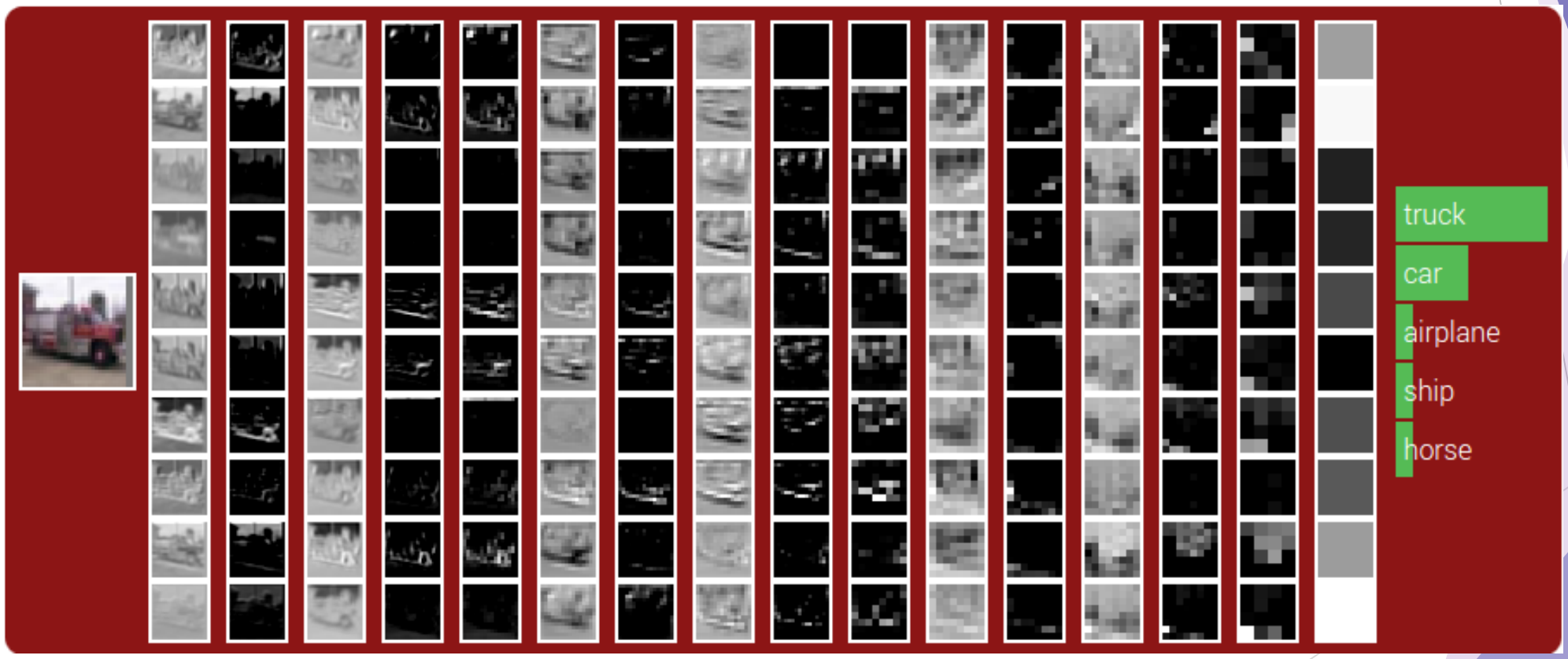


[1]

Beispiel (2/2)

- Feature Maps in verschiedenen Phasen:

[http://vision.stanford.edu/teaching/cs231n/]



Training des Faltungsnetzwerkes

- Überwachtes Lernen
- D.h. es werden Bilder mit der bekannten Klasse geliefert
- Gewichte werden in Richtung kleinsten Fehler geändert
  - „bergab im Fehlergebirge“
- Pro Trainingsschritt (Muster) werden die Faltungskerne aller Schichten aktualisiert:
  - Vom Output-Layer zum Input-Layer rückwärts

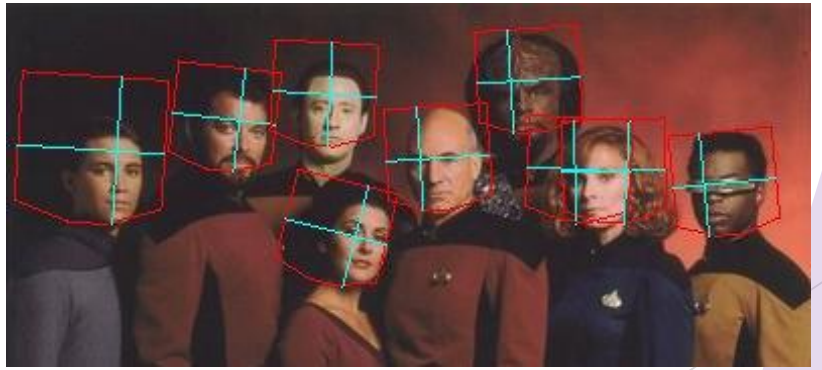
MLP vs. CNN

MLP

- Funktionsapproximation
- Preprocessing nötig
- Ohne Abstraktionsfähigkeit
- Nicht möglich:

CNN

- Gesichtserkennung
- Vierbeinererkennung inkl. Rasse/Gattung/Art
- Robustheit bzgl. geometrischen Transformationen



[http://yann.lecun.com/exdb/publis/index.html]

Aktueller Forschungsstand

Faltungsnetze heute

- Schneiden bei allen namhaften Wettbewerben für Objektklassifikation am besten ab:
  - Bei ImageNet ILSVRC-2012 → 16,4 % Fehlerrate (2. Platz: 26,2 %) [4, 11]
  - Bei ILSVRC 2013 → 11,7 % (2. Platz: 13,0 %)
  - Bei ILSVRC 2014 alle Top-Ergebnisse waren mithilfe der CNNs erreicht (1. Platz: 6,7 % Fehler)
  - Bei MNIST-Datenbank beträgt die Fehlerrate sogar 0,23 % [10] …
- Gründe dafür:
  - Leistungsfähige GPUs, die schnelleres Training ermöglichen
  - Größere Benchmarks [4, 11]
  - ???

Visualisierung der CNNs

- Mithilfe der Deconvolutional Neuronal Networks möglich [6]
- Beantwortet die Fragen:
  - Warum wird ein Objekt richtig erkannt?
  - Welche Bildbereiche sind dafür verantwortlich?
- Ermöglicht:
  - die Suche nach besseren Architekturen
  - Performanzanalyse der Schichten
  - besseres Verständnis der internen Abläufe

[6]

Ziele

Ziele Projekt 1

- Tieferes Verständnis der CNN verschaffen
- Trainingsumgebung kennenlernen und erweitern
- Erste Implementierung auf vorhandenen Daten (MNIST, ImageNet)
- Was gewinnt man durch Nutzung von 3D-Daten?
- Risiken:
  - Eventuell Wechsel der Arbeitsumgebung notwendig (Lush, Matlab, C++)
  - Trainingsdaten beschaffen (Menge, Qualität, Aussagekraft)

Ziele Masterarbeit

- Auswirkung verschiedener Parameter des Faltungsnetzwerkes feststellen
  - Architektur/Konfiguration
  - Schichten
  - Trainingsset
  - Filterparameter usw.
- Fehlertoleranz verbessern
- Auf den Benchmarks bessere Ergebnisse erzielen

http://funny-pictures.picphotos.net/sie-sind-hier-mobil-fun-x-rahmenprogramm-bogenschie-en/mobil-fun-x.de*images*articles*21e4843551f35eba5512b1b1ea8185f5_4.jpg/

Literatur (1/2)

1. Krizhevsky, A., Sutskever, I., and Hinton, G.E. Imagenet classfication with deep convolutional neural networks. In NIPS, 2012.
2. LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. Neural Comput., 1(4):541-551, 1989.
3. LeCun, Y., Kavukvuoglu, K., and Farabet, C. Convolutional Networks and Applications in Vision, Proc. International Symposium on Circuits and Systems (ISCAS'10), IEEE, 2010.
4. O. Russakovsky et al. ImageNet Large Scale Visual Recognition Challenge. In CoRR, 2014.

Literatur (2/2)

5. M. Zeiler. Hierarchical Convolutional Deep Learning in Computer Vision. Diss. PhD, 2014.
6. M. Zeiler and R. Fergus. Visualizing and understanding convolutional neural networks. In ArXiv, 2013.
7. http://pascallin.ecs.soton.ac.uk/challenges/VOC/
8. http://www.vision.caltech.edu/Image_Datasets/Caltech256/
9. http://www.cs.toronto.edu/~kriz/cifar.html
10. http://yann.lecun.com/exdb/mnist/
11. http://www.image-net.org/challenges/LSVRC/2014/

Benchmarks für Objektklassifizierung

- Pascal VOC datasets (ca. 10.000 Bilder, 20 Klassen, [7])
- Caltech256 (60.000 Bilder, 256 Klassen, [8])
- CIFAR (60.000 Bilder, 10 oder 100 Klassen, [9])
- MNIST database (70.000 Bilder, [10])
- ILSVRC dataset (1,2 Mio. Bilder, 1000 Klassen, [11])

Fragen ???