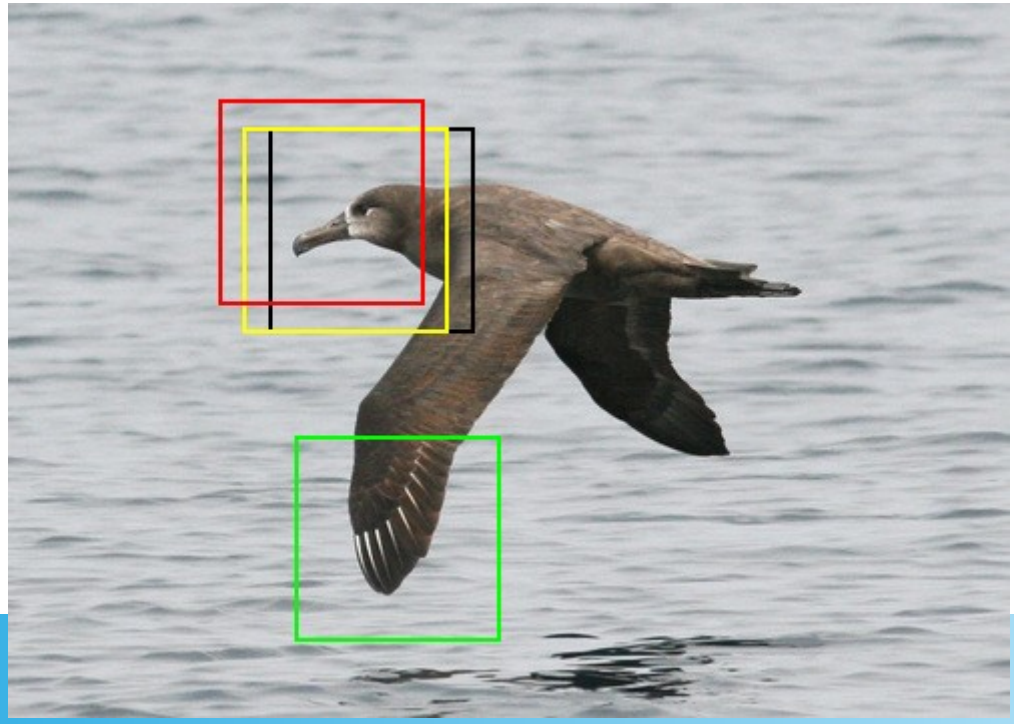
Learning Multi-Attention CNNs for Fine Grained Image Recognition

Multi-Attention CNNs

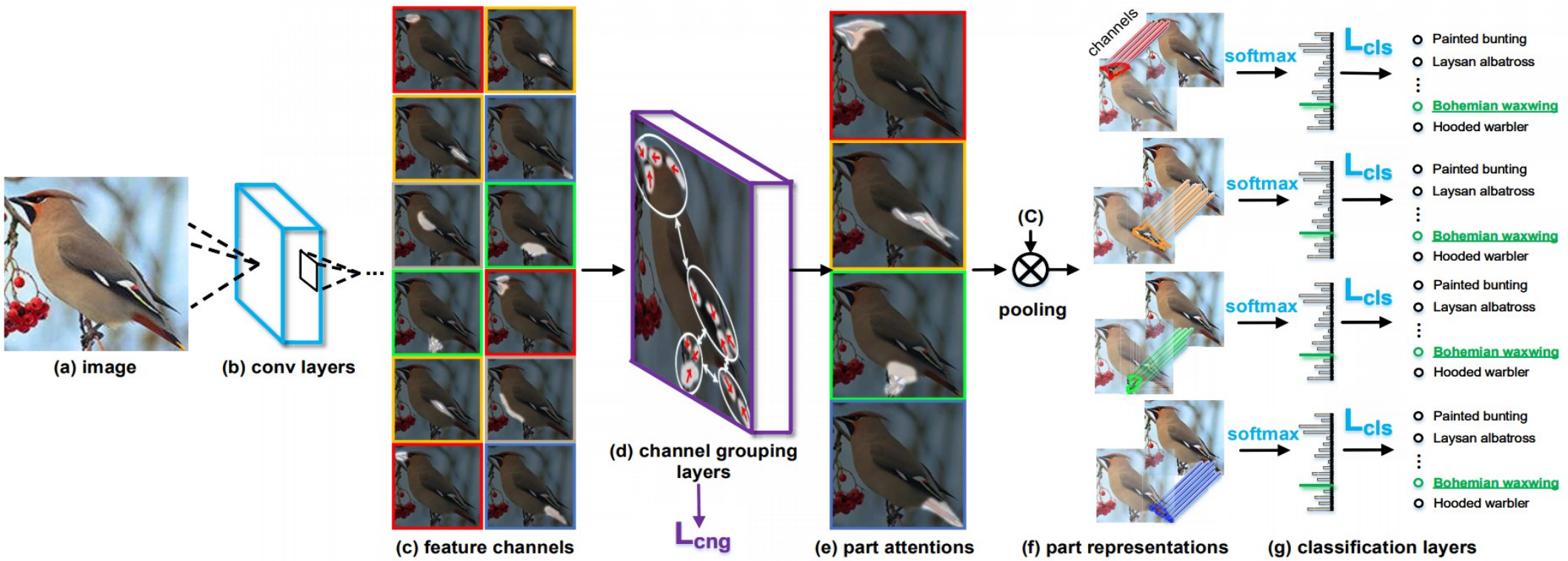
- Eingabe des Originalbildes plus mehrerer Aufmerksamkeitsspots
- Soll Klassifikation erleichtern



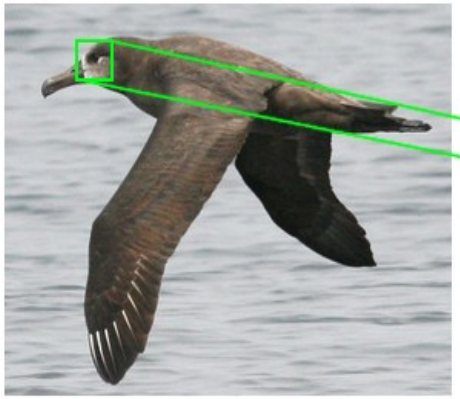
Fine Grained Image Recognition

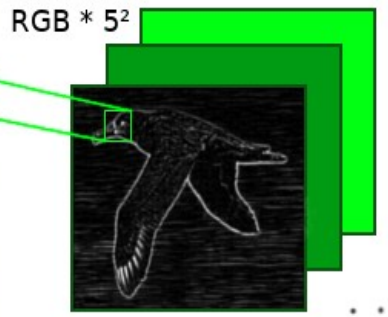
Statt grundlegend verschiedenen Dingen nur etwas unterschiedliche:

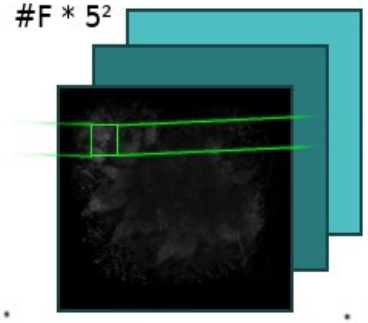
Das Netzwerk

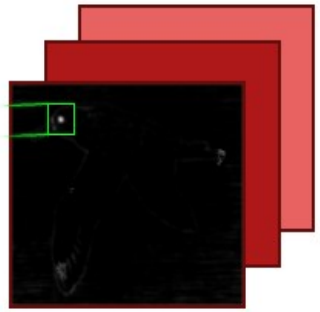


Convolutional Neural Networks



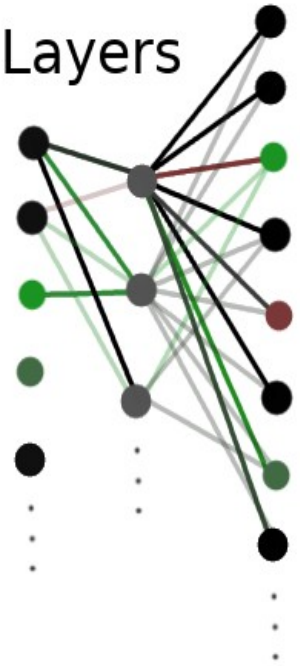






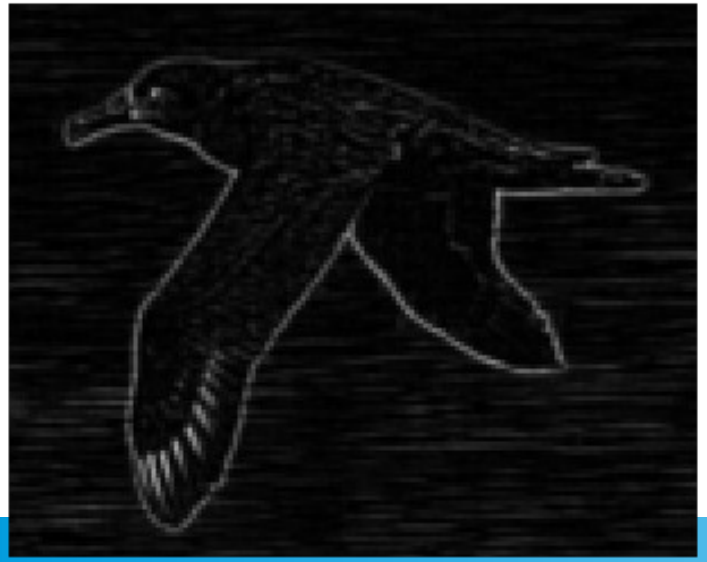
$$\begin{pmatrix} 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 4 & -1 & 0 \\ -1 & 4 & -4 & 4 & -1 \\ 0 & -1 & 4 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 \end{pmatrix}$$

oft tiefere Schichten kleiner,
dafür mehr an der Zahl



Pooling

- Oft Maximum der Pixel-Nachbarschaft
- Zur Vergrößerung der Nachbarschaft der Features, für bedeutendere Features
- Ohne mehr Ebenen weniger Inputs für die FC-Layer

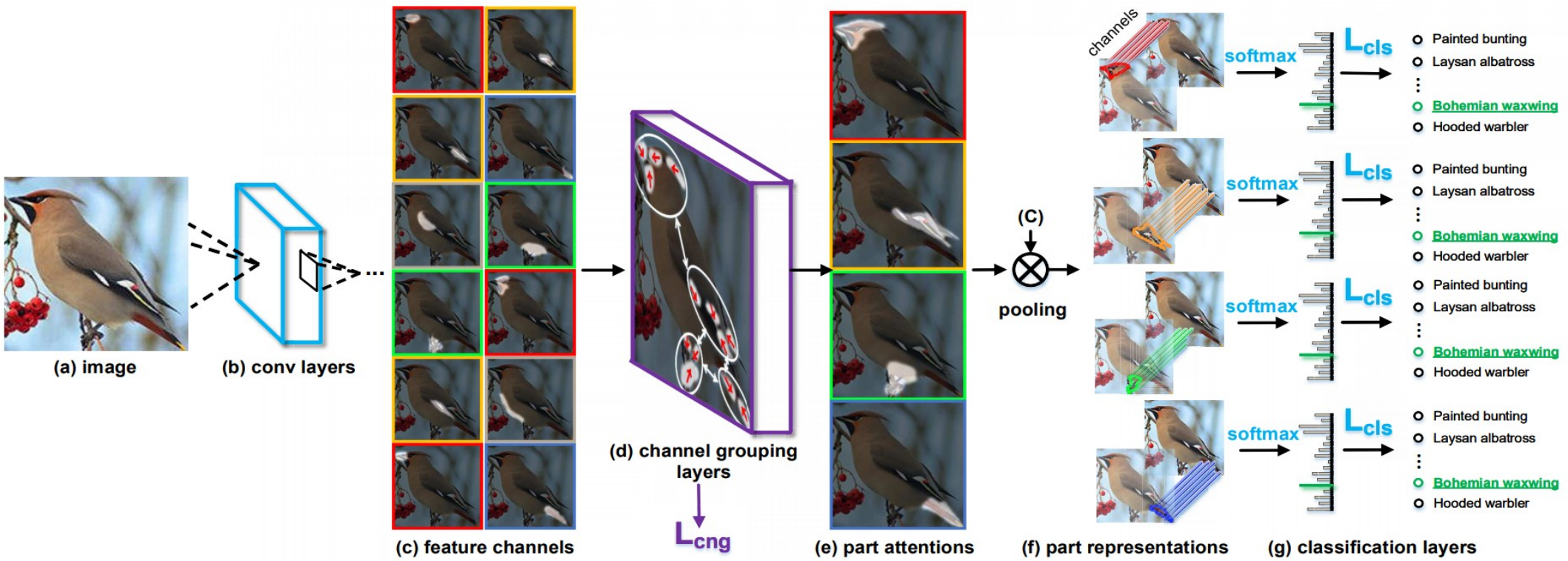


Deep Dream

- Verstärkung von Features durch Backpropagation
- Ermöglicht abgewandelt auch Stilübertragungen

Selbst ausprobieren: deepdreamgenerator.com

Das Netzwerk



Klassifikation

- I.d.R. entscheidet der stärkste Ausschlag
- Einzelner FC-Layer ist wie Korrelationstabelle
- Zum Trainieren (Optimieren) Fehlerfunktion erforderlich: Softmax gerne verwendet, Werte W täuschen allerdings Sicherheit vor:

$$W_k = \frac{\exp\left(NetOutput_k\right)}{\sum_i \exp\left(NetOutput_i\right)}$$

$$K = \underset{k}{argmax}\, W_k = \underset{k}{argmax}\, NetOutput_k$$

Learning Multi-Attention Convolutional Neural Network for Fine-Grained Image Recognition

Heliang Zheng[1*], Jianlong Fu[2], Tao Mei[2], Jiebo Luo[3]

[1]University of Science and Technology of China, Hefei, China

[2]Microsoft Research, Beijing, China

[3]University of Rochester, Rochester, NY

[1]zhenghl@mail.ustc.edu.cn [2]{jianf, tmei}@microsoft.com [3]jluo@cs.rochester.edu

Das Paper

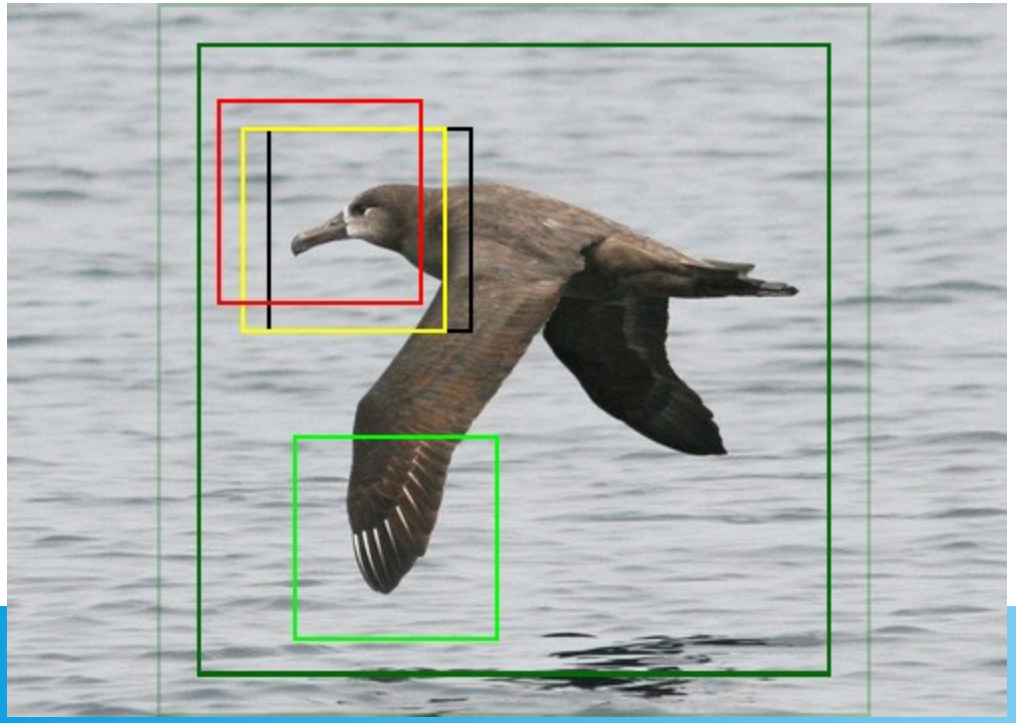
Abstract

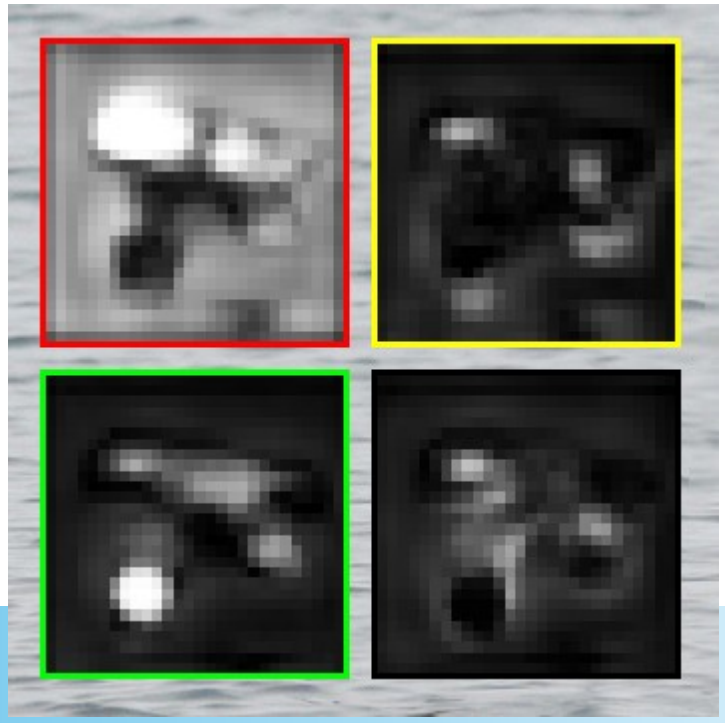
Recognizing fine-grained categories (e.g., bird species) highly relies on discriminative part localization and part-based fine-grained feature learning. Existing approaches predominantly solve these challenges independently, while neglecting the fact that part localization (e.g., head of a bird) and fine-grained feature learning (e.g., head shape) are mutually correlated. In this paper, we propose a novel part learning approach by a multi-attention convolutional neural network (MA-CNN), where part generation and feature learning can reinforce each other. MA-CNN con-

Heliang Zheng (Hefei Uni.),
Jianlong Fu (Microsoft Research, Beijing),
Tao Mei (Microsoft Research, Beijing),
Jiebo Luo (Rochester Uni., New York)

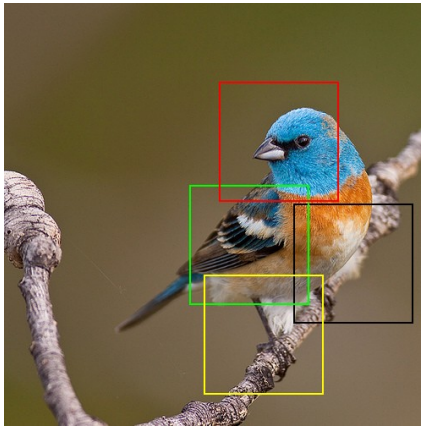
Bestimmung der Spots

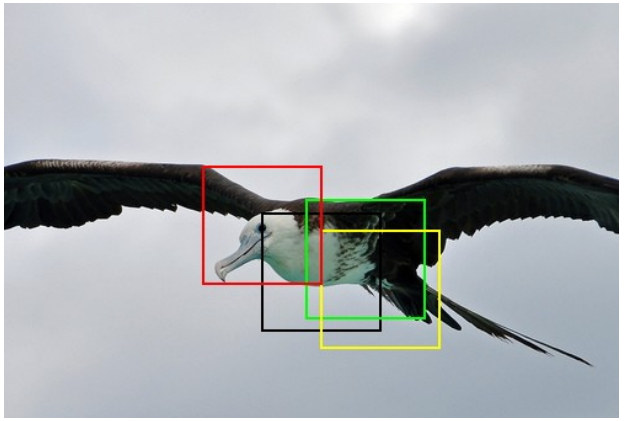
- Bisher oft durch Menschen annotiert
- Im Paper nun Aufgabe des Netzwerkes, die Spots auszuwählen: über Heatmaps

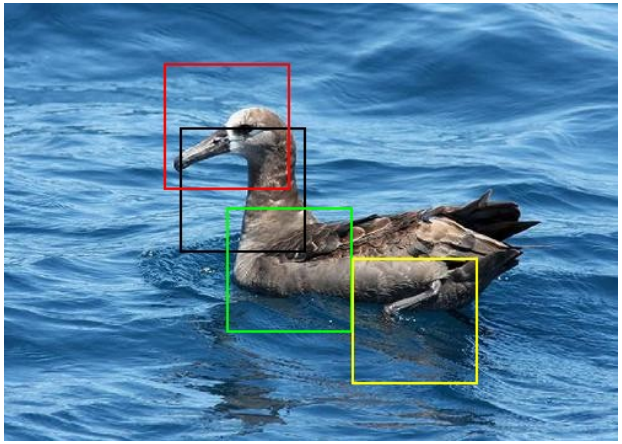


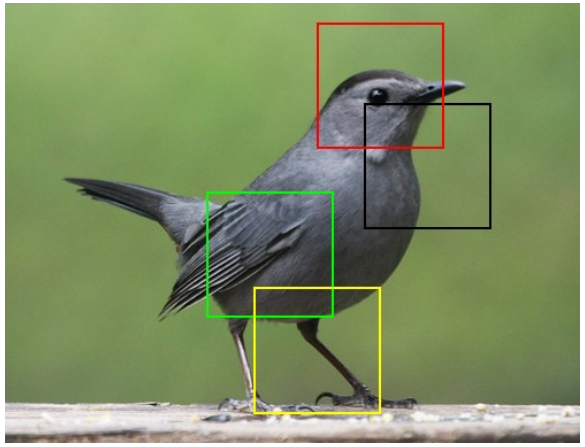


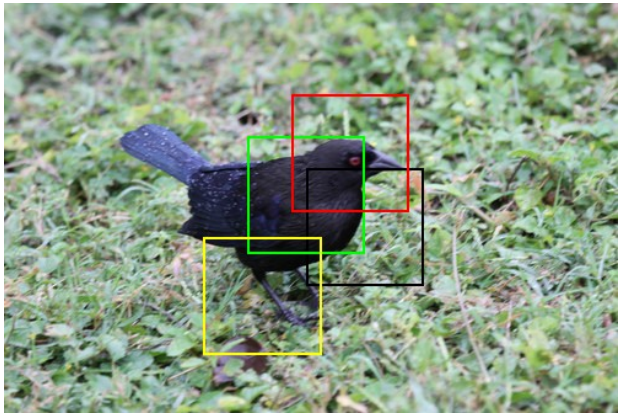
Weitere Beispiele

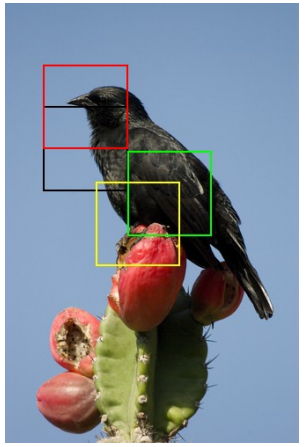












Lernen der Gewichte

- Zwei Möglichkeiten, die sich jedoch ähneln:

  1. Abwechselnd Aufmerksamkeits-Netzwerk und Klassifikator, oder

  2. End-to-end, d.h. im Ganzen

- Fehlerfunktion entscheidend, da sie Lernaufgabe vorgibt:

  Was sind gute Ausschnitte?

- CNN-Initialisierung per

  vortrainiertem ImageNet (VGG-19)

Fehlerfunktion

- Klassifikator-Fehler über Softmax
- Aufmerksamkeitsnetzwerk möglichst:
  1. wenig Überlappung
  2. lokal, d.h. nicht zu breit
- Realisierung durch Linearkombination von:
  1. $L_{\text{div}} = \sum_{x,y} A_i(x,y) \cdot \max_{j \neq i} \left(A_j(x,y) - margin \right)$
  2. $L_{dst} = \sum_{x,y} A_i(x,y) \cdot \left((x - x_c)^2 + (y - y_c)^2 \right)$

$$L_{cng} = \lambda \cdot L_{\text{div}} + L_{dst} \qquad L = L_{cng} + L_{cls}$$

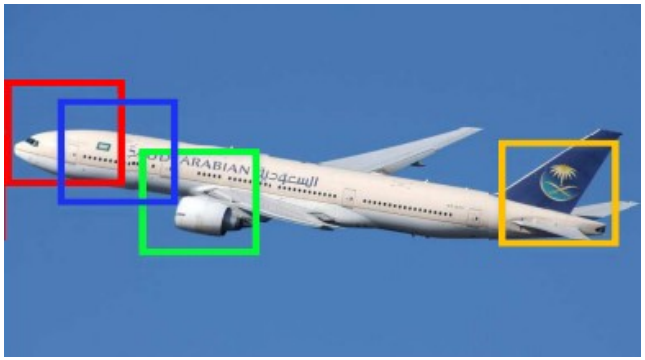
Hier (empirisch): $\lambda = 2, margin = 0{,}02$

Ergebnisse auf CUB 200

Approach	Train Anno.	Accuracy
PN-CNN(AlexNet) [1]	✓	75.7
Part-RCNN(AlexNet) [34]	✓	76.4
PA-CNN [14]	✓	82.8
MG-CNN [27]	✓	83.0
FCAN [18]	✓	84.3
B-CNN (250k-dims) [17]	✓	85.1
Mask-CNN [29]	✓	85.4
TLAN(AlexNet) [31]		77.9
MG-CNN [27]		81.7
FCAN [18]		82.0
B-CNN (250k-dims) [17]		84.1
ST-CNN (Inception net) [10]		84.1
PDFR [35]		84.5
RA-CNN [5]		85.3
MA-CNN (2 parts + object)		85.4
MA-CNN (4 parts + object)		**86.5**

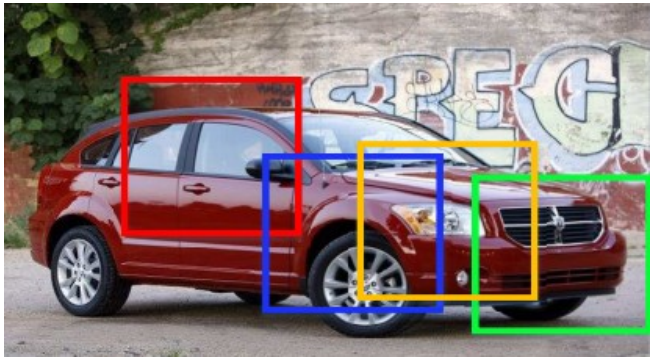
Weitere Ergebnisse

- FGVC Aircraft:



Approach	Train Anno.	Accuracy
MG-CNN [27]	✓	86.6
MDTP [28]	✓	88.4
FV-CNN [7]		81.5
B-CNN (250k-dims)[17]		84.1
RA-CNN [5]		88.2
MA-CNN (2 parts + object)		88.4
MA-CNN (4 parts + object)		**89.9**

- Stanford Cars:



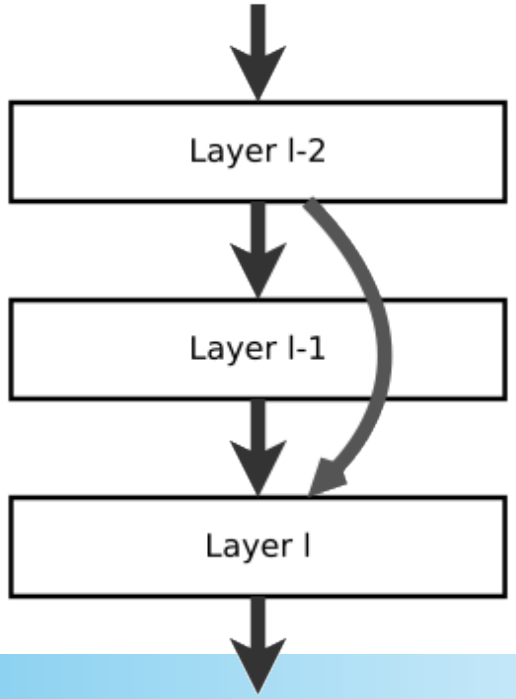
Approach	Train Anno.	Accuracy
R-CNN [6]	✓	88.4
FCAN [18]	✓	91.3
MDTP [28]	✓	92.5
PA-CNN [14]	✓	92.8
FCAN [18]		89.1
B-CNN (250k-dims) [17]		91.3
RA-CNN [5]		92.5
MA-CNN (2 parts + object)		91.7
MA-CNN (4 parts + object)		**92.8**

Verbesserungsideen

- Anpassung der Ausschnitte an die Vogelgröße
- Übergabe des Aufmerksamkeits-Ausschnittes auch für Auto- und Flugzeug-Datensatz
- ResNet oder Inception-v4 statt VGG

Residual Neural Networks

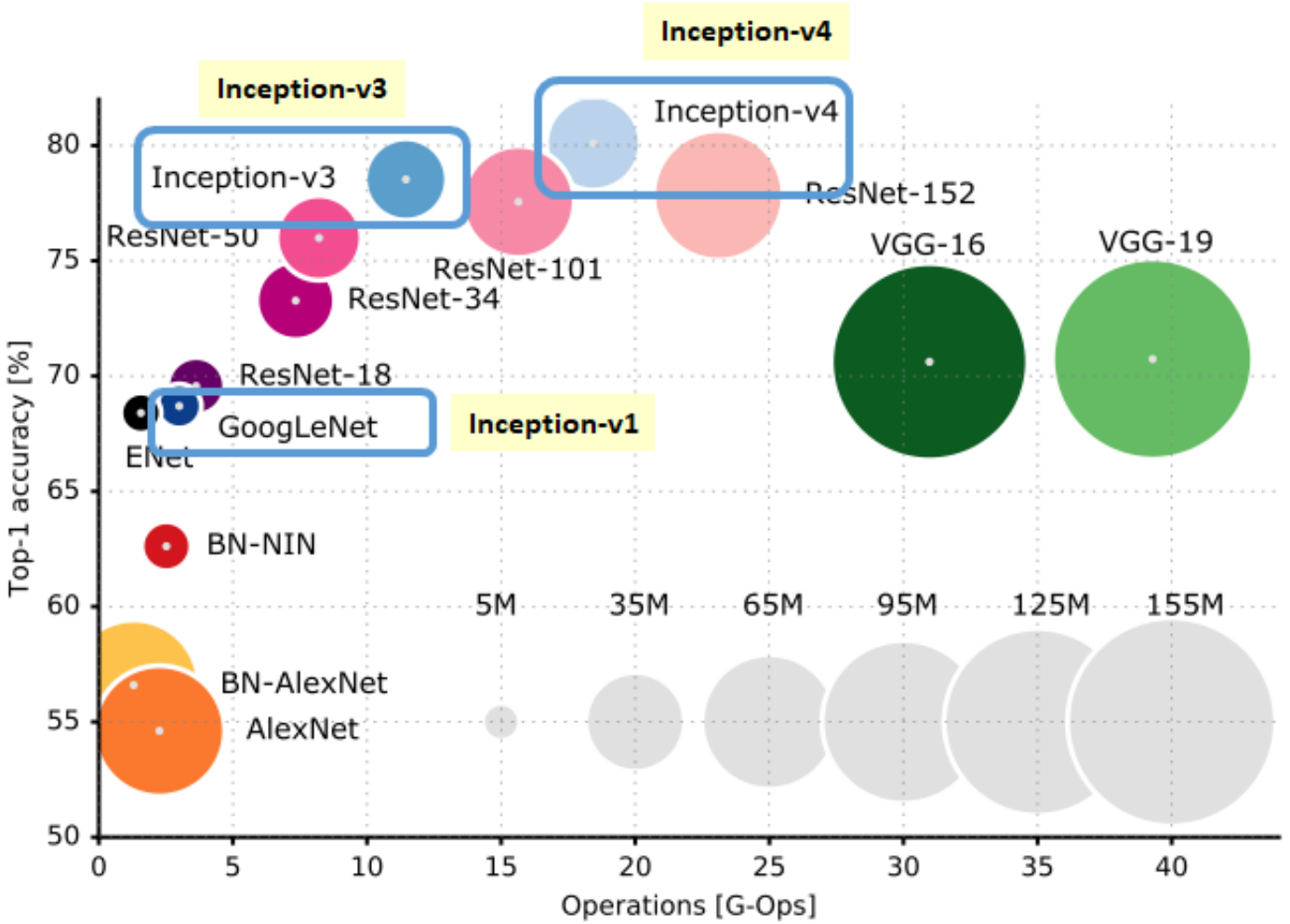
- Information darf Schichten “überspringen”
- Schnelleres Lernen, Lösungsansatz für “Verschwindenden Gradienten” in tiefen Netzen
- Inspiriert durch “Pyramidenzellen” in Großhirnrinde



https://en.wikipedia.org/wiki/File:ResNets.svg

Warum eine andere Architektur?

Erfolg in der ImageNet Large Scale Visual Recognition Challenge



https://cdn-images-1.medium.com/max/800/1*A0JzlOwTokGwhcBhT89tDQ.png

Quellen

Literatur:

- Das Paper selbst
- Verschiedene allgemeine Seiten zu z.B. CNNs wie towardsdatascience.com, Wikipedia

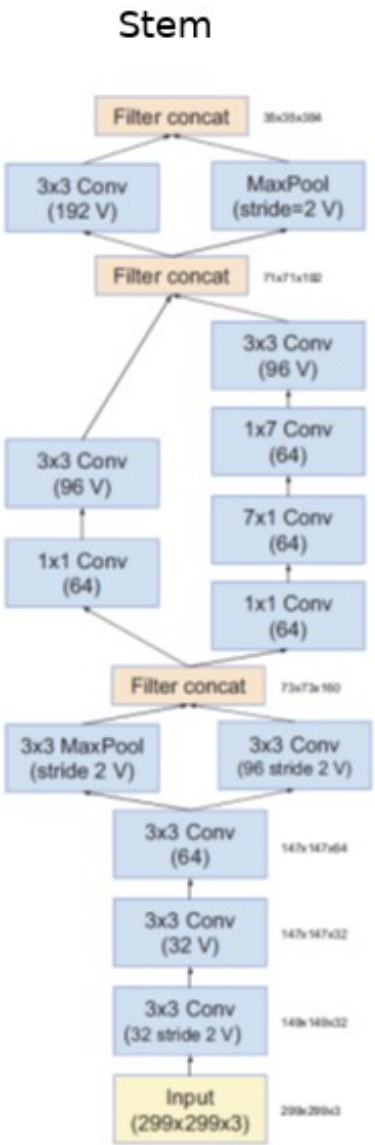
Bilder:

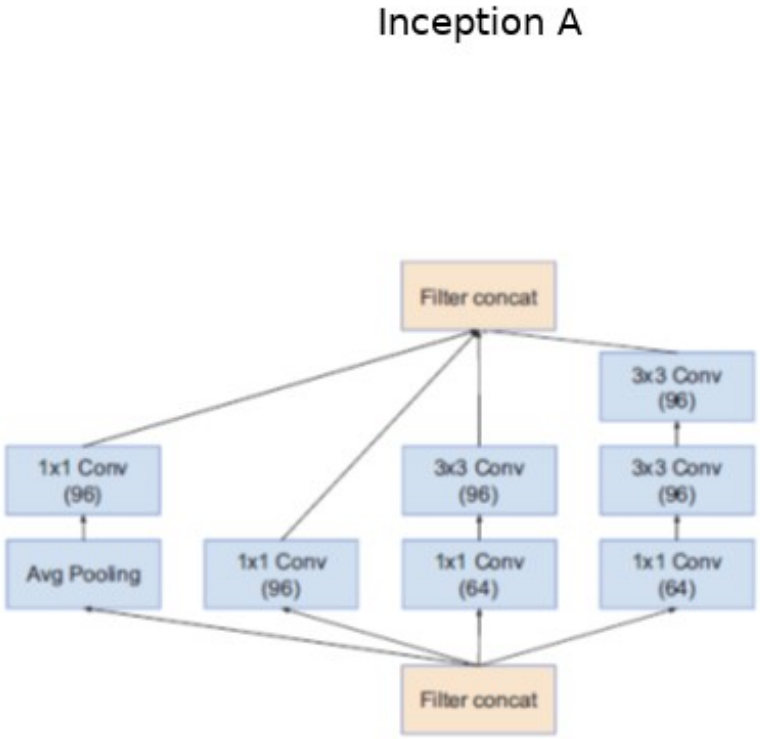
- CUB 200, das Paper
- Als Vorlage https://i0.wp.com/vinodsblog.com/wp-content/uploads/2018/10/CNN-2.png
- Generierte von deepdreamgenerator.com
- Müllauto: https://www.einfach-heidelberg.de/wp-content/uploads/2016/10/M%C3%BCllauto.png
- Waschbär: https://bilder.bild.de/fotos-skaliert/die-ersten-waschbaeren-in-europa-brachen-mitte-des-20-jahrhunderts-aus-zuchtfarmen-aus-und-wurden-aus-200937471-59224494/12,w=8192,q=high,c=0.bild.jpg

ImageNet LSVRC

- ImageNet Large Scale Visual Recognition Challenge
- Lokalisation für 1000 Kategorien
- Detektion für 200 Kategorien
- Detektion aus Video für 30 Kategorien

Inception-v4





https://cdn-images-1.medium.com/max/2560/1*HJ3CNNGz6v76H38s7-OTSA.png

https://towardsdatascience.com/review-inception-v4-evolved-from-googlenet-merged-with-resnet-idea-image-classification-5e8c339d18bc

Inception-v4