Kann man verstehen, wie intelligente Algorithmen entscheiden?

Ist künstliche Intelligenz gefährlich?

Andreas Haller

Universität Heidelberg

24. Mai 2017

1. Vertrauen in Entscheidungen
2. Gefahren
3. Erklärungen zu Entscheidungen
   - lokale Approximation
   - VQA - Visuell Fragen beantworten
4. Täuschung
5. Zusammenfassung

Mensch

- Sinneseindrücke seperat verarbeiten
- linke Gehirnhälfte nutzt Erfahrung zur Vereinheitlichung
- Unterbewusste Entscheidung wird von Interpreter zu Geschichte verarbeitet

Mensch

- Sinneseindrücke seperat verarbeiten
- linke Gehirnhälfte nutzt Erfahrung zur Vereinheitlichung
- Unterbewusste Entscheidung wird von Interpreter zu Geschichte verarbeitet

Computer

- Input (Wort, Bild) in hidden Layers verarbeitet
- mehrere letzte Layer vereinheitlichen
- Begründung der Entscheidung?

Vertrauen:

- Performance des Modells
- Robustheit
- Modell wird verstanden

$\Rightarrow$ Konsens der Anwender & Entwickler:
Interpretierbarkeit ist Grundlage für Vertrauen

1. Transparenz vs. Black-Box

- Konvergenz, eine Lösung, Oberfläche des Fehlers
- Repräsentation von Parametern
- komplett nachvollziehbar
- Ergebnis wiederholbar durch Mensch in annehmbarer Zeit

1. Transparenz vs. Black-Box

- Konvergenz, eine Lösung, Oberfläche des Fehlers
- Repräsentation von Parametern
- komplett nachvollziehbar
- Ergebnis wiederholbar durch Mensch in annehmbarer Zeit

2. nachträgliche Erklärungen

- natürliche Sprache (Merkmale, Bildunterschriften, "Sieht aus wie . . . ")
- Visualisierungen (Repräsentationen, Aufmerksamkeiten)

?Transparent $\Rightarrow\Leftarrow$ Intelligent?

$\Rightarrow$ Fokus auf Erklärungen

1 Vertrauen in Entscheidungen

2 Gefahren

3 Erklärungen zu Entscheidungen
- lokale Approximation
- VQA - Visuell Fragen beantworten

4 Täuschung

5 Zusammenfassung

Grenzen Maschinellem Lernens

- Ziel: Minimierung des Errors $\Rightarrow\Leftarrow$ Komplexität der Realität
- Verallgemeinerung $\widehat{=}$ $\mathcal{L}oss$(Test) - $\mathcal{L}oss$(Training)
  aber Test und Trainings-Set aus gemeinsamer Distribution
- übertriebenes Vertrauen in Performance des Modells
  wegen Erfolg auf Validierungs-Set und nicht auf Realität
- Hohe Genauigkeit $\not\Rightarrow$ wichtige Features erkannt
- ohne Hintergrundinformationen können Prognosen irreführend sein

Diskriminierung

- 1. Problem: Datensätze mit sozialem Bezug
- Lösung:
  -Löschen von Features mit direktem Bezug zur Diskriminierung
  Problem: Korrelationen bleiben erhalten
  -Löschen aller korrelierenden Features
  Problem: Es bleiben keine sinnvollen Features übrig
- 2. Problem: Randgruppen wegen erhöhter Unsicherheit benachteiligt
- Lösung: Repräsentation von Randgruppen erhöhen

Diskriminierung

- 1. Problem: Datensätze mit sozialem Bezug
- Lösung:
  -Löschen von Features mit direktem Bezug zur Diskriminierung
  Problem: Korrelationen bleiben erhalten
  -Löschen aller korrelierenden Features
  Problem: Es bleiben keine sinnvollen Features übrig
- 2. Problem: Randgruppen wegen erhöhter Unsicherheit benachteiligt
- Lösung: Repräsentation von Randgruppen erhöhen

Wichtig:

- Wahl des Datensatzes
- Erklärungen, um Diskriminierung im Entscheidungsprozess auschließen zu können

1 Vertrauen in Entscheidungen

2 Gefahren

3 Erklärungen zu Entscheidungen
- lokale Approximation
- VQA - Visuell Fragen beantworten

4 Täuschung

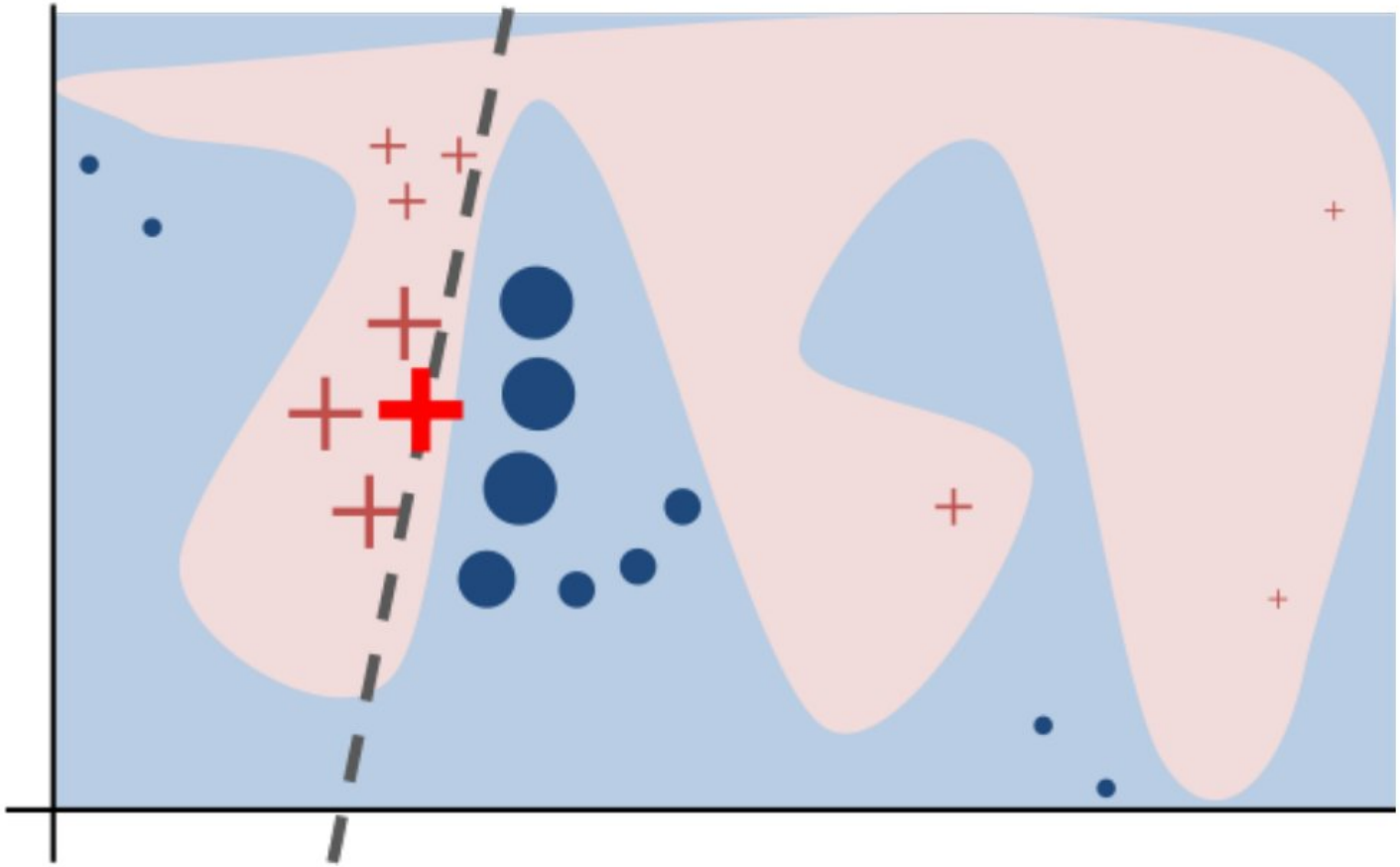
5 Zusammenfassung

LIME - Lokale interpretierbare, modellunabhängige Erklärungen

Ziel: lokale Approximation von Black-Box Modellfunktionen f durch interpretierbare Funktionen G mit Abstandsfunktion Π_x

$$\xi(x) = \arg \min_{g \in G} \underbrace{\mathcal{L}(f, g, \Pi_x)}_{\text{Loss der Approx.}} + \underbrace{\Omega(g)}_{\text{Komplexität}} \tag{1}$$

$$\text{z. B.} \qquad \Omega(g) = \begin{cases} \infty & \#\text{words} > K \\ 0 & \#\text{words} \leq K \end{cases} \tag{2}$$



[Ribeiro et al.(2016)]

$$\mathcal{L}(f, g, \Pi_x) = \sum_{z,z' \in Z} \Pi_x(z)(f(z) - g(z'))^2 \tag{3}$$

$$\text{mit} \quad z' \in N_{eigh.}(x') \tag{4}$$

$$\text{und} \quad t_{rafo}(z') = z \tag{5}$$

und mit einer Definitionsbereichstransformation $t_{rafo} : D(g) \rightarrow D(f)$

Nachteil

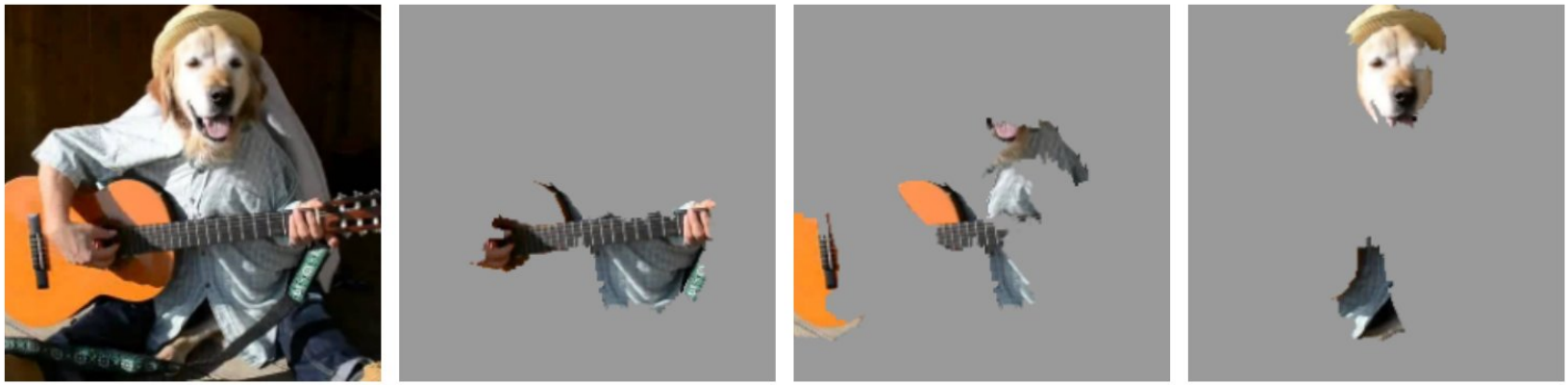
- nicht komplex genug
  Superpixel $\not\Rightarrow$ f(Sepia)= Retro
- f in Umgebung von x komplett nicht-linear
  $\Rightarrow$ g ist Müll

Nachteil

- nicht komplex genug
  Superpixel $\not\Rightarrow$ f(Sepia)= Retro
- f in Umgebung von x komplett nicht-linear
  $\Rightarrow$ g ist Müll

Vorteil

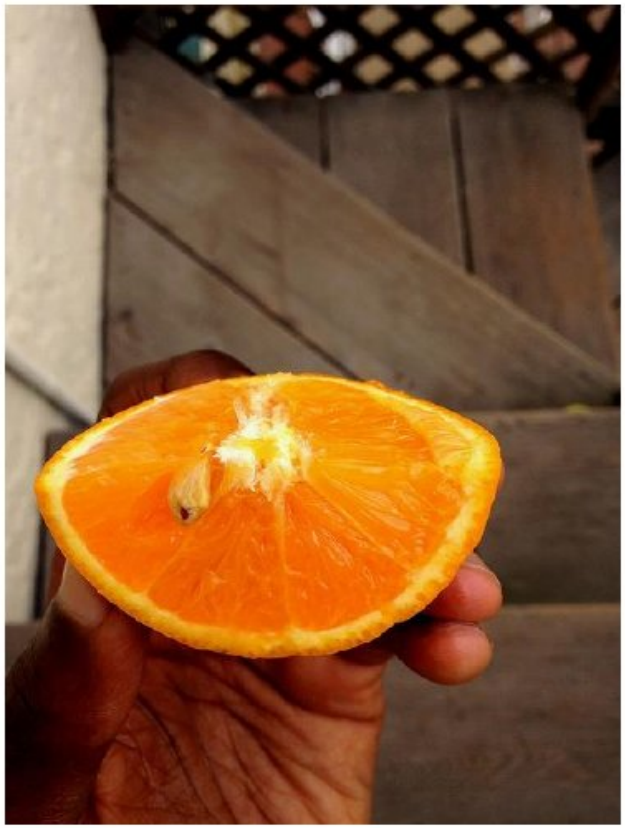
- modellunabhängig
- Erklärung vorhanden
  - Fokus auf Wörter
  - Fokus auf Bildbereiche



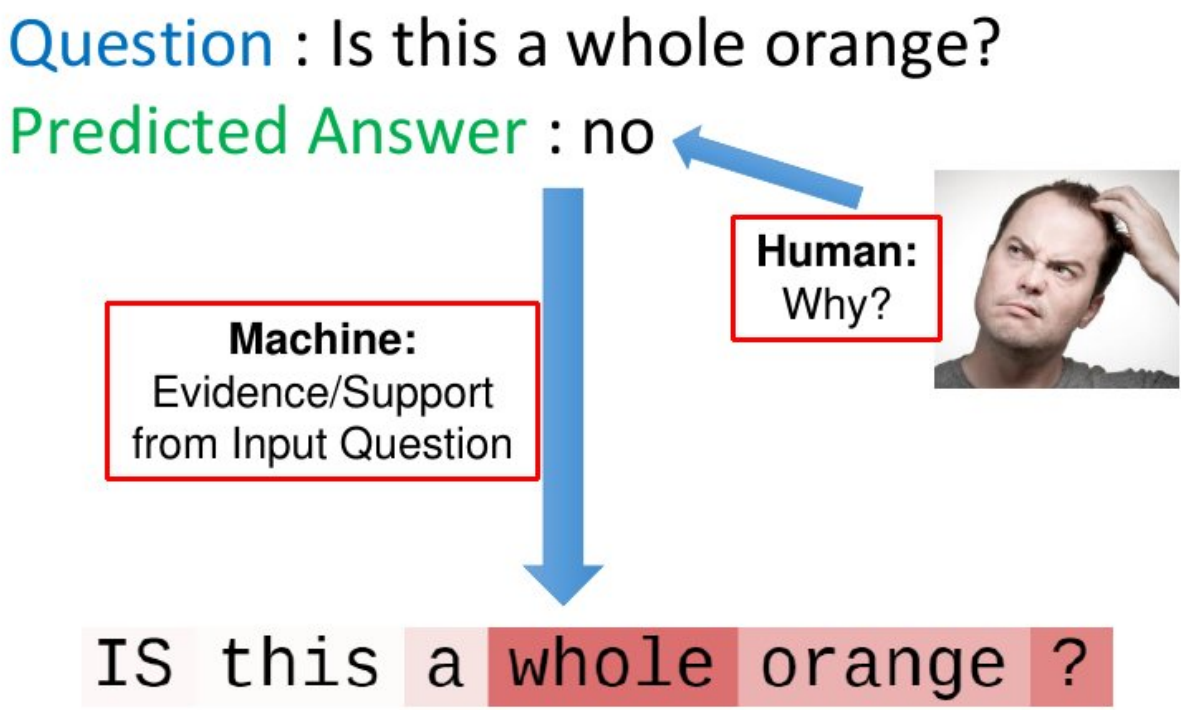
(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Top-3-Prognosen: Elektische Gitarre (p=0,32), Akustische Gitarre (p=0,24) & Labrador (p=0.21). Das Griffbrett erklärt die falsche Prognose für Elektrische Gitarre. Entnommen aus [Ribeiro et al.(2016)]

VQA - Visuell Fragen beantworten



[Goyal et al.(2016)]



Wörter löschen

Question : What vegetable is on the plate ?
Predicted Answer : broccoli

Question : What color is the plate ?
Predicted Answer : white

Question : Is there meat in this dish ?
Predicted Answer : no

Question : Where is the player ?
Predicted Answer : tennis court

Question : What does the man wear on his arms ?
Predicted Answer : tennis racket

Question : What sport is this ?
Predicted Answer : tennis

(a)

Superpixel mit ∅Farbe ersetzen

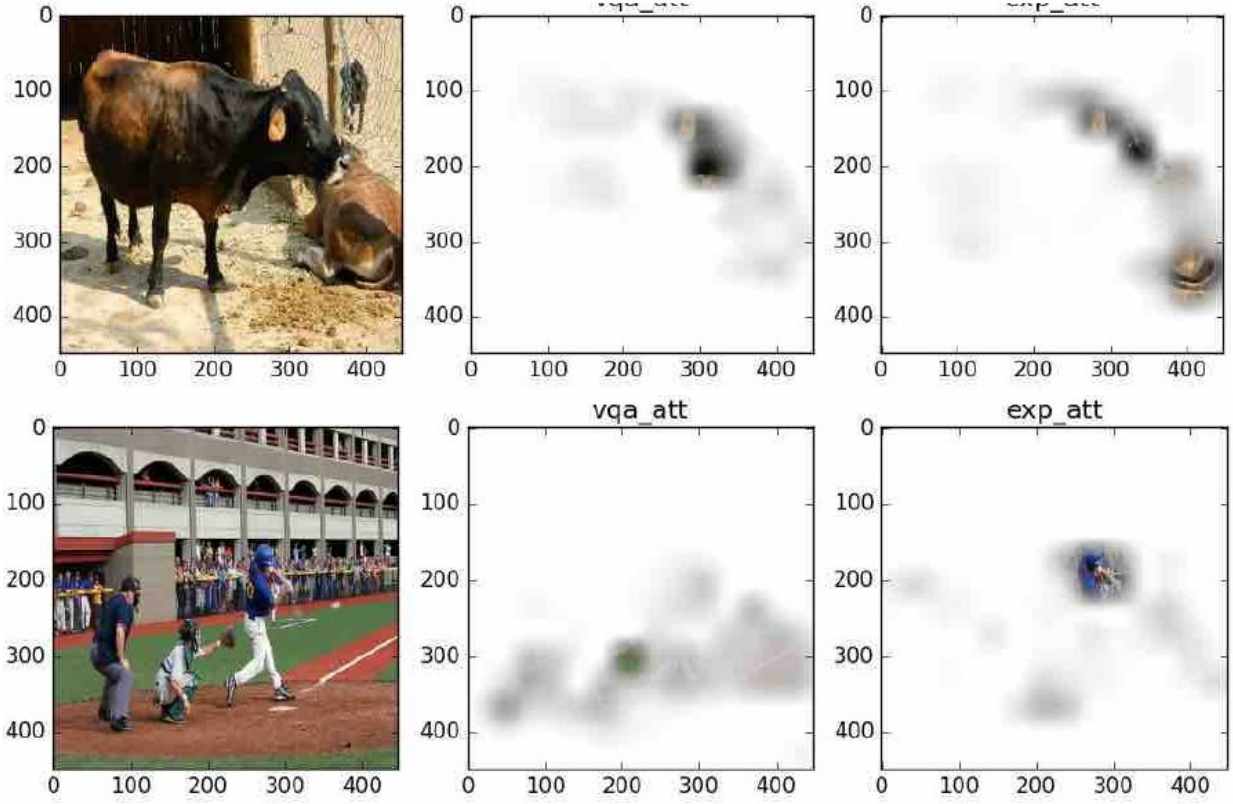
Question : What kind of bird is perched on the sill?
Predicted Answer : parrot

Question : What type of fruit is the plate?
Predicted Answer : banana

(b)

[Goyal et al.(2016)]
Wh-Wörter, Adjektive & Substantive besonders wichtig

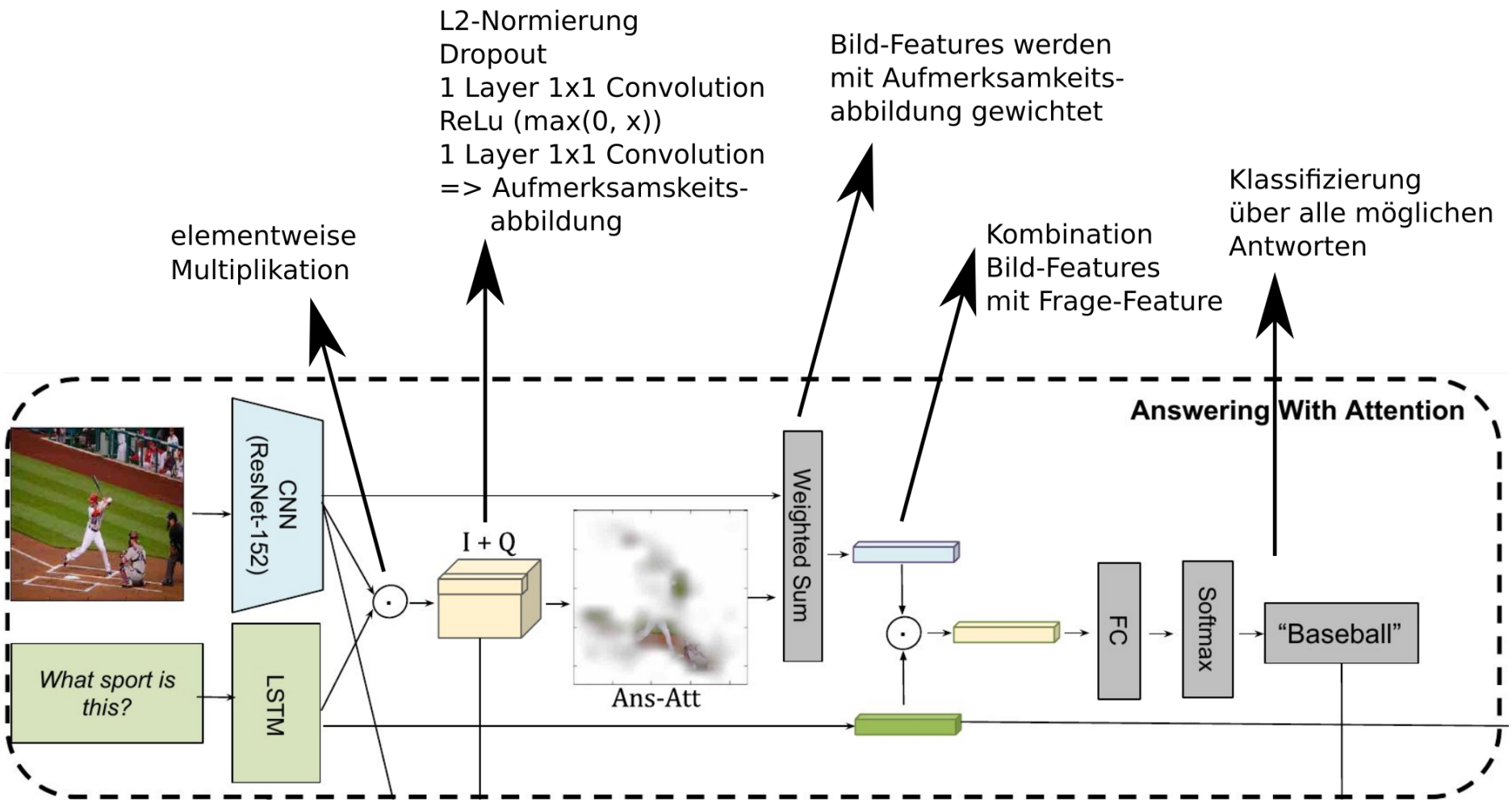
Antwort & Erklärung
jeweils in Sprache und mit Aufmerksamkeitsabblindung



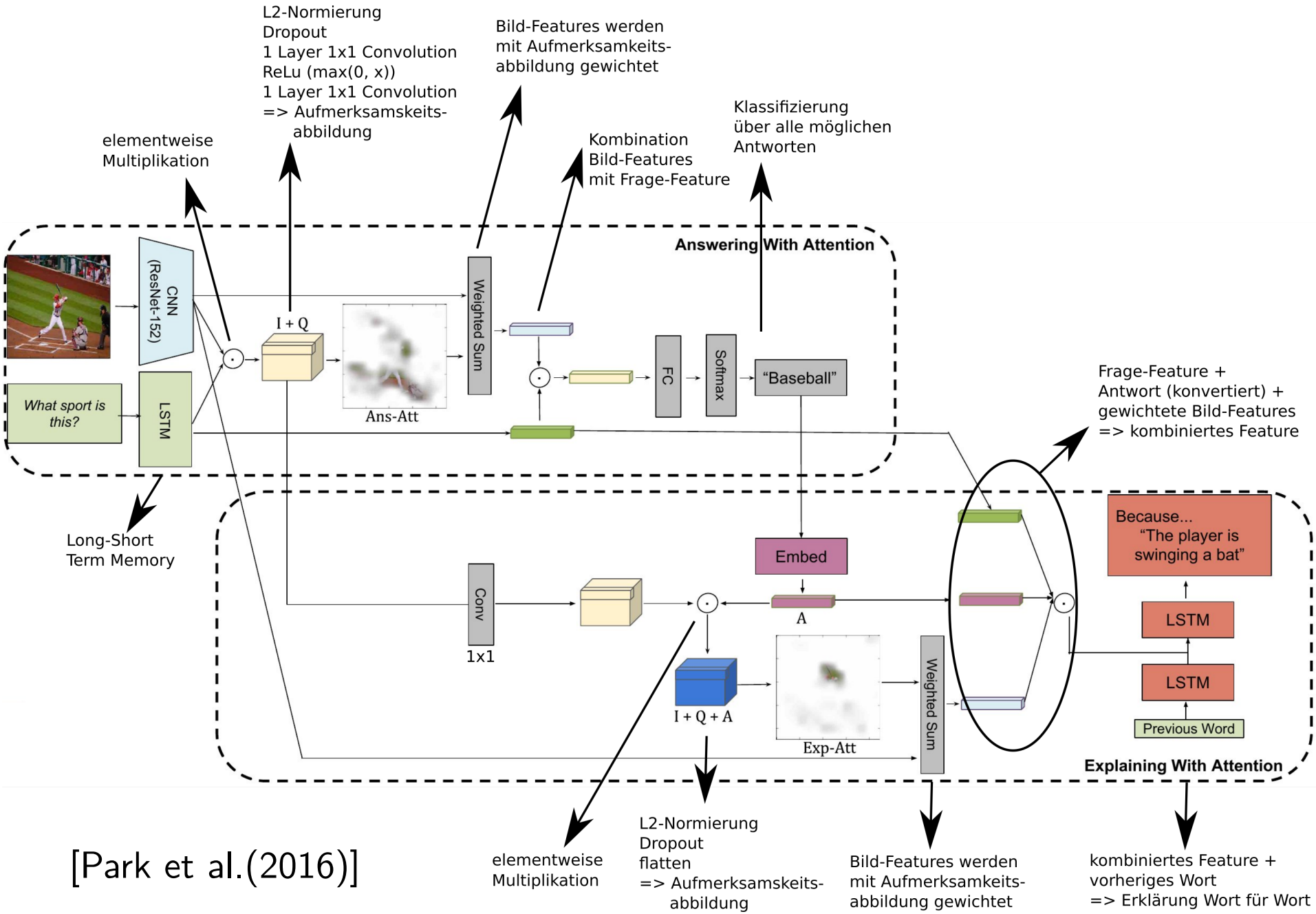
Q: What kind of animal is lying on the ground?
A: Cow. (correct)
E: Because it has four legs and looks like a cow.

Q: What game is this?
A: Baseball. (correct)
E: Because the player is holding a bat.

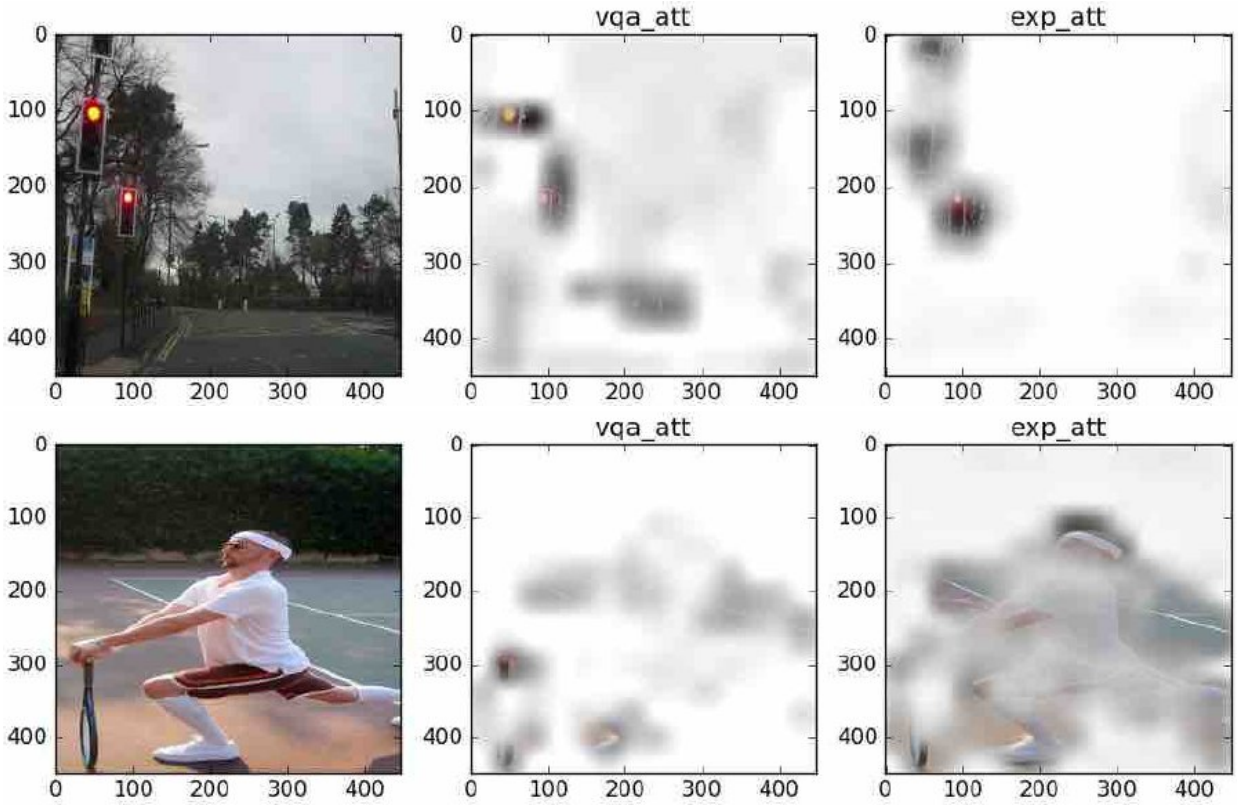
[Park et al.(2016)]



[Park et al.(2016)]



[Park et al.(2016)]



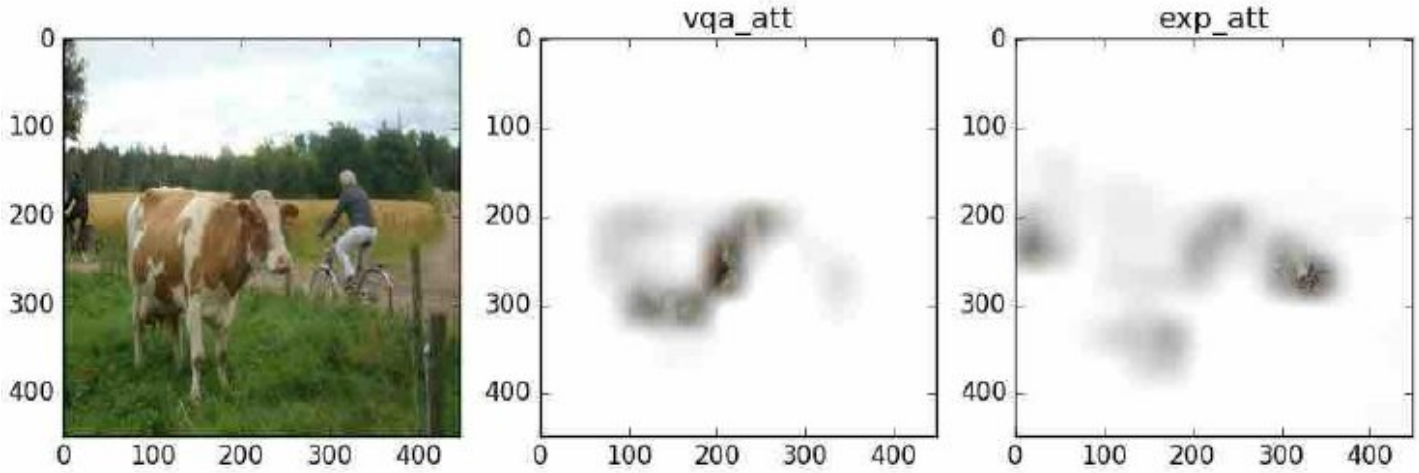
[Park et al.(2016)]

Q: Should we stop?
A: No. (wrong: Yes)
E: Because the light is green.

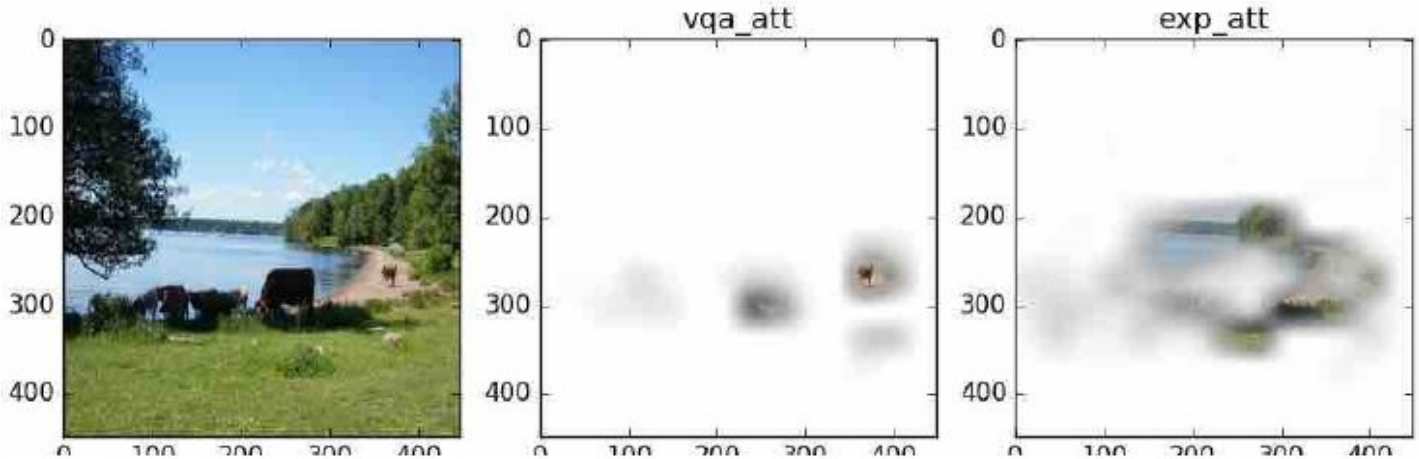
Q: What is the person doing?
A: Playing tennis. (wrong: Streching)
E: Because he is holding a tennis racket.

What kind of animal is this? Cow.

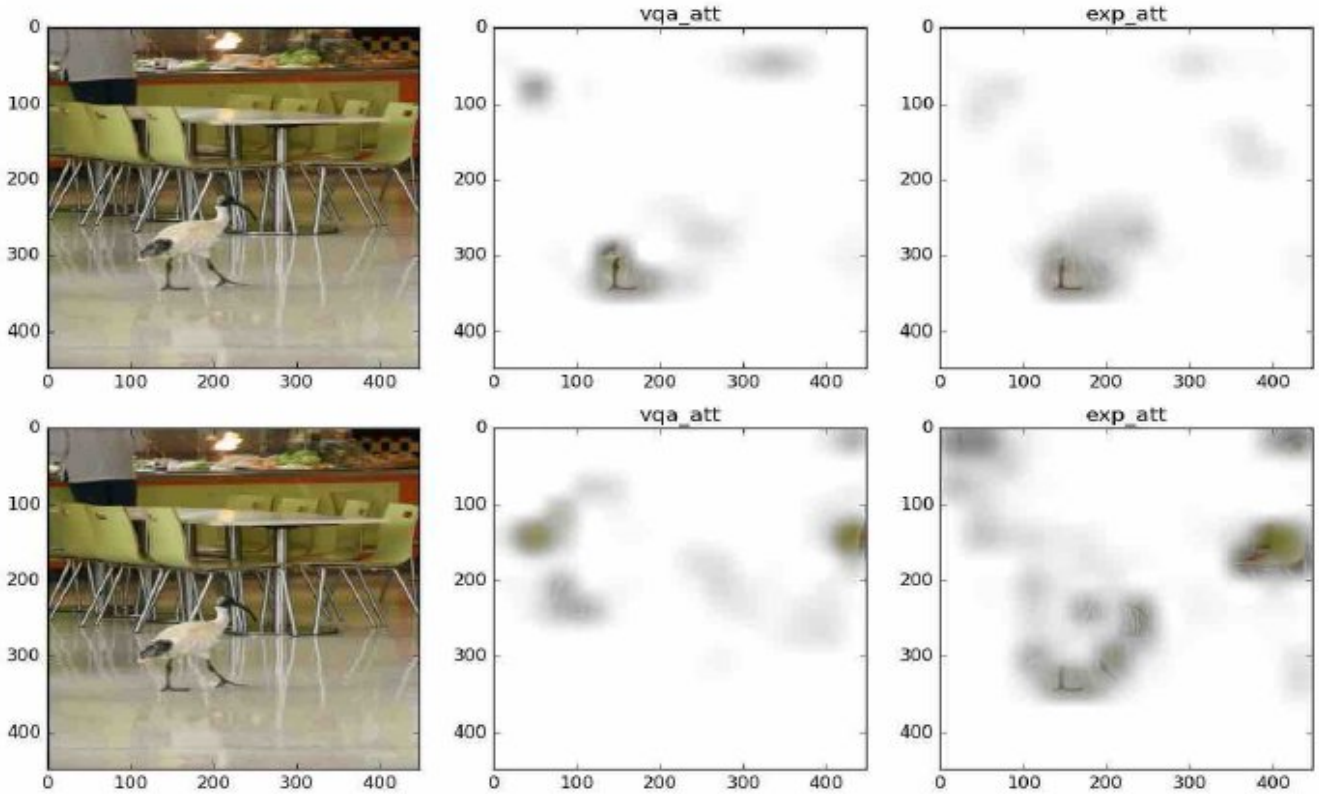
Because it has four legs and looks like a cow.



Because they are grazing in a field like cows.



Zwei Bilder, gleiche Antwort, unterschiedliche Begründungen [Park et al.(2016)]



What is the bird doing?
Walking.
Because they are on the ground.

What is the color of the seats?
Green.
Becuase they color of the trees and forest indicate.

[Park et al.(2016)]

Ein Bilder, zwei unterschiedliche Fragen

Verbesserungen von neuronalen Netzen aufgrund von Erklärungen bei Fehlprognosen

1. Mehr Instanzen hinzufügen
2. Klassifikator ungeeignet $\Rightarrow$ Klassifikator/Modell neu
3. Datensatz ungeeignet $\Rightarrow$ Ersatz
4. Mensch entscheidet & Erklärung ist Unterstützung

1 Vertrauen in Entscheidungen

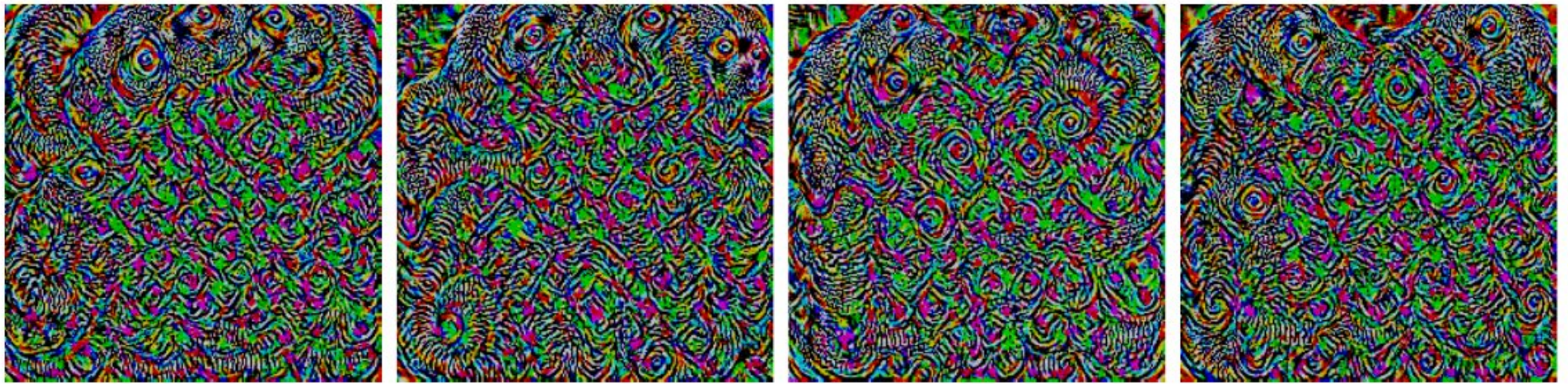
2 Gefahren

3 Erklärungen zu Entscheidungen
- lokale Approximation
- VQA - Visuell Fragen beantworten

4 Täuschung

5 Zusammenfassung

Bisher: individuell pro Bild ein Optimierungsproblem
Jetzt: eine Störung für alle Bilder eines Netzwerks



4 universelle Störungen eines Netzwerks [Moosavi-Dezfooli et al.(2016)]

Algorithmus 4.1 : Computation of universal perturbations from [Moosavi-Dezfooli et al.(2016)]

Data : Data points X,
classifier $\hat{k}$,
desired l_p norm of the perturbation ξ,
desired accuracy on perturbed samples δ

Result : Universal perturbation vector v

Initialize $v \leftarrow 0$.

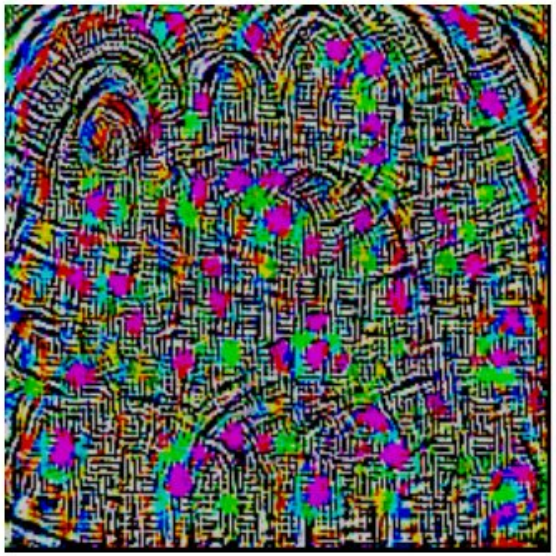
while $Err(X_v) \leq 1 - \delta$ **do**
  for x_i *in* X **do**
    if $\hat{k}(x_i + v) = \hat{k}(x_i)$ **then**
      $\Delta v_i \leftarrow \arg\min_r ||r||_2$ s.t. $\hat{k}(x_i + v + r) \neq \hat{k}(x_i)$
      $v \leftarrow \arg\min_{v'} ||v + \Delta v_i - v'||_2$ subject to $||v'||_p \leq \xi$
    end
  end
end

		CaffeNet [8]	VGG-F [2]	VGG-16 [17]	VGG-19 [17]	GoogLeNet [18]	ResNet-152 [6]
ℓ_2	X	85.4%	85.9%	90.7%	86.9%	82.9%	89.7%
	Val.	85.6	87.0%	90.3%	84.5%	82.0%	88.5%
ℓ_∞	X	93.1%	93.8%	78.5%	77.8%	80.8%	85.4%
	Val.	93.3%	93.7%	78.3%	77.8%	78.9%	84.0%

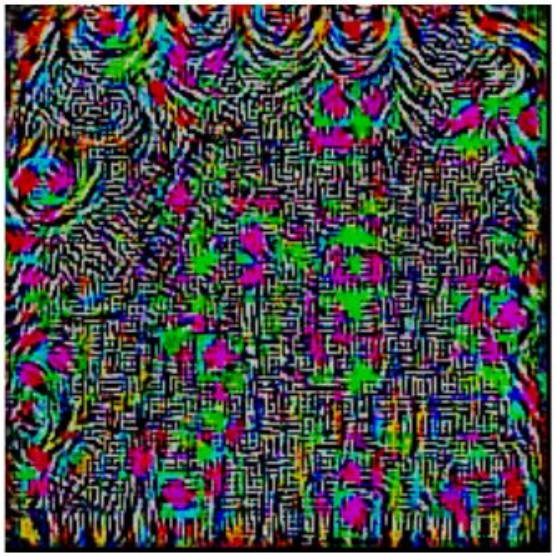
Fehlklassifizierungsraten universeller Störungen auf unterschiedlichen neuronalen Netzwerken.

	VGG-F	CaffeNet	GoogLeNet	VGG-16	VGG-19	ResNet-152
VGG-F	**93.7%**	71.8%	48.4%	42.1%	42.1%	47.4 %
CaffeNet	74.0%	**93.3%**	47.7%	39.9%	39.9%	48.0%
GoogLeNet	46.2%	43.8%	**78.9%**	39.2%	39.8%	45.5%
VGG-16	63.4%	55.8%	56.5%	**78.3%**	73.1%	63.4%
VGG-19	64.0%	57.2%	53.6%	73.5%	**77.8%**	58.0%
ResNet-152	46.3%	46.3%	50.5%	47.0%	45.5%	**84.0%**

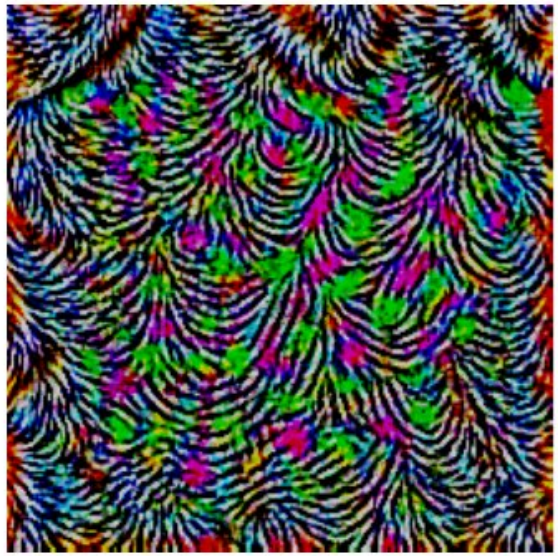
Spalte: universelle Störung aus gegebenem Netzwerk
Zeile: Ergebnis für dieses Netzwerk mit gegebenen Störungen
[Moosavi-Dezfooli et al.(2016)]



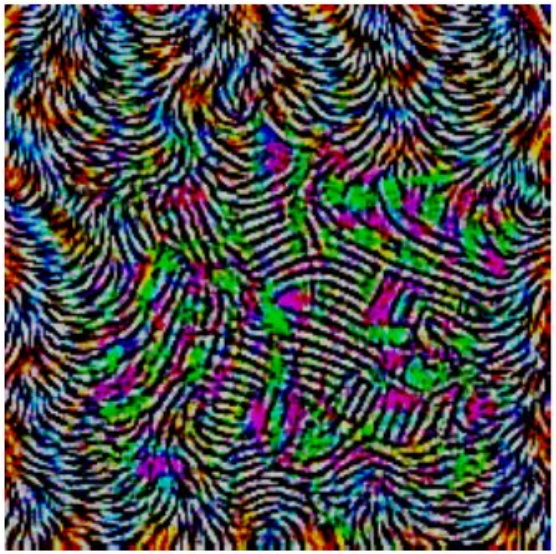
(a) CaffeNet



(b) VGG-F

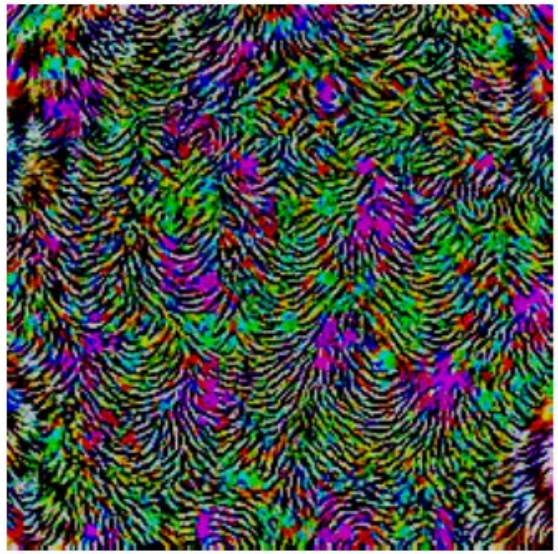


(c) VGG-16



(d) VGG-19

(e) GoogLeNet



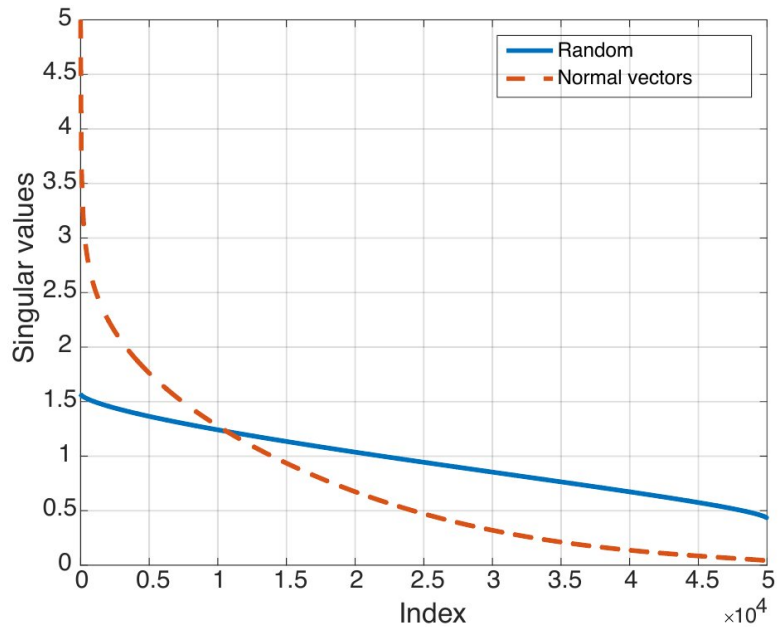
(f) ResNet-152

Universelle Störungen verschiedener neuronaler Netze

Mathematische Erklärung:

- Viele Instanzen fehlklassifiziert zu dominanten Labels mit großer Fläche im Bildraum; genauer:
- binärer Klassifikator hat 1 Normalvektor
- Matrix aus Normalvektoren aus Umgebung von n Instanzen zur Entscheidungsgrenze: $N = \left[\frac{r(x_1)}{||r(x_1)||_2} \cdots \frac{r(x_n)}{||r(x_n)||_2} \right]$

- Singularwerte von N nehmen stark ab
  $\Rightarrow$ große Korrelation & Redundanzen im Netz
  $\Rightarrow$ Unterraum U' mit $d' << d$ enthält meisten Normalvektoren
- Vektor aus U' erzielt 38% Fehlklassifizierungsrate (10% für zufälligen Vektor)
- Robustheit durch Lernen auf Störungsbildern: VGG-F 93,7% $\rightarrow$ 76,8%



[Moosavi-Dezfooli et al.(2016)]

1 Vertrauen in Entscheidungen

2 Gefahren

3 Erklärungen zu Entscheidungen
- lokale Approximation
- VQA - Visuell Fragen beantworten

4 Täuschung

5 Zusammenfassung

Nutzung intelligenter Algorithmen:

1. Verbrechensbekämpfung
2. Personenerkennung
3. militärische KI
4. gefährliche Aufgaben
5. Qualitätssicherung
6. Gewinnmaximierung
7. etc.

Gefahren:

1. Diskriminierung
2. Vortäuschung falscher Tatsachen
3. falsche Beratung bis hin zum Tod (Arzt)
4. Überwachung und Schrittverfolgung
5. Verurteilung im Gericht aufgrund von schlechten Datensätzen

Chancen durch besseres Verständnis:

1. Vertrauenssteigerung
2. Verbesserung der Algorithmen und Datensätze
3. Beratung für Ärzte
4. Automatisierung von Prozessen (da Vertrauen)
5. Problemlösung durch ungehinderten Fortschritt
6. Prozesse sind gerechter wg. größerer Datenbank und Vergleichen

Probleme im Verständnis:

1. nicht Open-Source
2. Ottonormalverbraucher versteht Code nicht
3. Multi-dim. mathematische Begründung
   $\Rightarrow\Leftarrow$ menschliches Denken

Verbesserung der Erklärung seitens der Politik:

- Verbraucherzentrale des Bundesverbandes: TÜV für Algorithmen
- Ab April 2018: General Data Protection Regulation (GDPR) EU-weit
  Pflicht zu nicht-diskriminierenden Algorithmen &
  Recht auf Erklärung:
  - dass und was an Daten gesammelt wird
  - wie eine Entscheidung zustande kommt
  - Möglichkeit, Prognosen korrigieren zu können

  Strafen von bis zu 4% des Umsatzes

Veständnis? - Nein, da nicht transparent.
Vertrauen? - Ja, aber nur bis zu einem gewissen Grad.

Verlust der Kontrolle ist gefährlich, da Entscheidungsfindungsprozess nicht bekannt ist.

Literatur

- Michael S Gazzaniga.
  The ethical brain.
  Dana press, 2005.
- Bryce Goodman and Seth Flaxman.
  European union regulations on algorithmic decision-making and a"right to explanation".
  arXiv preprint arXiv:1606.08813, 2016.
- Yash Goyal, Akrit Mohapatra, Devi Parikh, and Dhruv Batra.
  Towards transparent ai systems: Interpreting visual question answering models.
  arXiv preprint arXiv:1608.08974, 2016.
- Zachary C Lipton.
  The mythos of model interpretability.
  arXiv preprint arXiv:1606.03490, 2016.
- Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard.
  Universal adversarial perturbations.
  arXiv preprint arXiv:1610.08401, 2016.
- Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach.
  Attentive explanations: Justifying decisions and pointing to the evidence.
  arXiv preprint arXiv:1612.04757, 2016.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin.
  Why should i trust you?: Explaining the predictions of any classifier.
  In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.