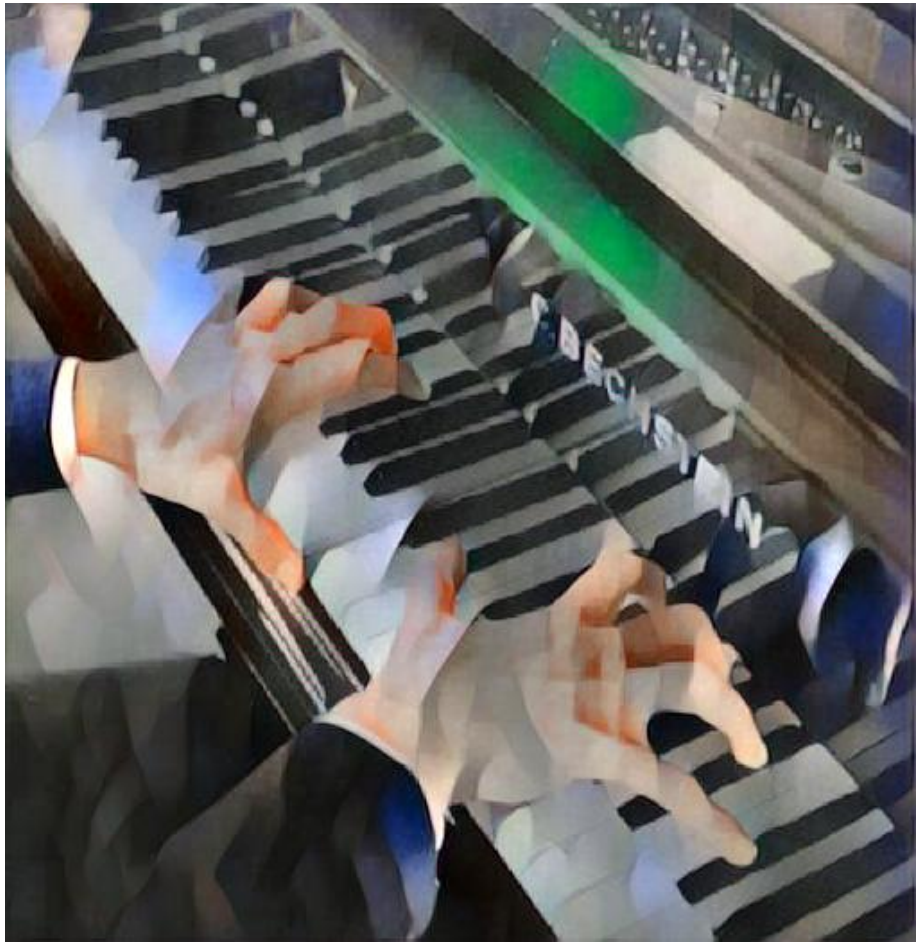
Der KI auf die Finger geschaut

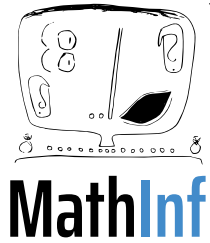
Ausgaben neuronaler Netze erklären



Thomas Viehmann, MathInf GmbH, tv@mathinf.eu

Berufenachmittag Actuarial Data Science, Göttingen, 7. Mai 2019

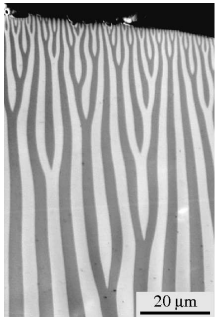
Über mich



Thomas Viehmann

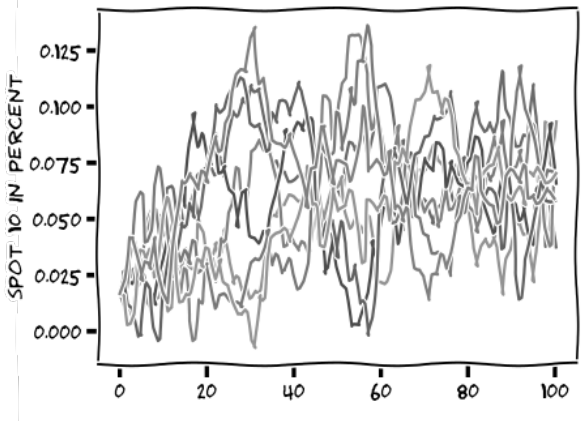
- Seit 2018: MathInf GmbH: Training, Coaching, Beratung, um Unternehmen beim Aufbau eigener KI zu unterstützen
- PyTorch Core Developer, mit >100 Features und Fehlerkorrekturen einer der weltweit führenden unabhängigen PyTorch-Entwickler (@tom bei PyTorch, @t-vi auf Github)
- aktuell u.a. Prototype Fund / BMBF-gefördertes Projekt LOTranslate für eine KI-Übersetzungsfunktion in LibreOffice
- ML blog: `https://lernapparat.de/`

Biographisches



- Mathematische Modellierung
- Promotion in Mathematik (Bonn) – Mathematische Untersuchung von Strukturen in Magneten

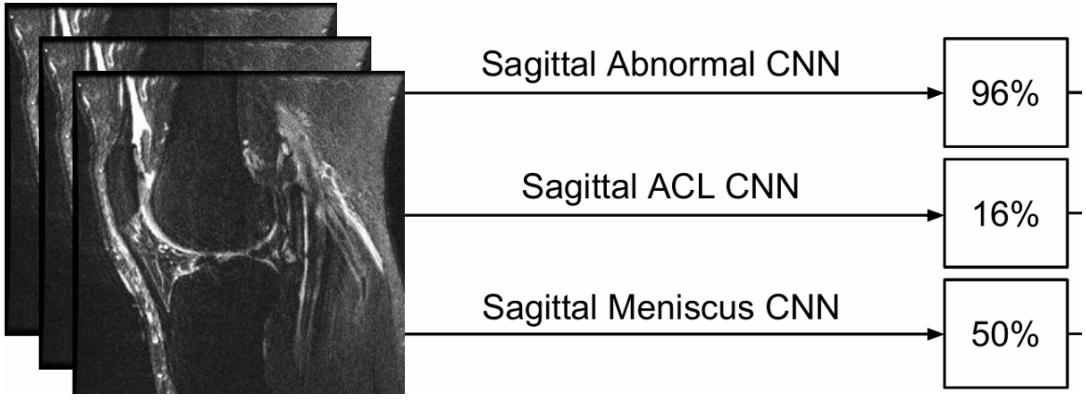
- 2009-2018 beratender Aktuar bei B&W Deloitte - vor allem statistische und ökonomische Modellierung (z.B. verschiedene Kapitalmarktszenario-Generatoren)



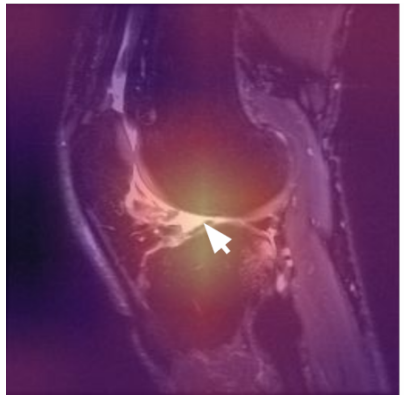
- DAV-Mitglied / Aktuar DAV seit 2011
- Verschiedene AGs (Data Science, Kreditrisiko, ...)
- Unterrichte im Certified Enterprise Risk Actuary-Programm
- altes Blog: https://interesting-rates.de/

Warum Ausgaben erklären?

- Business-Motivation: Black-Box für kritische Unternehmensentscheidungen ist bedenklich.
- KI als Unterstützung / zum Erkennen von Mustern: Prognose Radiologische Diagnose “Das ist ein Knie mit Meniskusriss” ist vermutlich viel weniger nützlich als “hier sieht es auf dem Bild wie ein Meniskusriss aus”.



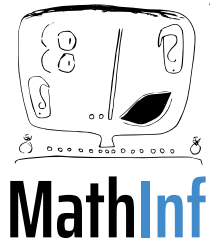
oder



(Quelle: Bien et al.: Deep-learning assisted diagnosis)

- Regulatorik: Kunden / Aufseher haben einen ggf. Anspruch darauf, zu wissen, warum eine Entscheidugn getroffen wird.

Unterstützung für die Schadenbearbeitung in der Krankenversicherung



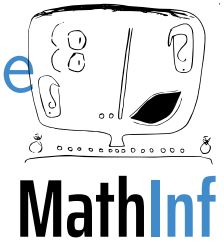
- Eingabe: Rechnung als Einzelpositionen, Tarifmerkmale, Informationen aus der Leistungs-Historie
- Ausgabe: Vorschlag (pro Position), z.B. volle Leistung / gekürzte Leistung / Ablehnung

Das ist noch nicht so nützlich...
...aber mit:

- folgendes Tarifmerkmal schließt die Leistung aus, oder
- diese beiden Positionen können nicht zusammen abgerechnet werden,

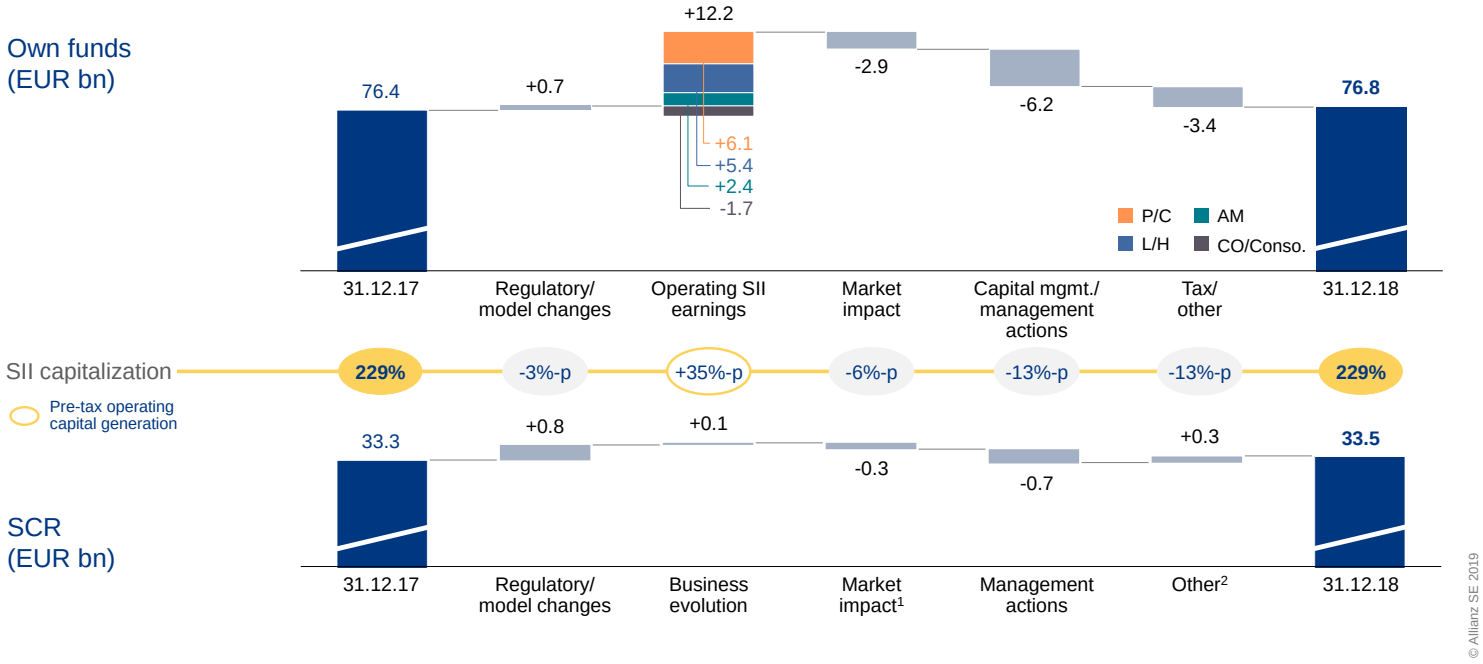
wird eher ein Schuh draus.

Ausgaben erklären für Aktuare – Veränderungsanalyse



B. Group financial results 2018

Group: excellent capital generation



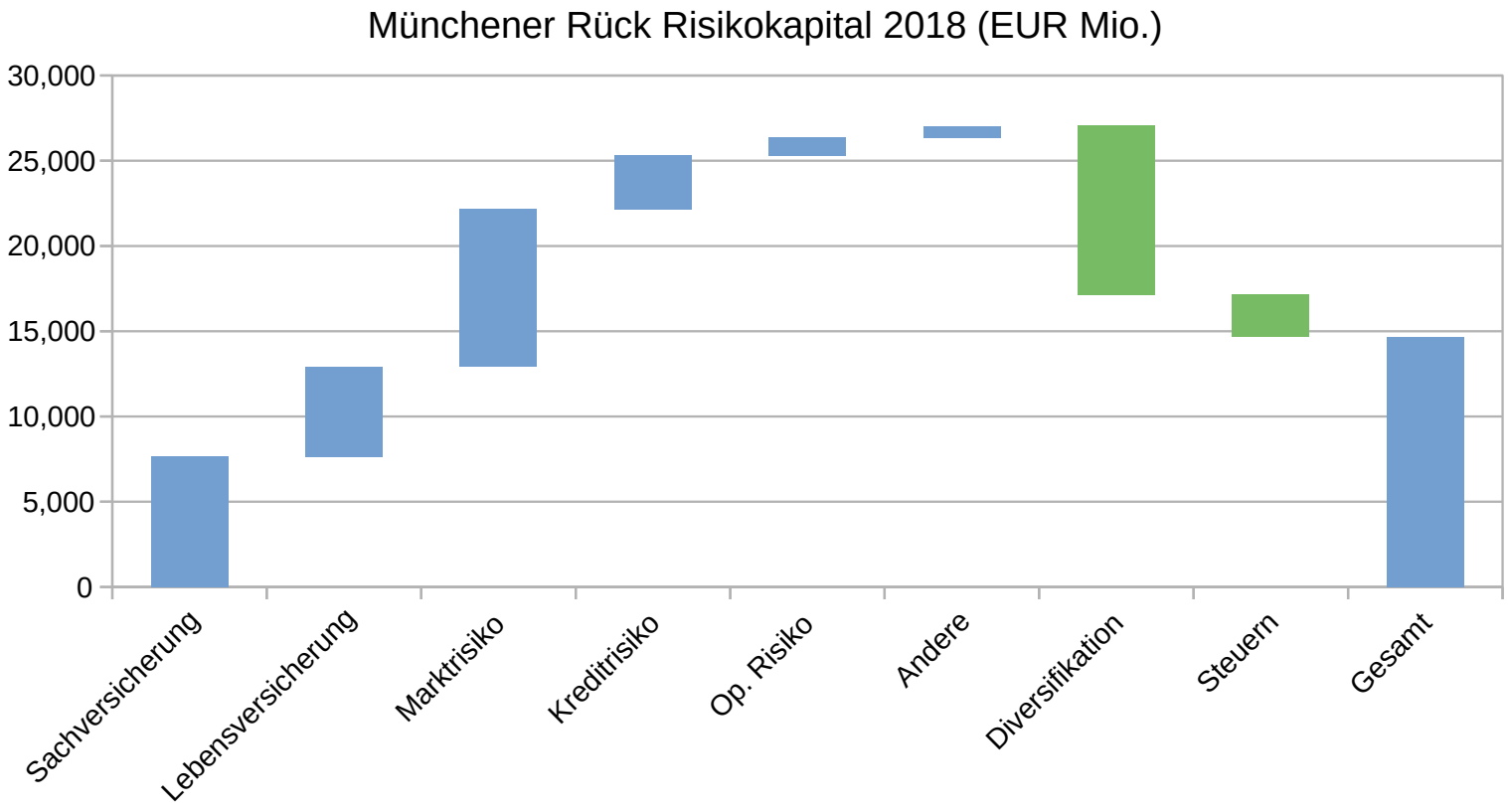
1) Including cross effects and policyholder participation
2) Other effects on SCR include diversification effects

B 9

Gegeben Eingaben (und Modell) zu Zeitpunkten t_1 und t_2, wie zerlege ich Differenz der Ausgaben nach Unterschieden der Eingaben?

Quelle: Allianz Analystenpräsentation zu den Ergebnissen 2018

Ausgaben erklären für Aktuare – Kapitalallokation

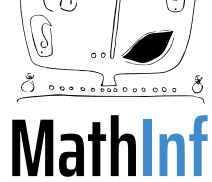


Gegeben Einzel-Risikokapital RK_i und gesamtes RK, wie allokiere ich Kapitalien AK_i so dass $\sum AK_i = RK$?

Mit anderen Worten: Wie verteile ich die Diversifikation?

Quelle: MR Jahresbericht 2018, eigene Grafik

Ausgaben erklären – allgemeine Problemstellung



Gegeben

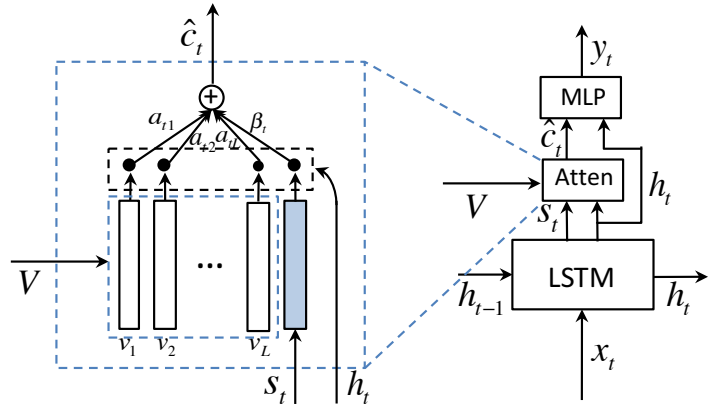
- Modellfunktion f,
- Eingabe X_i für $i = 1, ..., N$ (Pixel, Wörter, ...) und
- Ausgabe $f(X)$ (z.B. Wahrscheinlichkeitsverteilung über Kategorien).

Welche Eingaben bestimmen $f(X)$ wesentlich?

Wir werden zwar die Modellstruktur ausnutzen, versuchen aber letztlich nicht, den Berechnungsweg zu erklären.

Attention – wo sieht das Modell hin?

Manche KI-Modelle benutzen einen Attention-Mechanismus, bei dem Eingabemerkmale gewichtet in die Berechnung einfließen.
“Gewichte = Aufmerksamkeit”



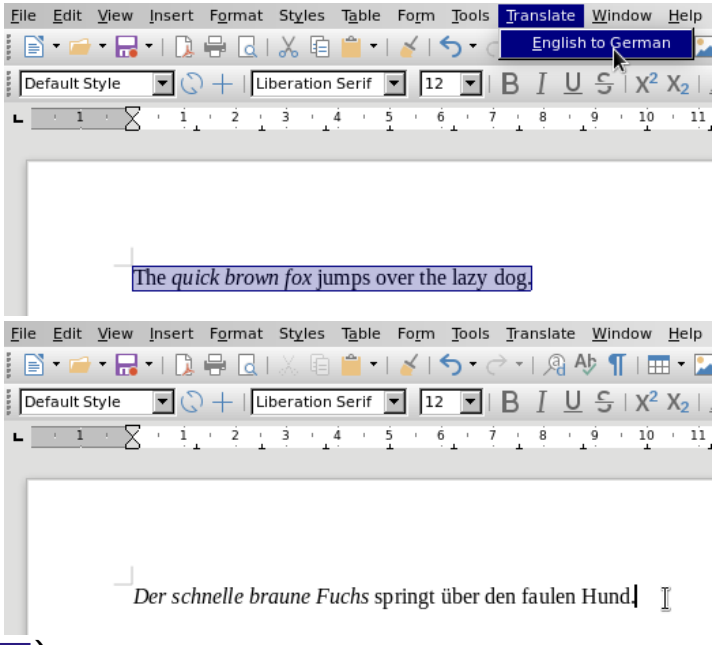
V = Information pro Pixelblock
moralisch: Skalarprodukte mit Suchinformation S

Quelle: J. Lu et al., Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning

Attention II – wo sieht das Modell hin?

...aber wenn das neuronale Netz Möglichkeiten hat, sich anderweitig etwas zu merken, sieht es schon auf die nächsten Eingaben.
Zum Beispiel rekurrentes neuronales Netz (RNN) zur Übersetzung

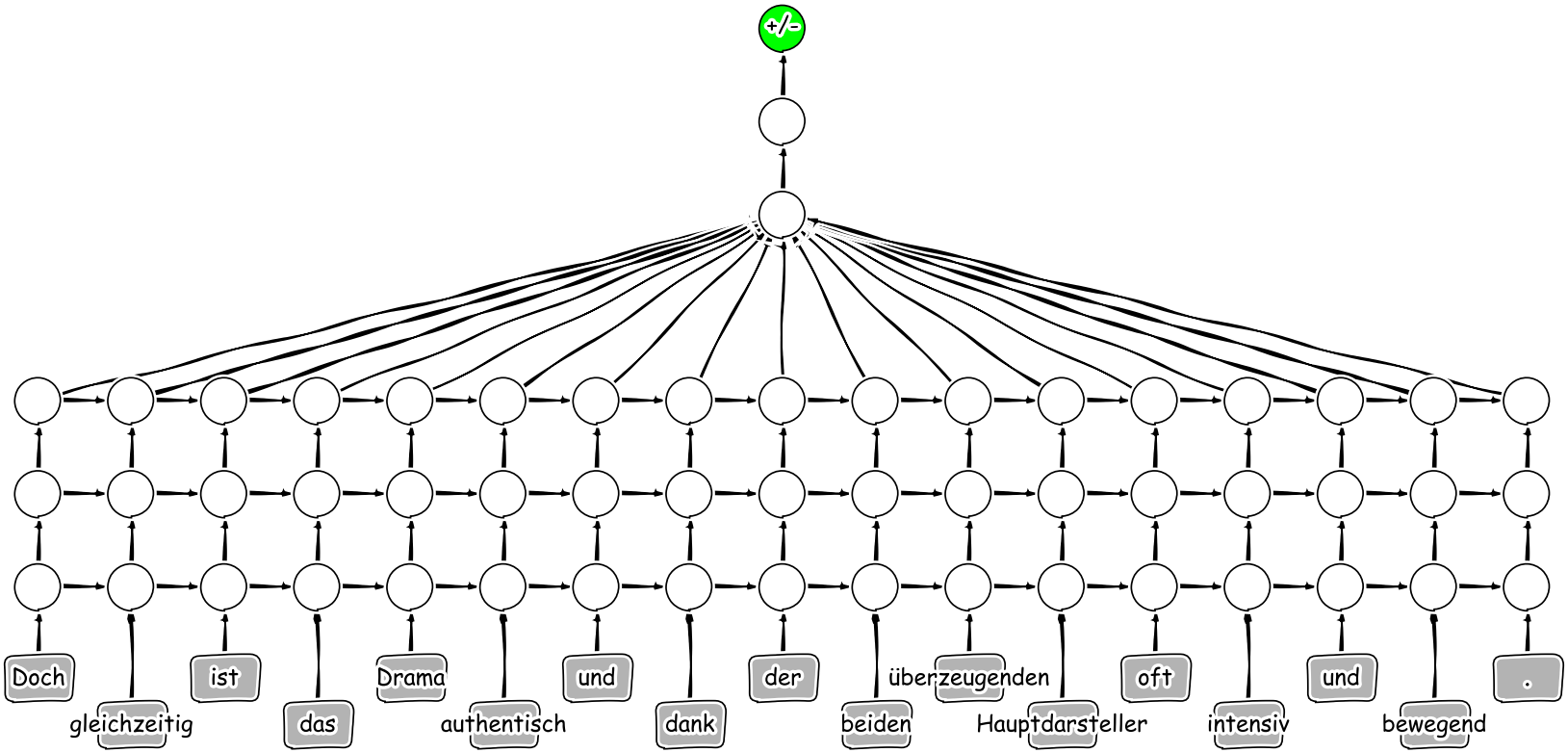
	_The	_quick	_brown	_fo	x	_jump	s	_over	_the	_la	zy	_dog	.
_Der	0.07	0.42	0.13	0.03	0.01	0.03	0.02	0.02	0.01	0.02	0.01	0.00	0.24
_schnelle	0.02	0.76	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.01
_bra	0.00	0.03	0.90	0.02	0.03	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01
une	0.00	0.00	0.02	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.94
_F	0.00	0.00	0.06	0.02	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07
uch	0.00	0.00	0.00	0.75	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.04
s	0.01	0.01	0.01	0.02	0.05	0.02	0.01	0.01	0.00	0.00	0.01	0.01	0.85
_spring	0.00	0.00	0.00	0.00	0.00	0.86	0.08	0.03	0.00	0.00	0.00	0.00	0.02
t	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.93
_über	0.00	0.03	0.02	0.00	0.00	0.04	0.03	0.03	0.04	0.11	0.02	0.01	0.67
_den	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.05	0.32	0.19	0.01	0.40
_fa	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.08	0.84	0.00	0.03
ul	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.06	0.73	0.00	0.20
en	0.00	0.01	0.03	0.01	0.00	0.06	0.01	0.03	0.01	0.04	0.03	0.06	0.71
_Hund	0.00	0.01	0.01	0.00	0.02	0.01	0.00	0.02	0.01	0.02	0.08	0.04	0.77
.	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.01	0.00	0.01	0.00	0.01	0.93
</s>	0.01	0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.00	0.95



(eigenes Projekt LOTranslate, basierend auf OpenNMT)

Modularität von neuronalen Netzen

Wie vieles bei NNs kann man auch Erklärbarkeit herunterbrechen.
Ein neuronales Netz zur Textklassifikation erkennt eine positive Filmkritik, aber woran?

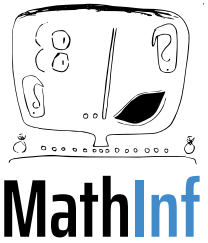


Kreise = Module = "elementare" Funktionen

Pfeile = Informationsfluss = Verkettung von Funktionen

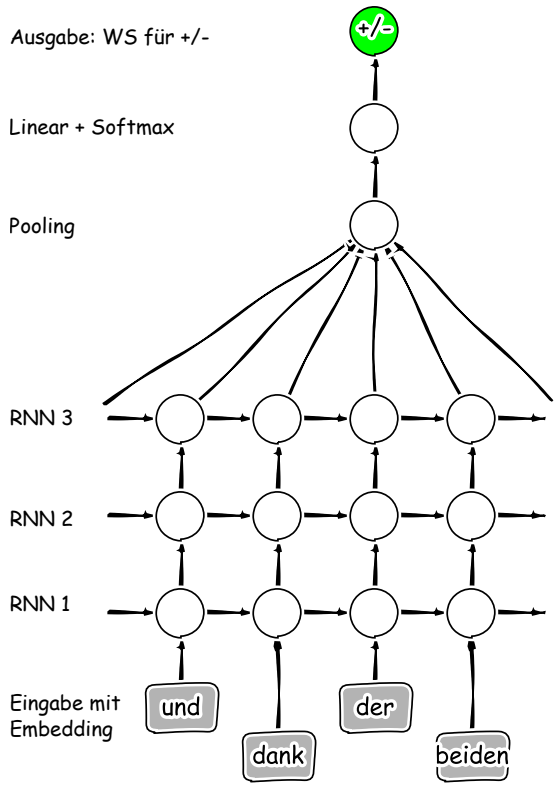
(Modell Howard/Ruder: ULMFiT, eigenes Training für deutsche Sprache)

Ein kurzer Blick auf die Details



Schichten:

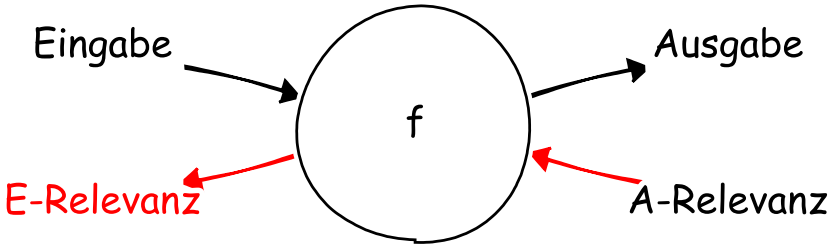
- Embedding – ordnet jedem Wort des Lexikons einen Vektor zu
- RNN (rekurrent): berechnet in jedem Kreis einer horizontalen Ebene jeweils die selbe Funktion. Eine der Eingaben ist das Ergebnis des letzten "Zeit-"Schritts – so "merkt" das NN sich Dinge.
- Pooling – Macht aus den vielen Ausgaben eine (Maximum, Mittelwert, letzter Wert)
- (affin) Linear wie erwartet. Softmax: Abbildung $\mathbb{R}^n$ auf Verteilung zwischen n Alternativen.
- zwischendurch nichtlineare elementweise Funktionen (Aktivierungen)



ULMFiT Trick: Trainiere Embedding + RNN auf Wikipedia o.ä. das jeweils nächste Wort vorherzusagen, erst dann auf das eigene Problem.

Erklären von Ausgaben - Propagation-Methoden

Relevanz für ein Modul:



Erste Idee:

Ableiten: Wo muss ich wackeln, um die Ergebnisse zu beeinflussen?

Also: E-Relevanz $R_{in} = \nabla f \cdot R_{out}$ (Simonyan et al: Deep inside Convolutional Networks)

- für Aktuare: wie Euler-Allokation ohne Homogenität
- "Dunkle Seite": adversarial Attacks

f zum Beispiel

- (affin) linear: $f(x) = Ax + b$,
- elementweise nichtlinear $f(x) = tanh(x)$.

CERA Ausbildung Modul D
13./14. März 2019 in München

DAA DEUTSCHE AKTUAR-AKADEMIE GmbH

Performancemessung: Allokation von Kapital

Euler-Allokation (kontinuierliches Marginalprinzip)

- Jedes Risiko X erhält einen Beitrag proportional zum Risikobeitrag bei Ausbau der Position:

$$EC(X_j) = \frac{d}{d\lambda}\Big|_{\lambda=0} E(X + \lambda X_j)$$

- Für positiv homogene Risikomaße: Summe der Einzelkapitalien ist das Gesamtkapital (Satz von Euler für homogene Funktionen)
- kohärente Risikomaße → **die** Allokation, die die Denault-Axiome erfüllt
- Für Varianz, TVaR → Varianz-Covarianz-Prinzip, TvaR-Prinzip
- Ableitung in der Praxis schwierig zu handhaben
  - z.B. Euler Allokation für VaR ist Erwartungswert über niederdimensionalen Unterraum im Risikoraum

80

Erklären von Ausgaben - Propagation-Methoden II

Zweite Idee:

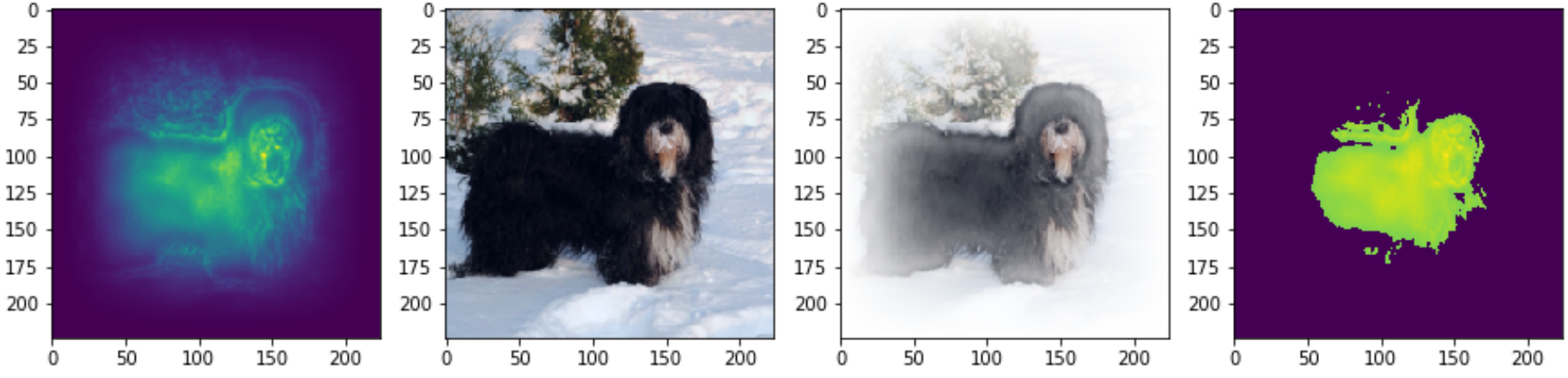
1. Ordnung-Taylor Entwicklung: $f(X) - f(0) \approx \nabla f(X) \cdot X$

$\rightarrow$ Bach et. al: Layerwise Relevance Propagation

Montavon et al.: Deep Taylor Decomposition

Varianten:

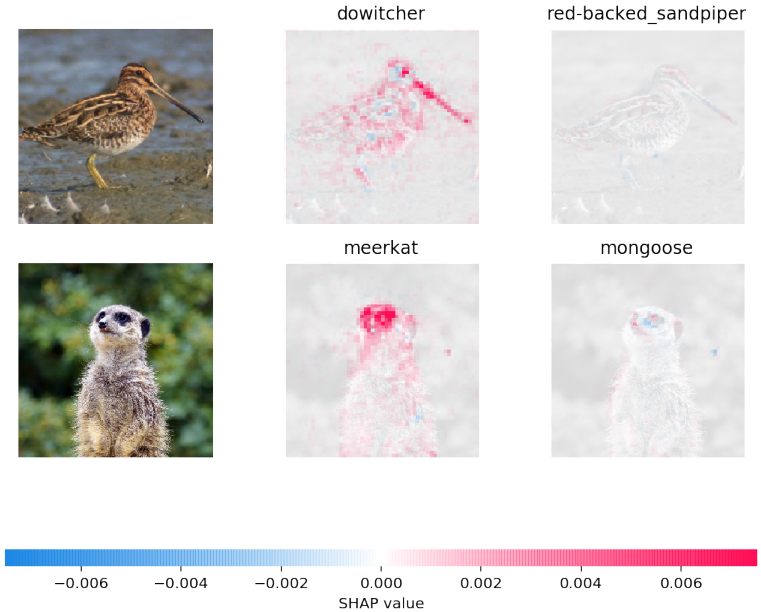
- geeignete Normierung der Relevanz
- ad-hoc Umgehung der Unstetigkeit für ResNets (Eigenentwicklung)



LRP für ResNet34 (eigene Implementierung, torchvision ResNet)

Erklären von Ausgaben - Propagation-Methoden III

DeepLIFT (Shrikumar et al.) ändert die Referenz z.B. auf die mittlere Eingabe jedes Moduls und baut - ähnlich Backpropagation ein (wie oben normiertes) Relevanzmaß darauf auf.
DeepSHAP (Lundberg/Lee) schließlich liefert eine abstraktere Begründung für DeepLIFT als lineare Approximation eine Shapley-Allokation von Relevanz. (+ Verbesserungen im Detail aus dieser Anschauung)



CERA Ausbildung Modul D
13./14. März 2019 in München

DAA
DEUTSCHE AKTUAR-AKADEMIE GmbH

Performancemessung: Allokation von Kapital

Shapley-Algorithmus

- Idee: Wie beim Marginalprinzip wird der Beitrag zum Gesamtrisiko berechnet, den jede Einheit liefert, wenn diese zu dem Gesamtunternehmen ohne diese Einheit hinzugefügt wird. Im Gegensatz zum Marginalpinzip werden aber alle Permutationen berücksichtigt.
- Berechnung:
  Sei B eine Teilmenge mit m Elementen aus allen betrachteten Segmenten, $m < n$
  - Sei $EC(B)$ das Risikokapital dieser Teilmenge B
  - Risikokapital-Beitrag eines Segments i, welches nicht zu B gehört, zur Teilmenge B:

$$\Delta_i(B) = EC(B \cup \{i\}) - EC(B)$$

  - Berechne diesen Beitrag über alle $n!$ möglichen Teilmengen und bestimme so für jedes Segment i das allokierte Risikokapital

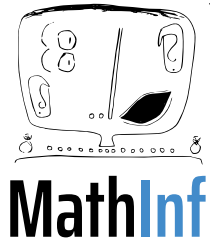
$$EC_i^A = \sum_{B \subset \{1,\ldots,i-1,i+1,\ldots,n\}} \frac{\#B!(n-1-\#B)!}{n!} \cdot \Delta_i(B)$$

Vorteil	Nachteil
Erfüllt immer risikolose Allokation und Symmetrie	Verstößt bei VaR gegen Exzessverbot
Einheiten mit hohem Beitrag zum Diversifikationseffekt, erhalten geringeres Kapital allokiert.	Komplexe Berechnung auf Grund Vielzahl von Kombinationen
Theoretisch gerechte Verteilung	Deutlich aufwändiger als alle anderen Verfahren
	Kollektives Exzessverbot nur bei additiven Risikomaßen erfüllt

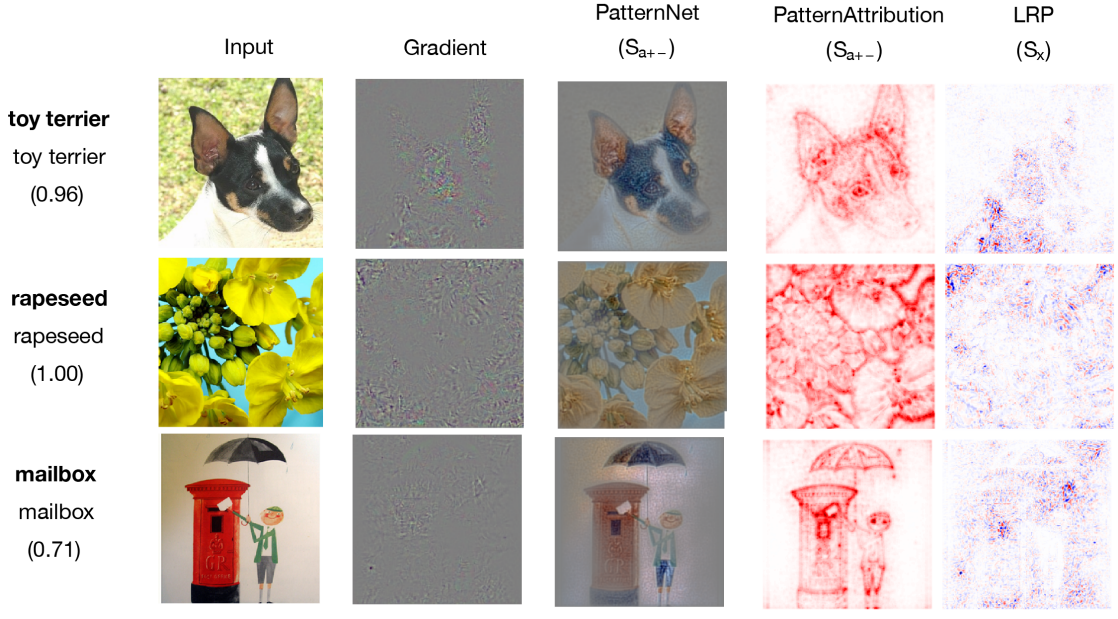
87

(Quelle r.: Code Lundberg et al, A Unified Approach to Interpreting Model Predictions)

Weitergehende Analysen - über den Einzelfall hinaus

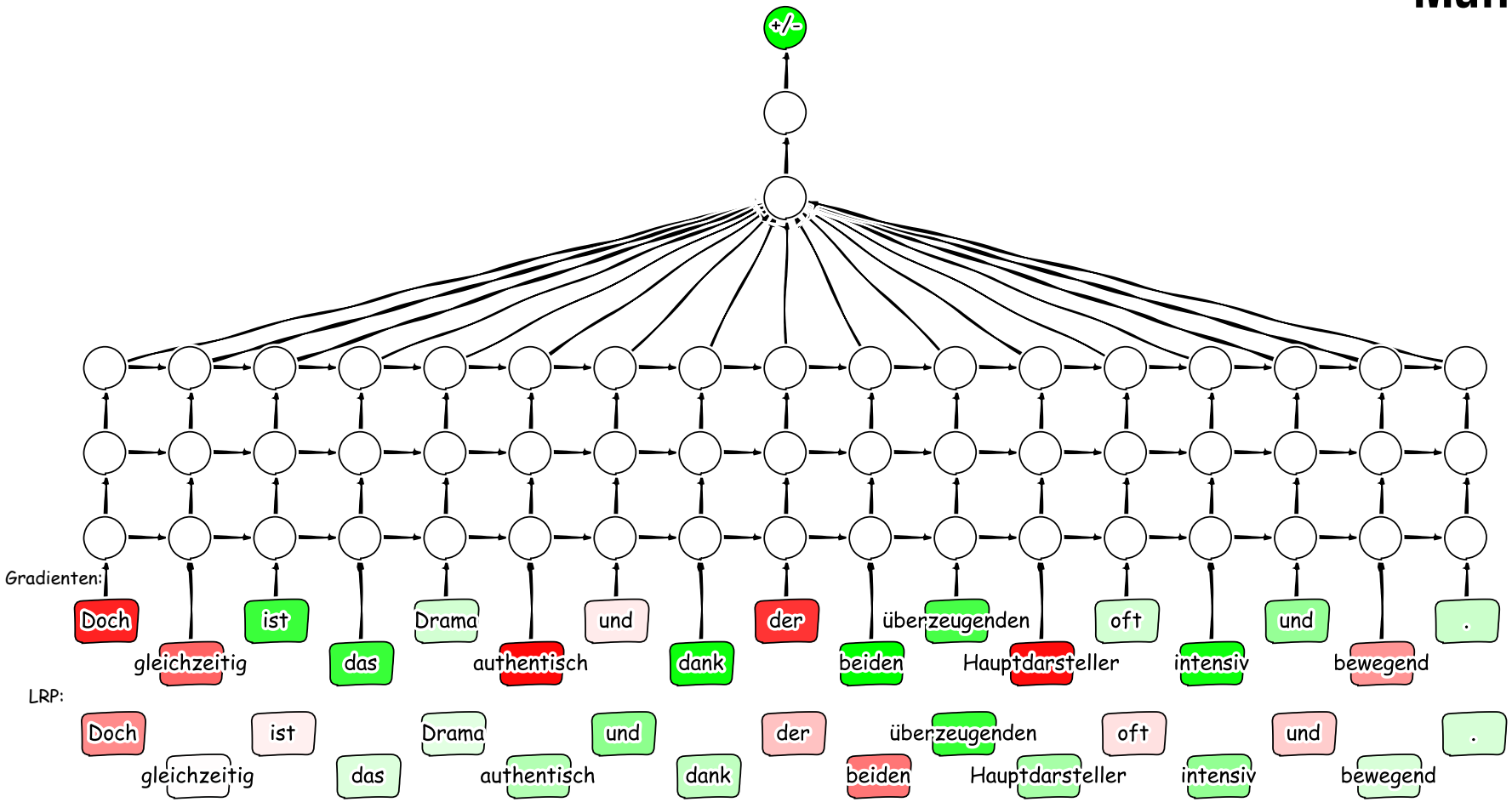


- PatternNet / PatternAttribution (Kindermans et al) – Zerlege fallweise Relevanz in Signal und Rauschen – hängt mit der Wahl des Referenzpunktes zusammen



- SpRAy (Lapuschkin et al) – Clustering von Erklärungen für einen Datensatz, um Auffälligkeiten / Ausreißer zu markieren

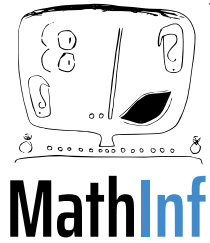
Für unsere Textklassifikation



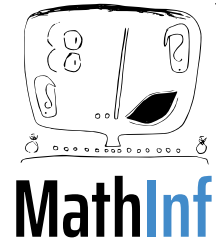
LRP deutlich plausibler als Gradienten.

Eigene Implementierung für ULMFiT nach Arras et al: Explaining Recurrent Neural Network Predictions in Sentiment Analysis

Fazit



- Oft sind Ausgaben nützlicher oder überhaupt erst nützlich, wenn man weiß “woran es liegt”.
- Die Hürde, Erklärungen im Sinne von “an welchen Eingaben” zu haben, ist niedriger als ein umfassendes Verständnis des Modells.
- Wir haben einige Methoden zur Erklärung angerissen:
  - Attention-Mechanismen sind manchmal zur Erklärung geeignet, aber nicht immer zuverlässig.
  - Für Modelle ohne Attention bieten sich Verfahren an, die an den Backpropagation-Algorithmus angelehnt sind.
  - Techniken sind allgemein, hier vorgeführt für Bilder und Text.
  - Wir haben Parallelen zur Kapitalallokation gesehen.



Vielen Dank!

Ihre Fragen und Anmerkungen

Kontakt: Thomas Viehmann, MathInf GmbH, tv@mathinf.eu

Schrift aus einem anderen Vortrag

Titelfolie: Hände von Dinu Lipatti + DeOldify + PyTorch Style Transfer