



Attention und Transformer Netzwerke

I. Attention Konzept 2. Transformer Architektur

Attention-Konzept: beim Gucken konzentriert man sich auch auf gewisse Objekte mehr
↪ Attention

Palm findet die nicht so nice, aber sind groß sehr aktuell / im Hype

Attention

Tokenization

„Sinfonie in Wörter und Interpretation unterteilen“

best-selling music artists

best - selling music artists

albums sold 500,000 copies

albums sold 500 , 000 copies

technically perfect, melodically correct

technically perfect , melodically ##ally correct

featuring a previously unheard track

featuring a previously un##heard track

Attention

Embedding

↳ sind vortrainiert

(≈ Vektoren)

↳ wörterbuch mit Embeddings

best	-	selling	music	artists
------	---	---------	-------	---------

$$\begin{pmatrix} 0.87 \\ -0.11 \\ \vdots \\ 0.13 \end{pmatrix} \begin{pmatrix} 0.34 \\ 0.12 \\ \vdots \\ -0.45 \end{pmatrix} \begin{pmatrix} -0.61 \\ 0.54 \\ \vdots \\ -0.32 \end{pmatrix} \begin{pmatrix} 0.42 \\ 0.14 \\ \vdots \\ 0.77 \end{pmatrix} \begin{pmatrix} 0.26 \\ -0.46 \\ \vdots \\ 0.27 \end{pmatrix}$$

mörd mit Vektoren weitergearbeitet

↳ man kann damit weiter.

↳ Embedding enthält Informationen über die die sich
Bedeutung des Tokens

© Peltarion

Prof. Dr. Christoph Palm
Regensburg Medical Image Computing (ReMIC)
Ostbayerische Technische Hochschule Regensburg (OTH Regensburg)

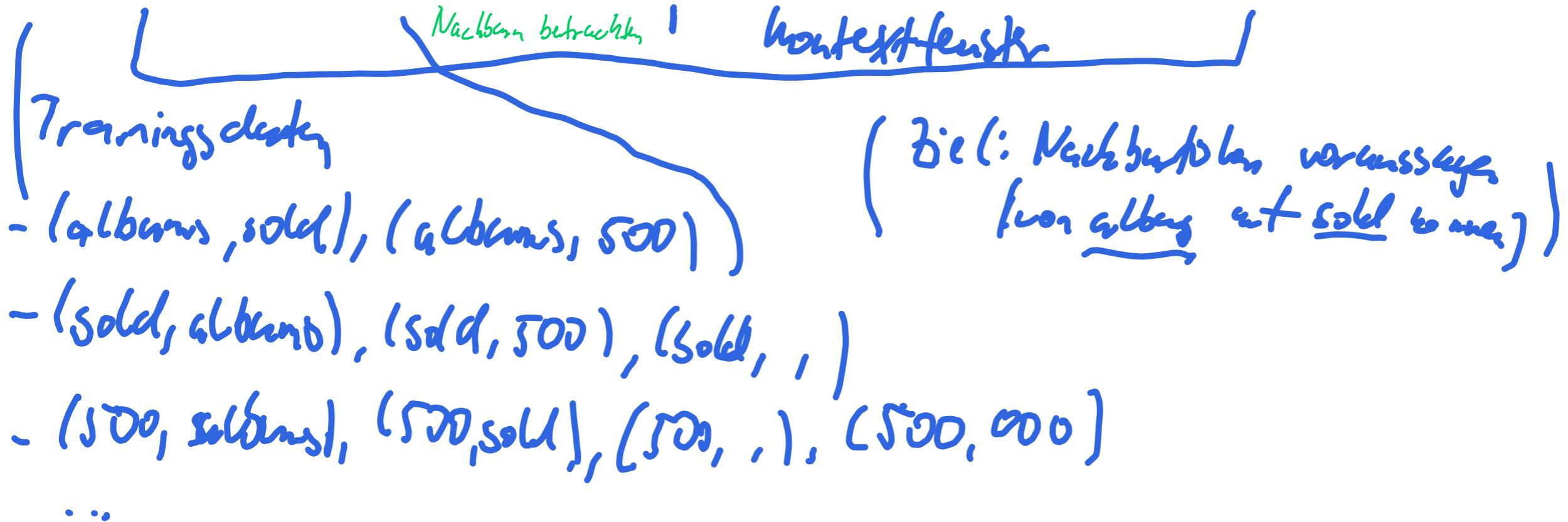
Attention

Word2Vec

mit kommt von der Embedding¹:

hilft man

albums	sold	500	,	000	copies
--------	------	-----	---	-----	--------

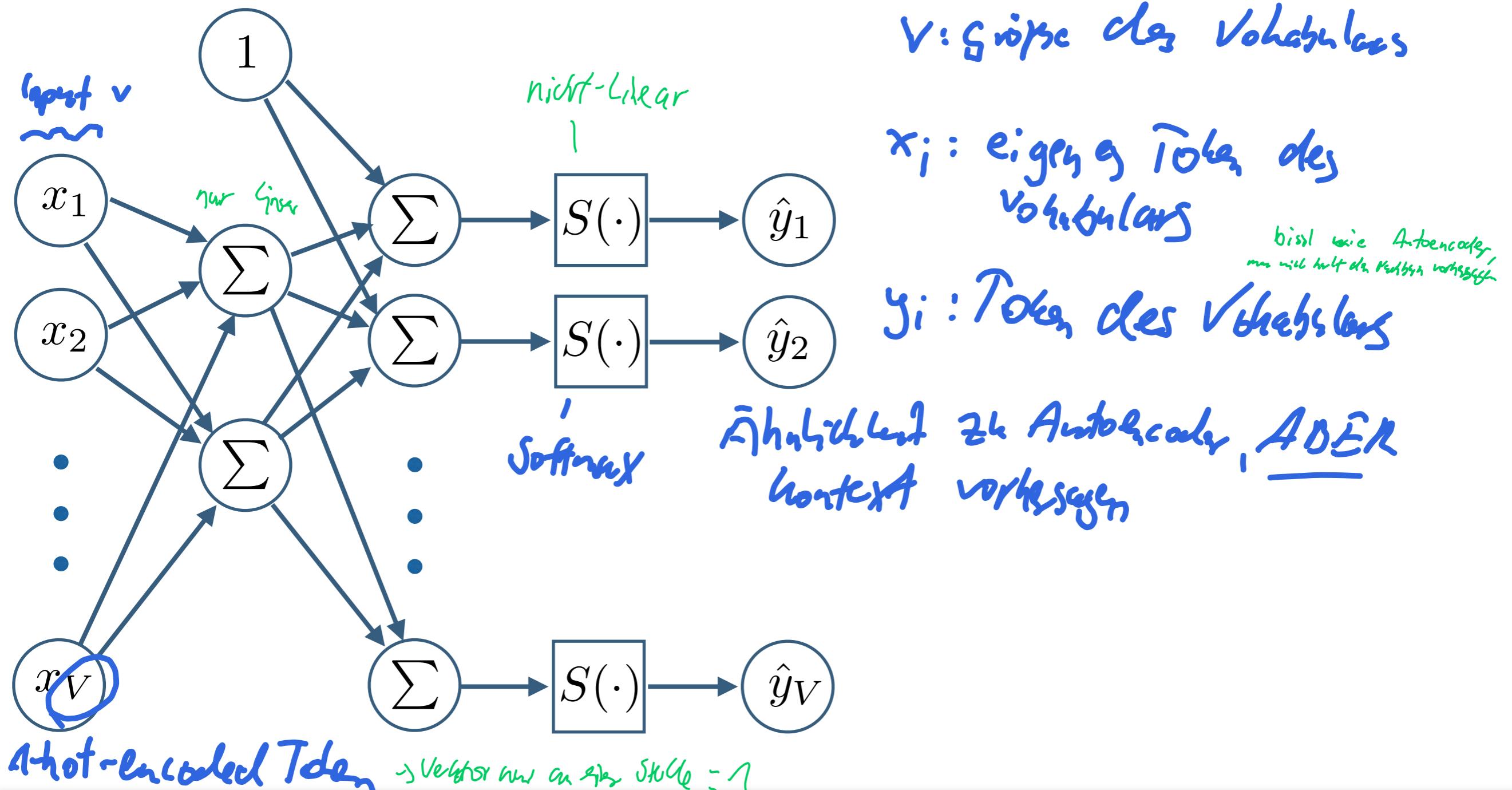


Idee: aus einem Token ein benachbartes Token vorhersagen

Attention

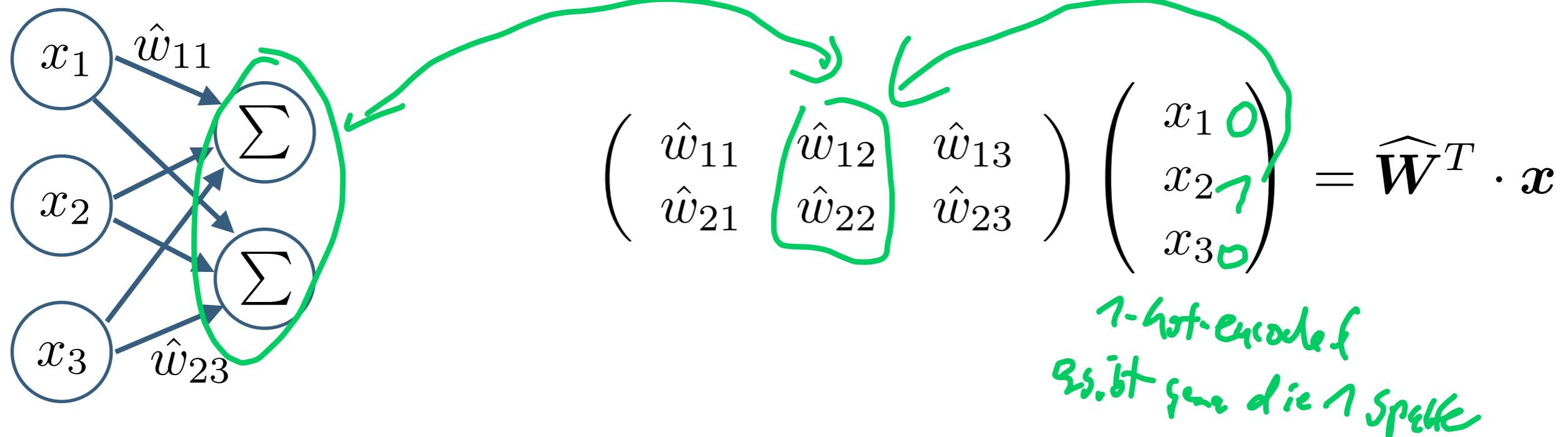
Word2Vec

albums sold 500 , 000 copies

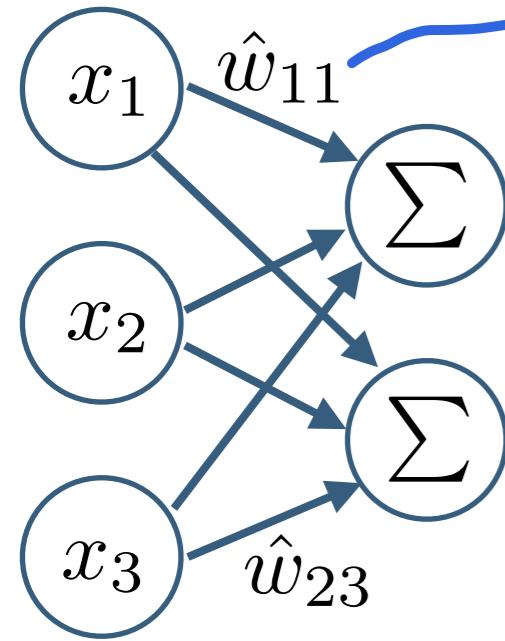


Wiederholung

Zeile \times Spalte



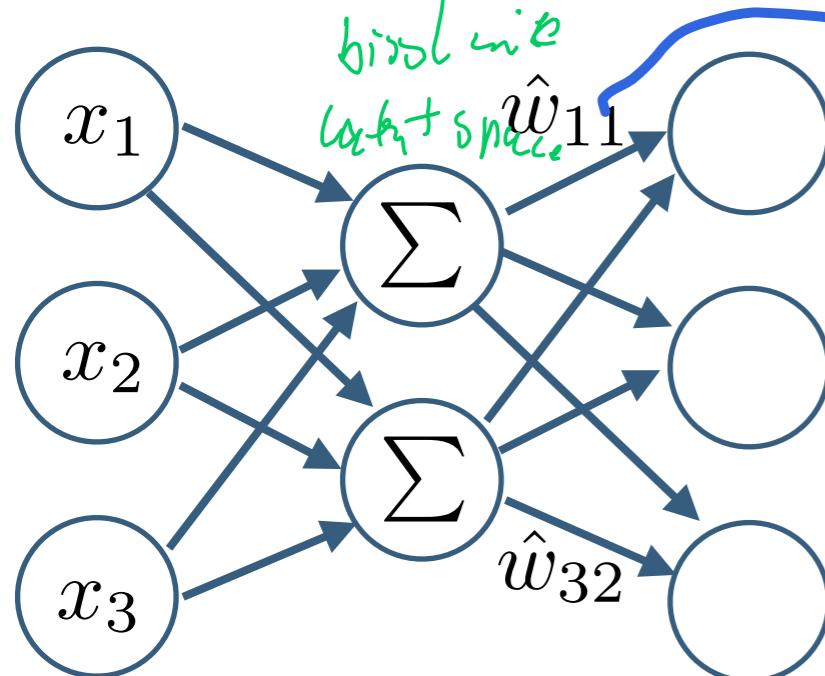
Wiederholung



Gewichtsmatrix 1. Schicht

$$\begin{pmatrix} \hat{w}_{11} & \hat{w}_{12} & \hat{w}_{13} \\ \hat{w}_{21} & \hat{w}_{22} & \hat{w}_{23} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \hat{\mathbf{W}}^T \cdot \mathbf{x}$$

not the same



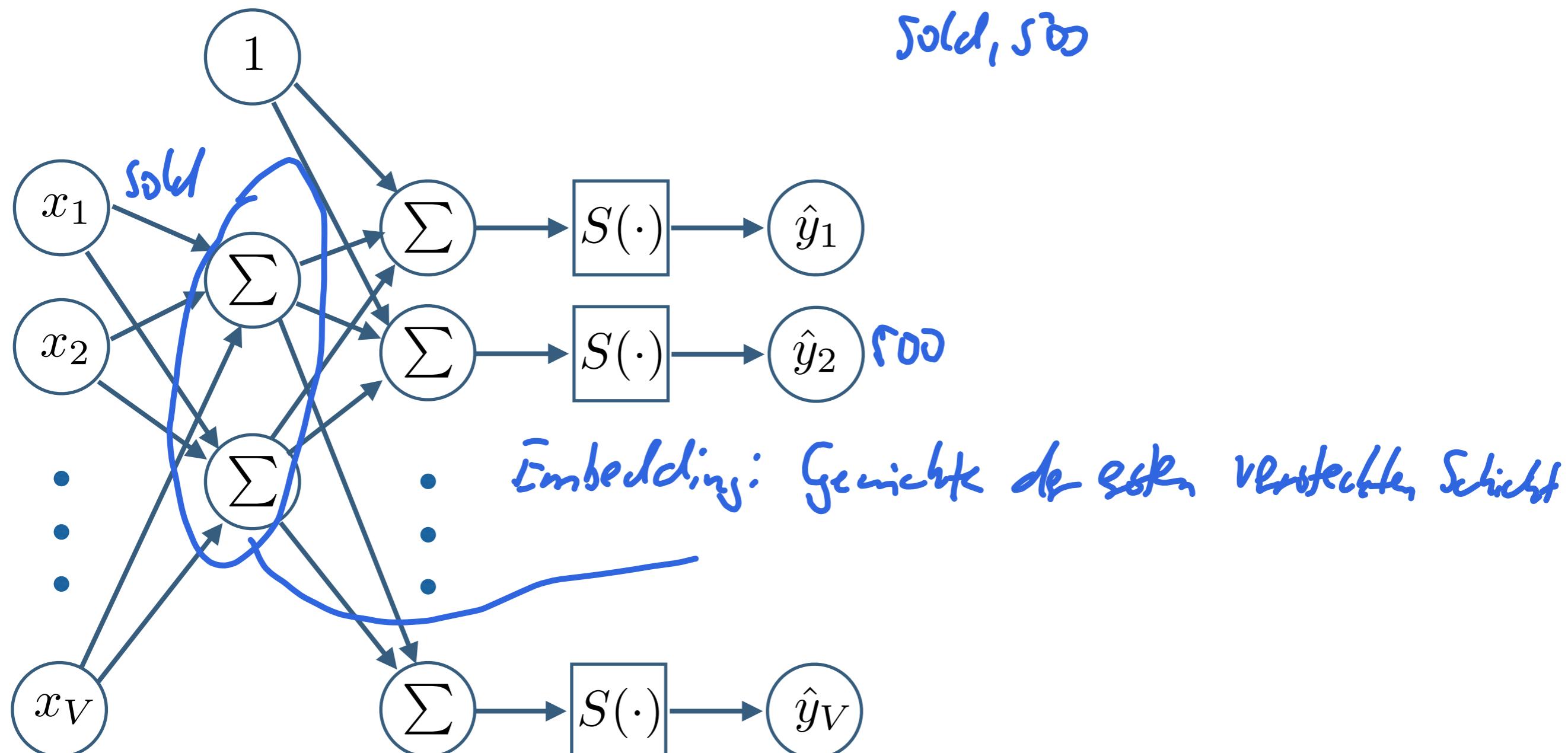
Gewichtsmatrix
2. Schicht

$$\begin{pmatrix} \hat{w}_{11} & \hat{w}_{12} \\ \hat{w}_{21} & \hat{w}_{22} \\ \hat{w}_{31} & \hat{w}_{32} \end{pmatrix} \begin{pmatrix} \sum_1 \\ \sum_2 \end{pmatrix}$$

Attention

Word2Vec

albums sold 500 , 000 copies



Attention

Embedding

$$\boxed{\text{king}} + \boxed{\text{woman}} - \boxed{\text{man}} = \boxed{\text{queen}}$$

Beispiele nicht übersehen, ob geht los, mit Bedeutungen zu rechnen

$$\boxed{\text{singing}} + \boxed{\text{yesterday}} - \boxed{\text{today}} = \boxed{\text{sang}}$$

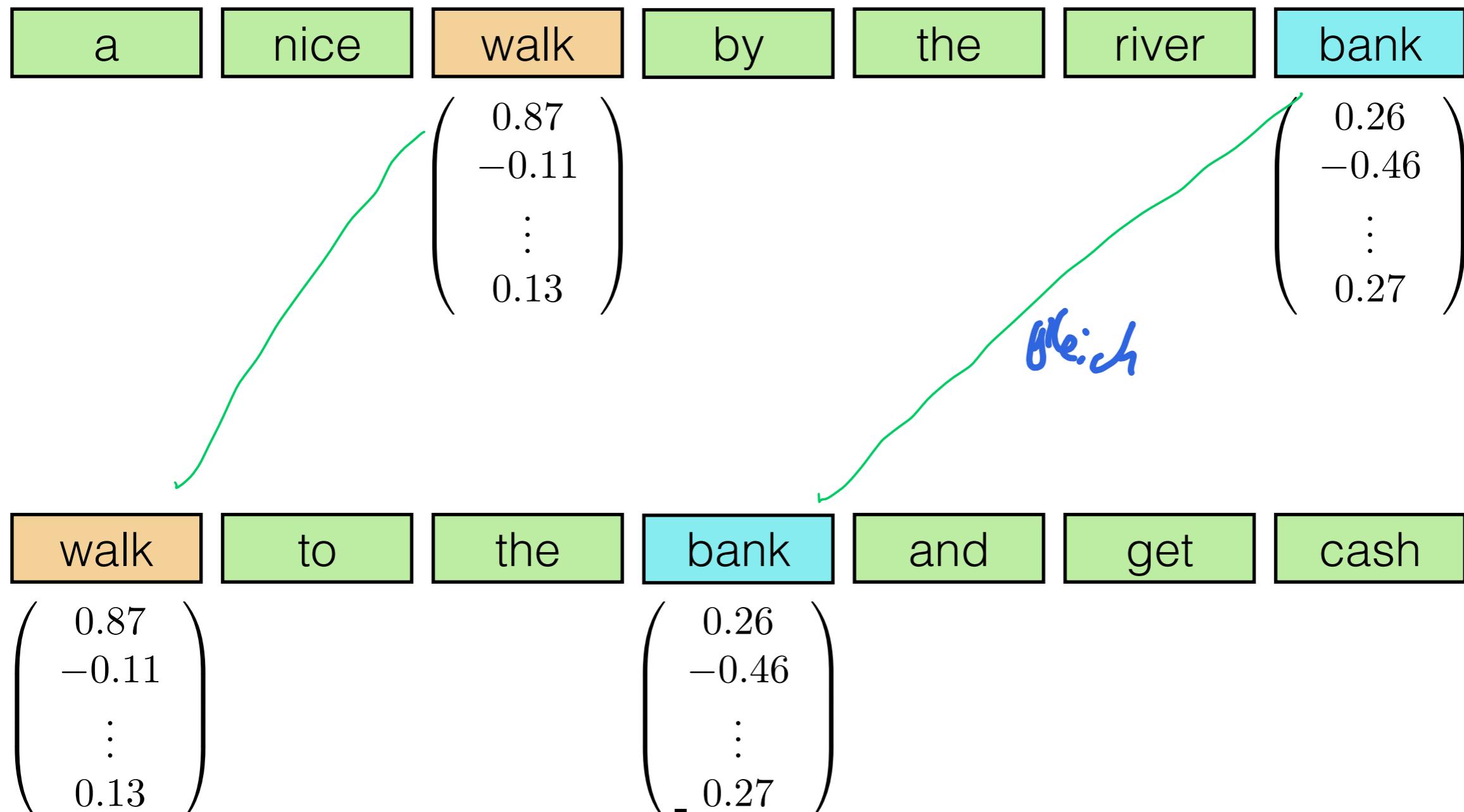
$$\boxed{\text{italy}} + \boxed{\text{paris}} - \boxed{\text{rome}} = \boxed{\text{france}}$$

$$\begin{pmatrix} 0.87 \\ -0.11 \\ \vdots \\ 0.13 \end{pmatrix} + \begin{pmatrix} -0.61 \\ 0.54 \\ \vdots \\ -0.32 \end{pmatrix} - \begin{pmatrix} 0.42 \\ 0.14 \\ \vdots \\ 0.77 \end{pmatrix} = \begin{pmatrix} 0.26 \\ -0.46 \\ \vdots \\ 0.27 \end{pmatrix}$$

Attention

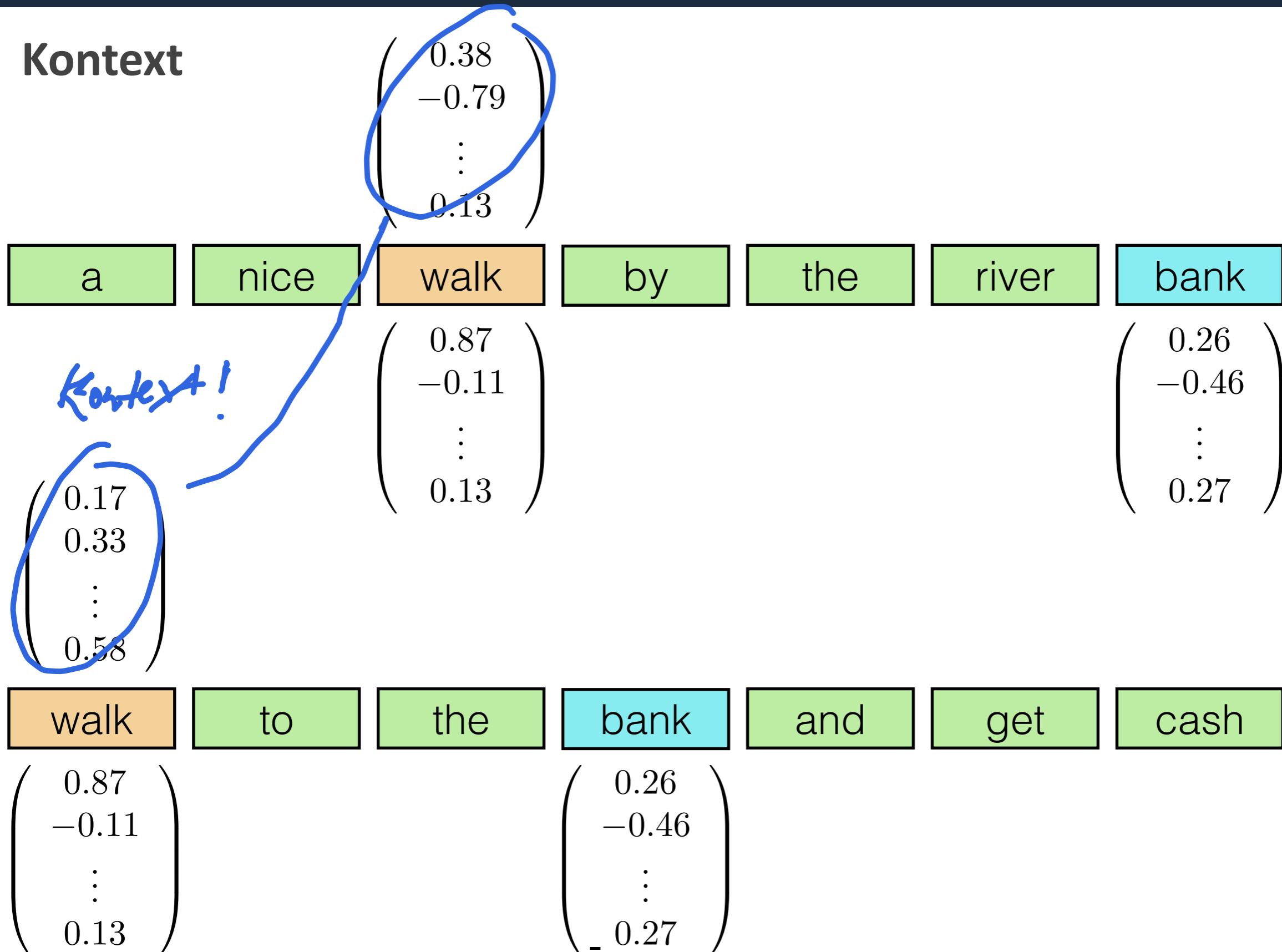
Kontext *mit einbezogen = Attention*

↪ Embeddings sind kontextunabhängig



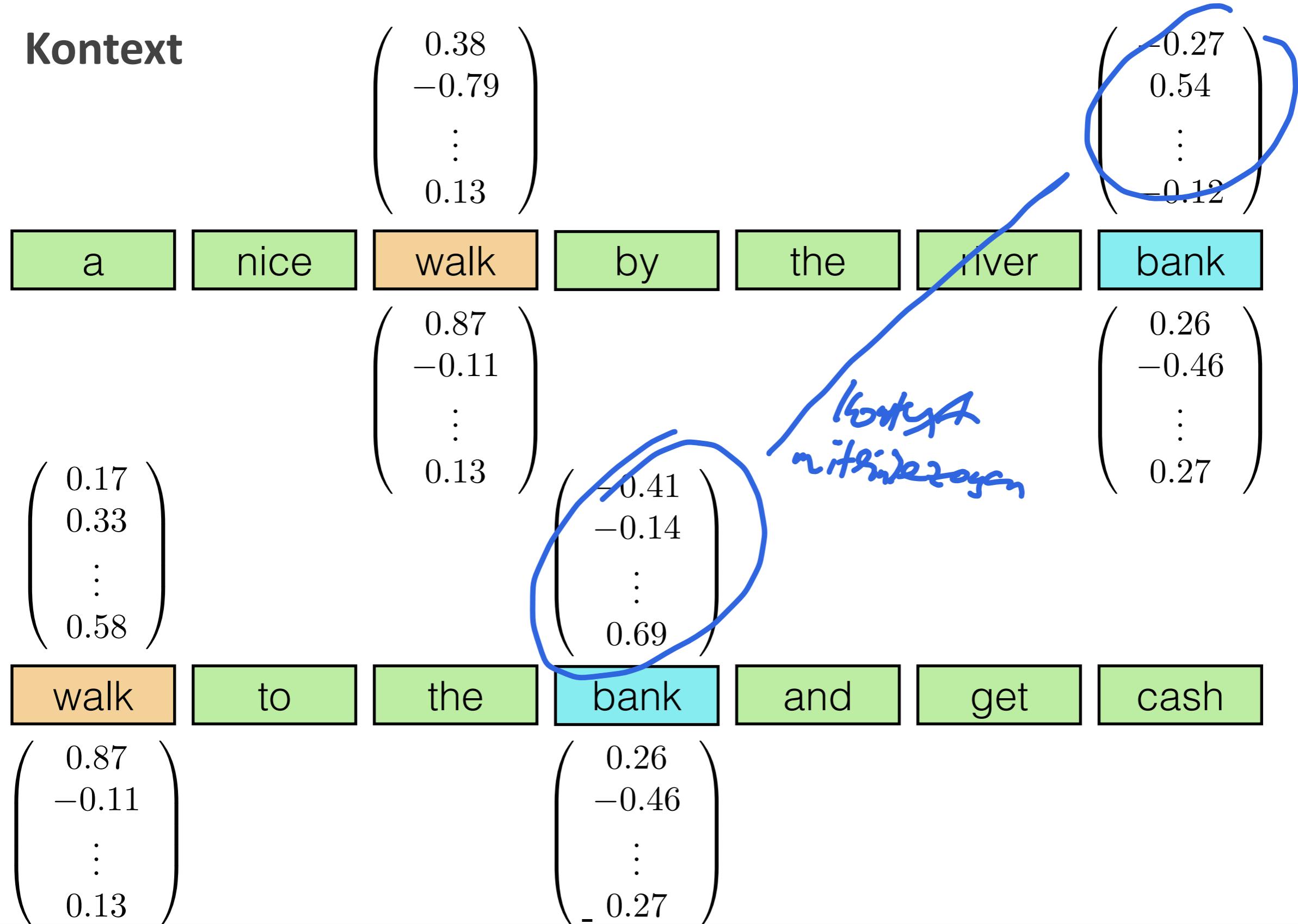
Attention

Kontext



Attention

Kontext



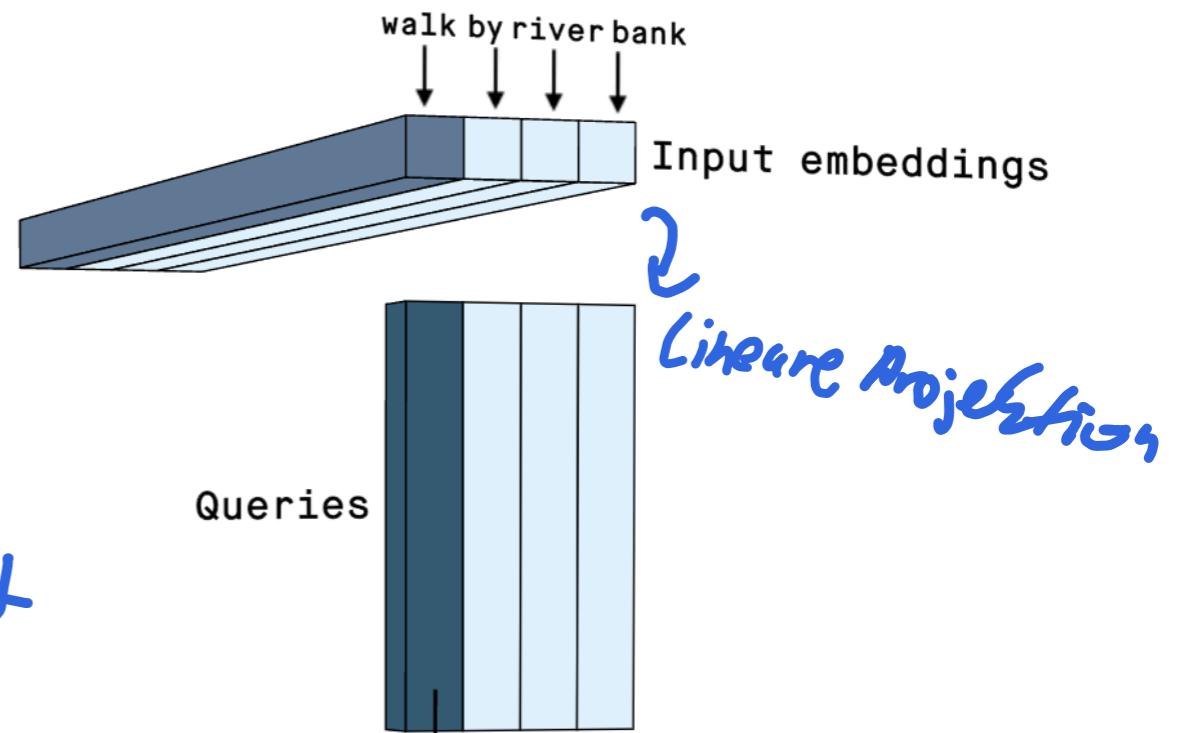
Attention

$Q = \text{Queries}$

Q : Lineare Projektion des Input Embeddings auf eine Query mit niedrigerer Dimension

geschaffenes
Modell

BERT: von 768 auf 64 Queries

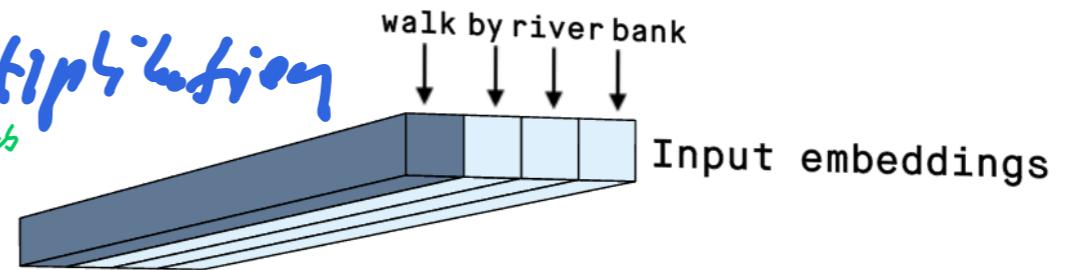


Attention

$Q \cdot K^T \rightarrow$ Ähnlichkeitsmaß, das Vektoren multipliziert

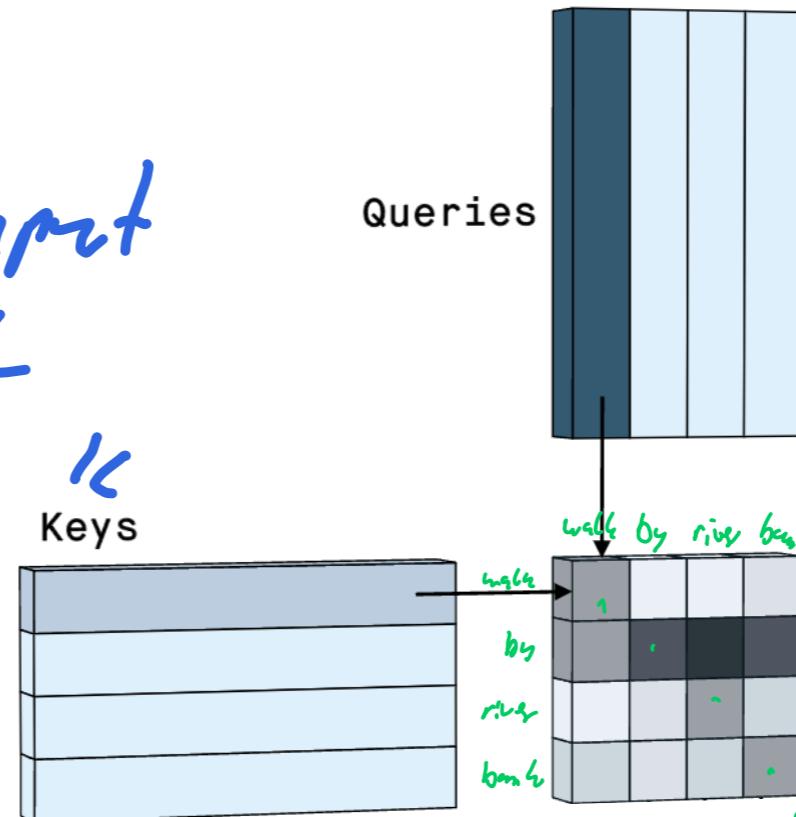
$$\frac{QK^T}{\sqrt{d_k}}$$

Scalarprodukt zwischen 2 Vektoren:
 - falls hoch hoch \rightarrow sehr ähnlich
 - niedrig \rightarrow sehr unähnlich



K : Lineare Projektion des Input Embeddings (again) auf einen Key

L) Scalarprodukt: groß bei ähnlichen Vektoren



- Warum ist PCA schlechter als t-SNE?
 - o Bei PCA wird Klasseinformation nicht verarbeitet
 - o Kriterium „Variance = Information“ nicht unbedingt immer korrekt (z.B. 4 vs 4 vs 4)

- SNE vs t-SNE
 - o Verteilung im niedrigdim. Raum \rightarrow Cauchy vs Gaußverteilung

- kein Cheat-Sheet, nur Taschenrechner

- ~ fragt eigt! keine Formeln ob (Formel für Loss-Fkt?)
eher sowas wie „das ist die Loss-Fkt. erklärt mir die mit“
 - ~ Motivation, Anwendung, Komponenten

- Verständnisfragen

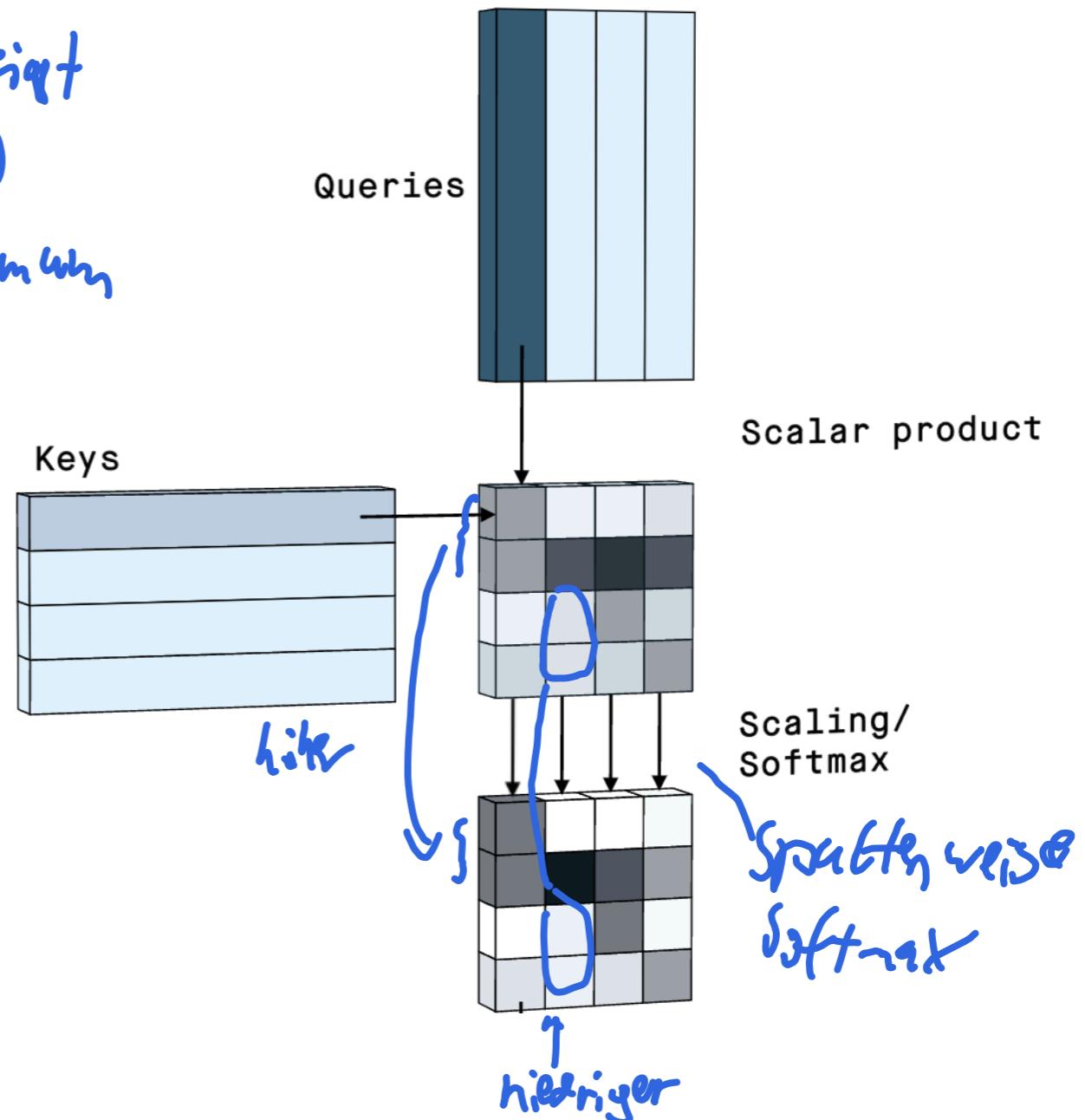
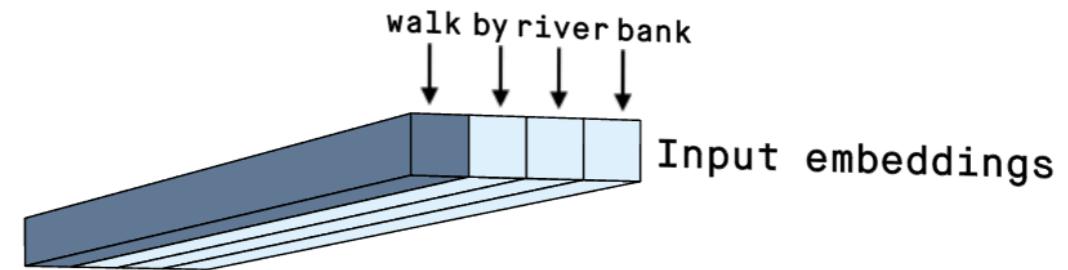
- keine Übersichtsblätter

Attention

$$S \left(\frac{QK^T}{\sqrt{d_k}} \right)$$

Problem: Variante des Skalarprodukts steigt mit steigender Dimension (mehr Vektoren)
↳ Softmax wird immer mehr zu Maximum

Lösung: Normierung auf $\sqrt{d_k}$



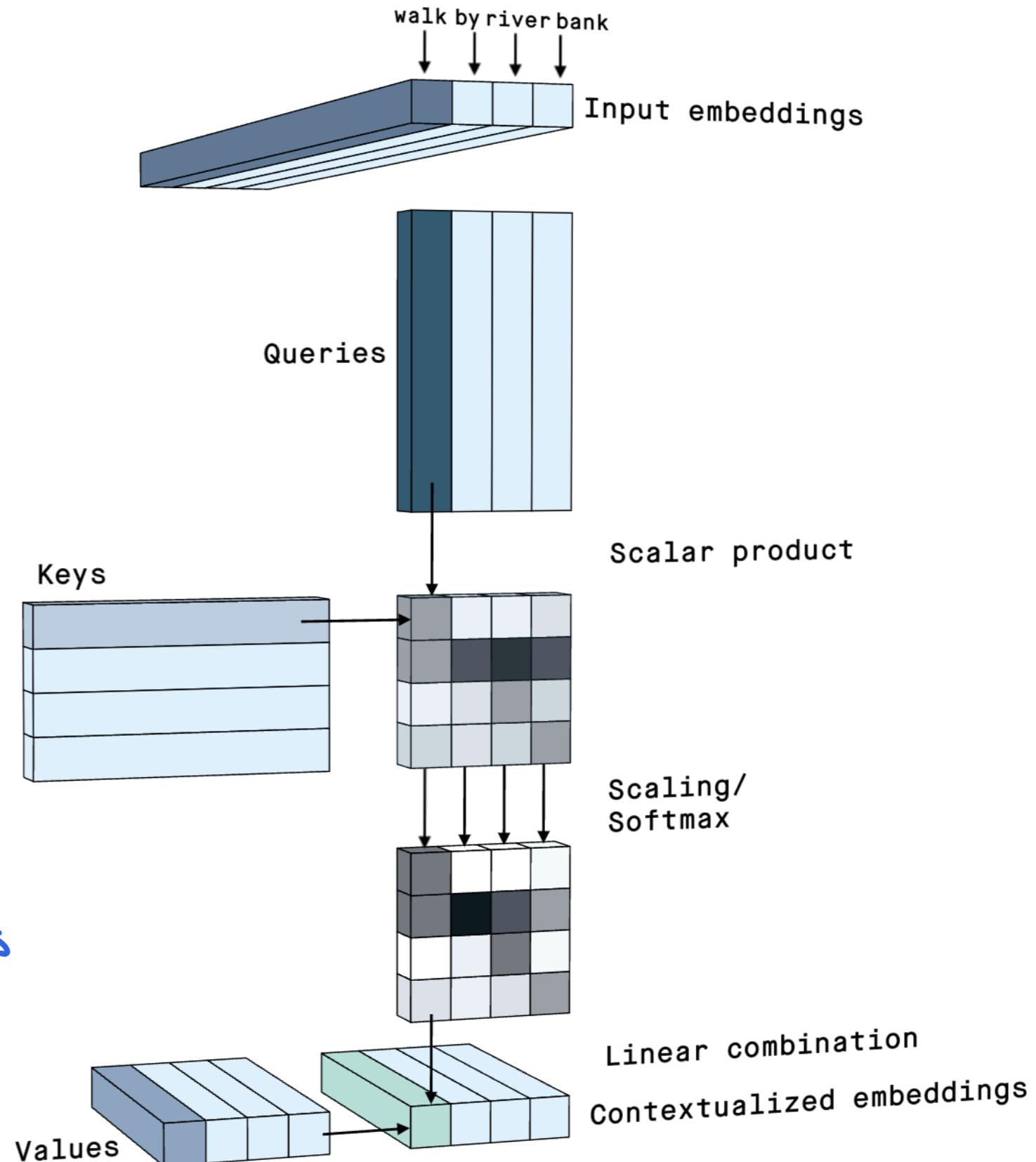
Attention

$$A(Q, K, V) = S \left(\frac{QK^T}{\sqrt{d_k}} \right) V$$

Kontextgewichtung

V: Values

- ↳ im Prinzip Input Embeddings
- ↳ werden durch Attention-mechanismus kontextualisiert

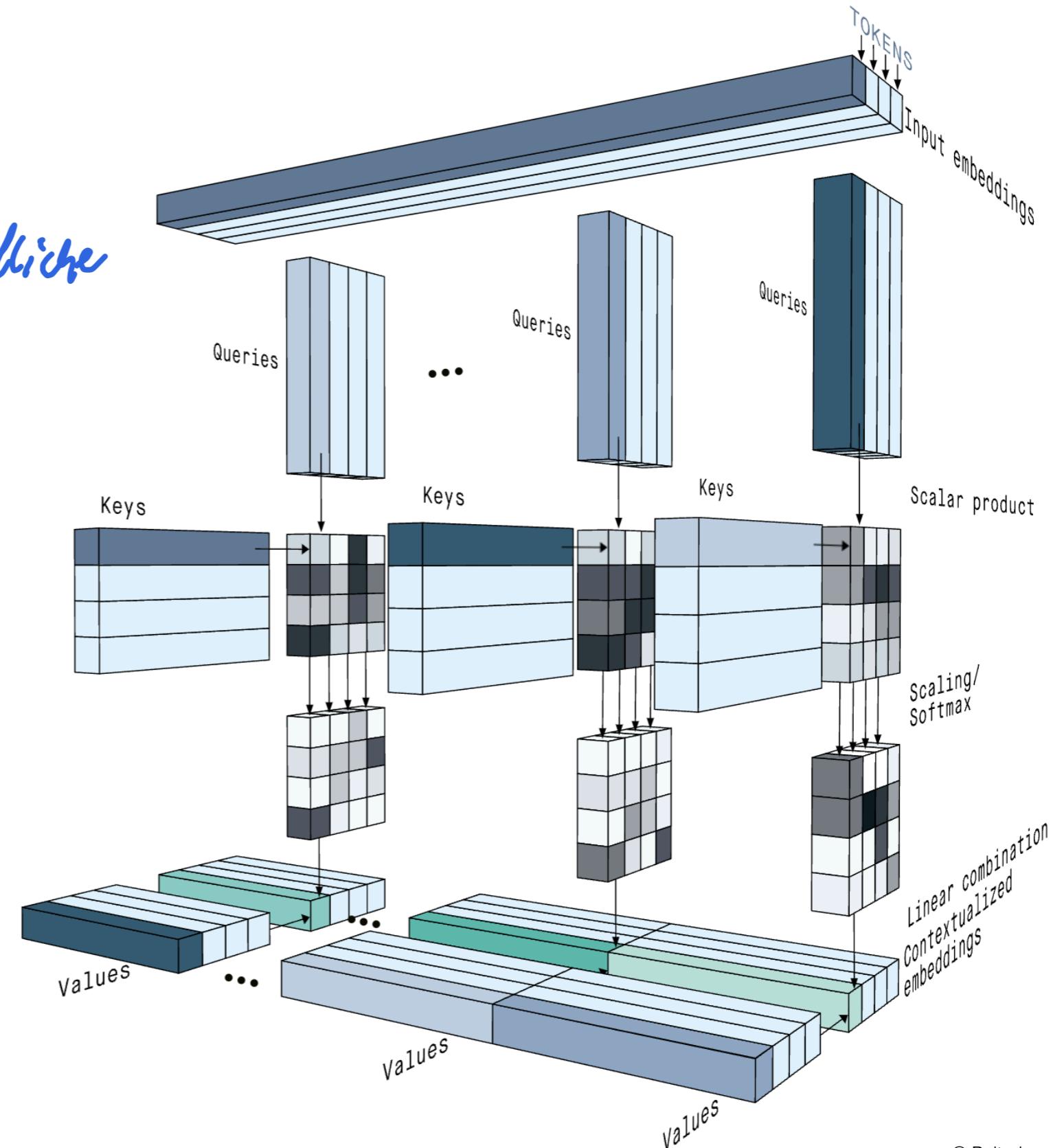


Attention

Multi-head Attention

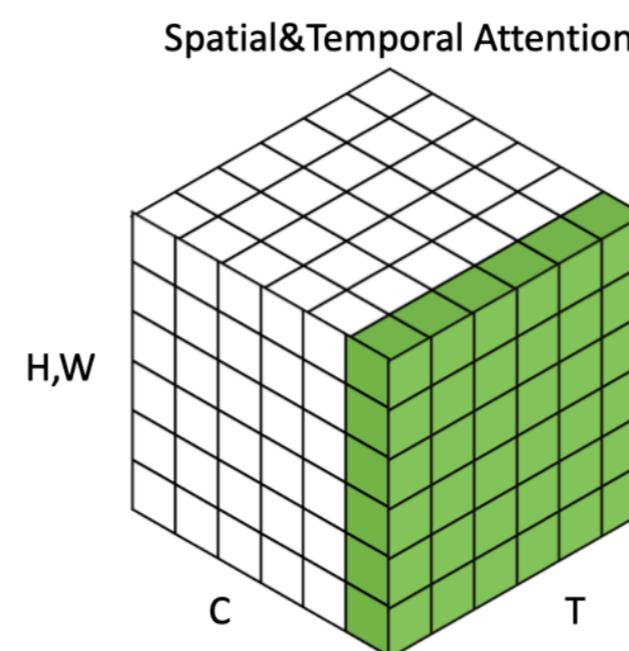
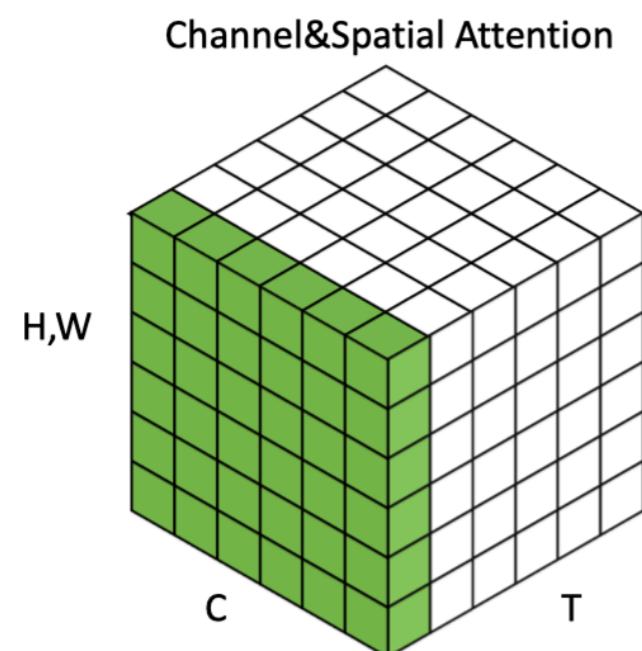
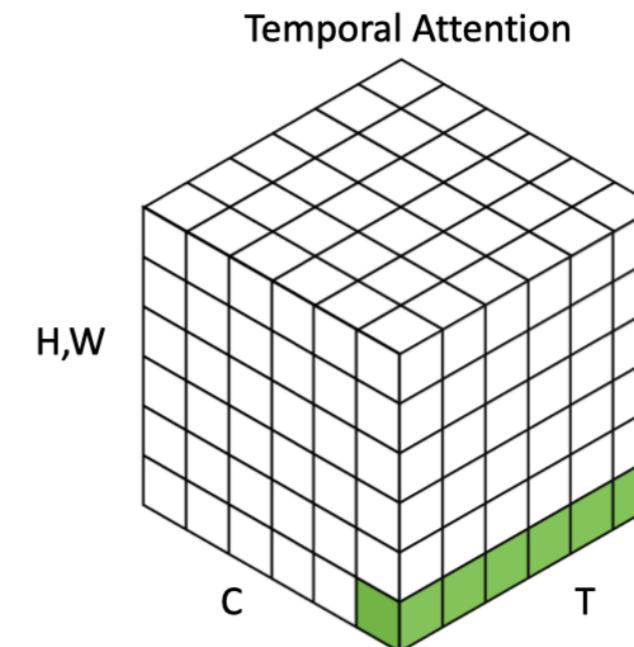
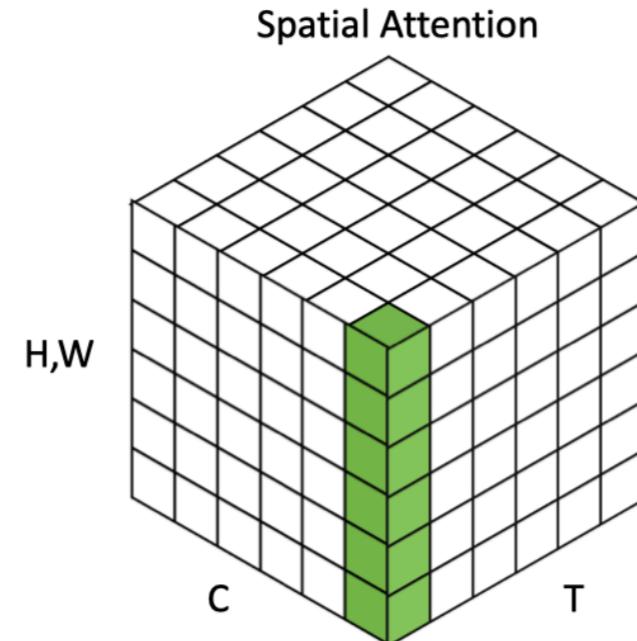
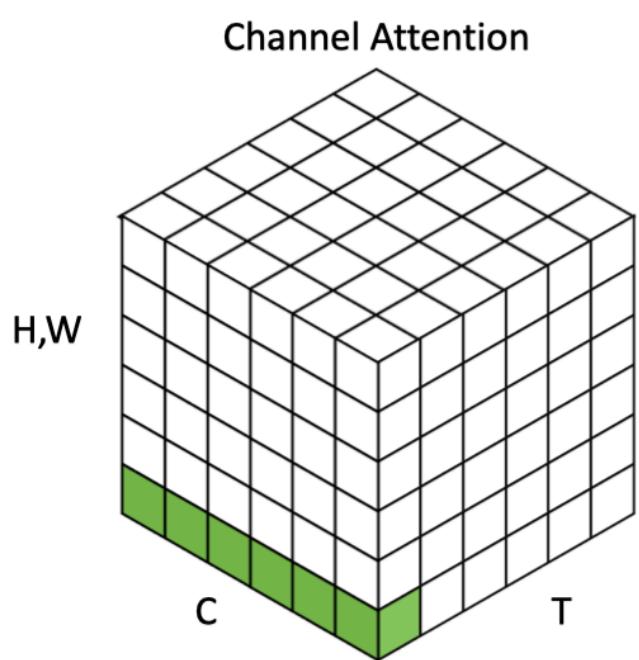
BERT: 72 Köpfe

↳ könnte sich auf unterschiedliche Kombinations-eigenschaften beziehen (z.B. Subjekt/Verb)



© Peltarion

Attention in der Bildverarbeitung



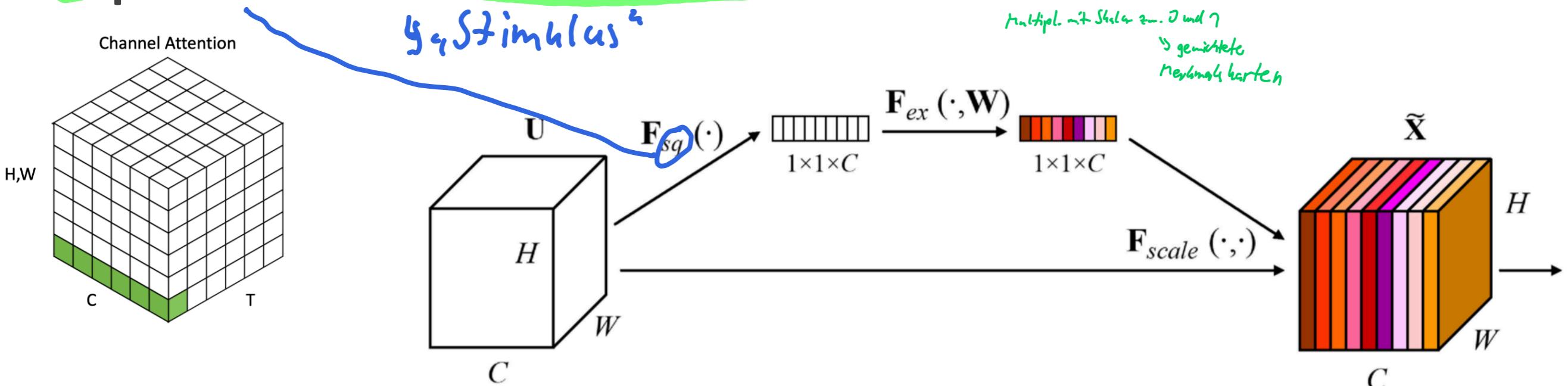
C: Channel
↳ Merkmalskarte
T: Zeit
H,W: Höhe, Breite

Attention in der Bildverarbeitung

Merkmalskarte bleibt; H,W auf
T reduziert

hohes Gewicht für bestimmte Kästle

Squeeze-and-Excitation Network



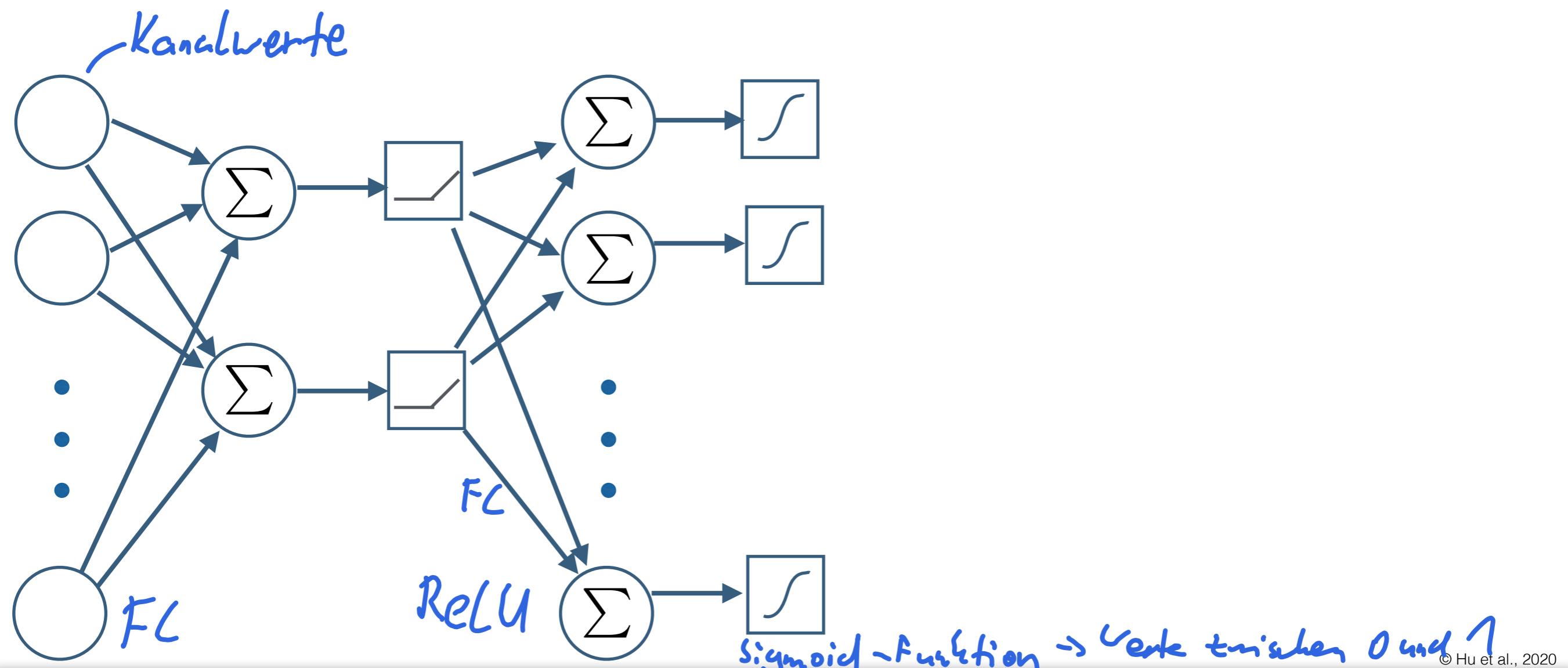
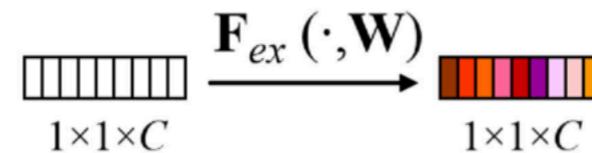
U: Block von Merkmalskarten irgendwo
in einem CNN

Squeeze: Minimum einer Merkmalskarte oder Mittelwert einer Merkmalskarte
 $F_{sq}(\cdot)$
=> Global Average Pooling

Attention in der Bildverarbeitung

Squeeze-and-Excitation Network

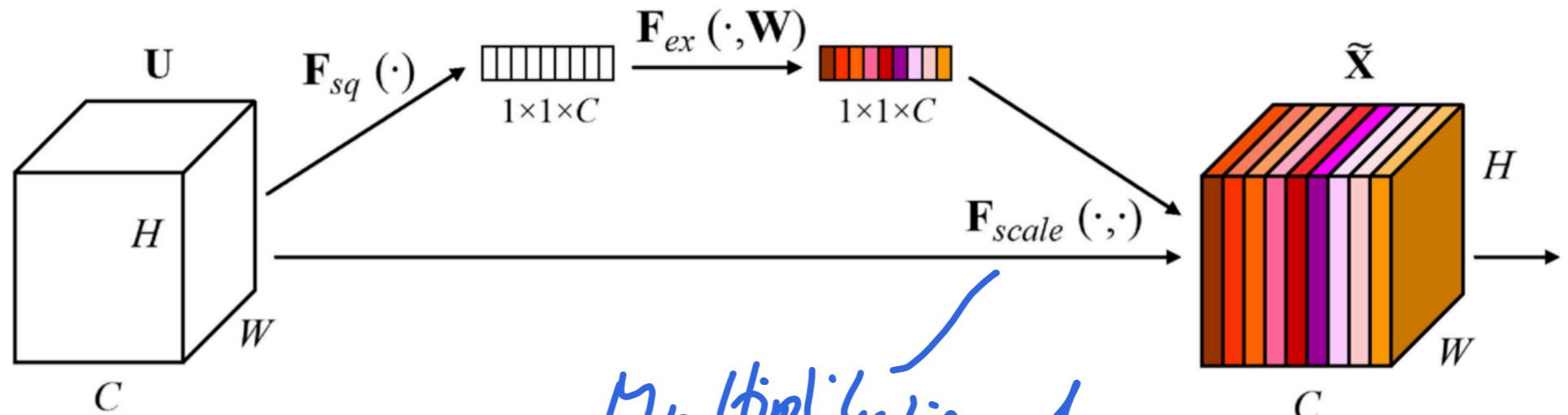
Bottleneck-Architektur



© Hu et al., 2020

Attention in der Bildverarbeitung

Squeeze-and-Excitation Network

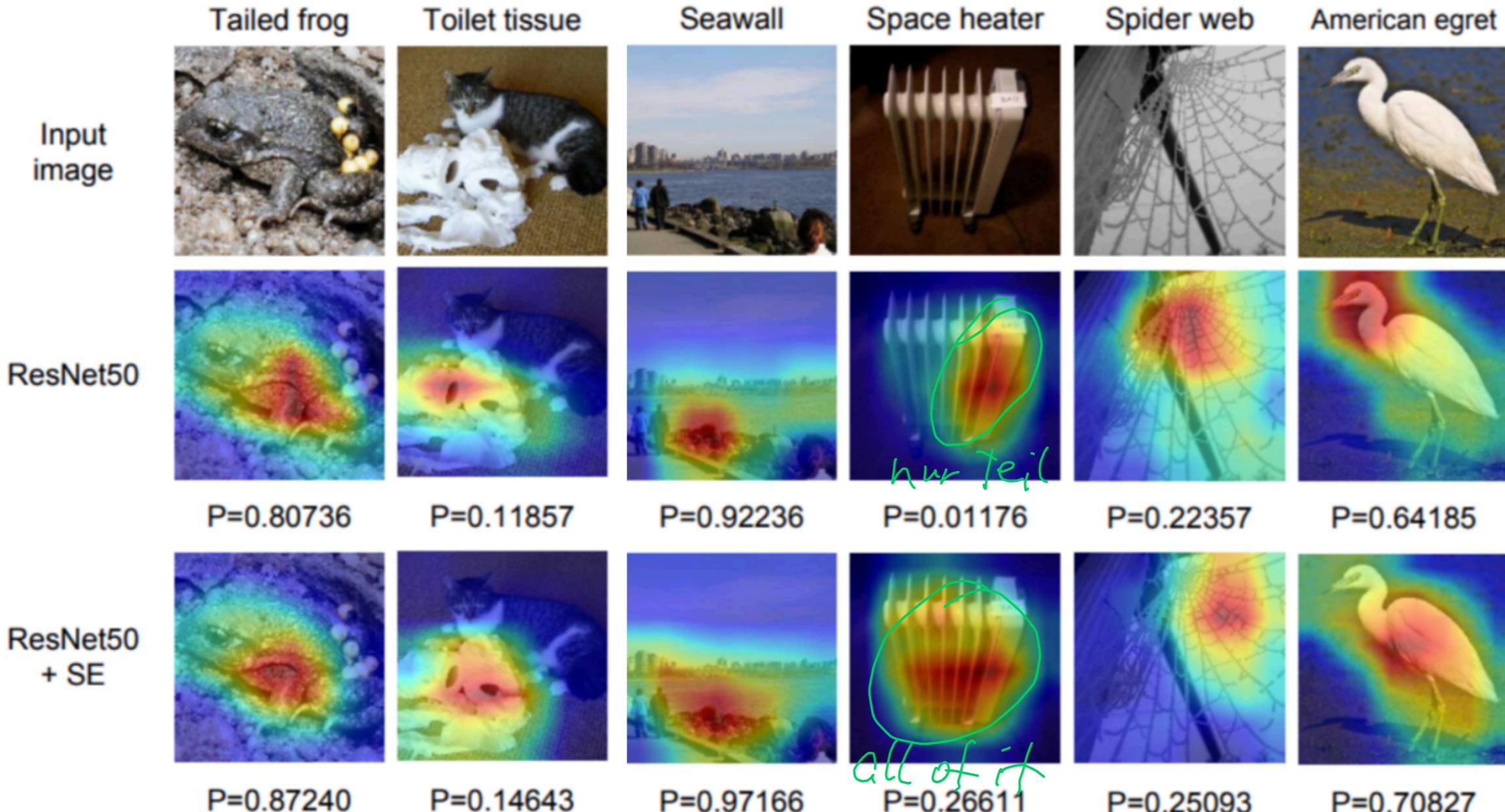


Multiplication of

current channel weight with
result of $F_{ex}(\cdot, \cdot)$

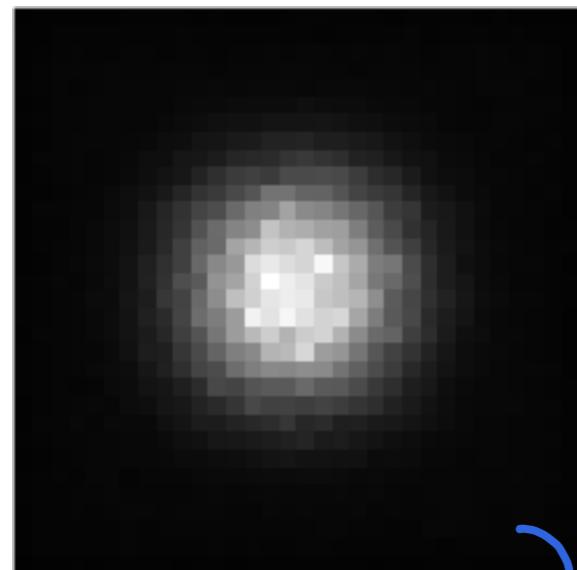
1. Skript räumliche Info zusammen
-Squeeze → nur noch Kanalinformationen vorhanden
2. -Excitation → nach Kontextualisierung (also kontextabhängige Gewichte)
3. -Scale → Ergebnis aus 2 benutzen, um Originale Kanalwerte zu kontextualisieren

Attention in der Bildverarbeitung

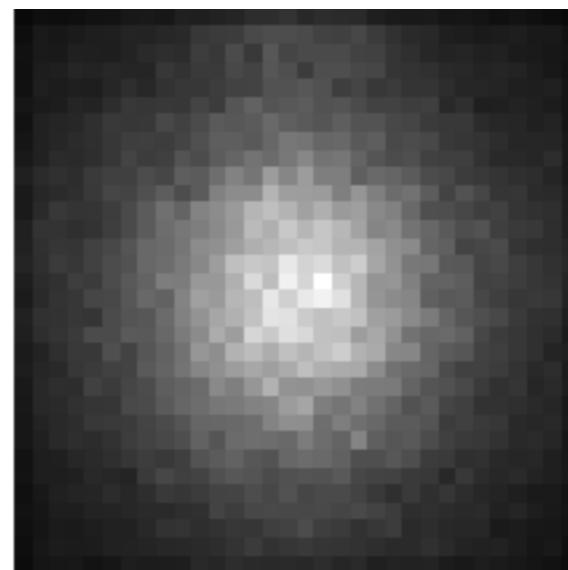


Attention in der Bildverarbeitung

Gather-Excite



vor Training



nach Training

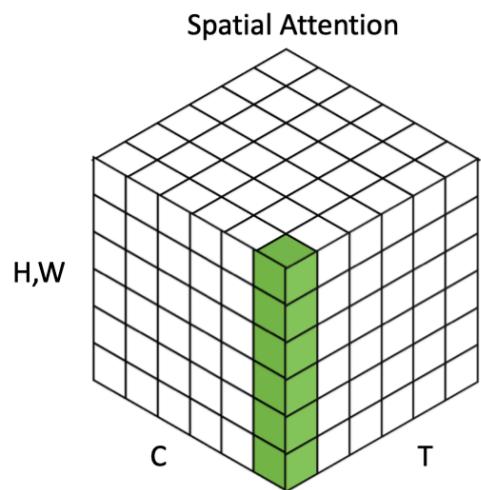
Effektives rezeptives Feld
für CIFAR70

Training so gestaltet, dass rezeptives Feld
größer ist

Intensität: Wie viel Einfluss hat ein Pixel auf den Wert weit
hinter in einem CNN

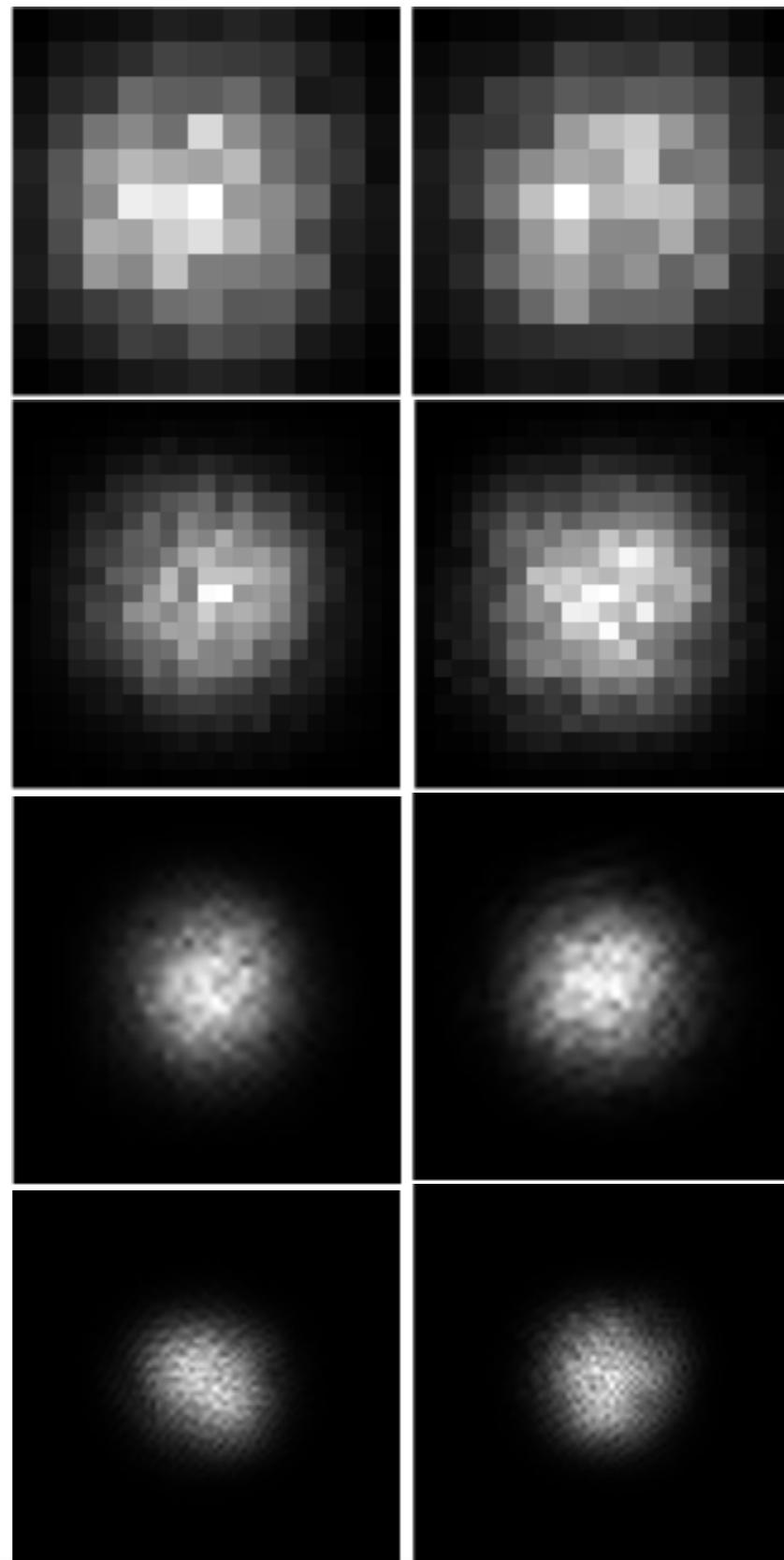
Attention in der Bildverarbeitung

Gather-Excite



links: alle Faltungs-
schichtenelemente: 1

rechts: Faltungsmaßen-
elemente zufüllig



5 Schichten

theoretisch mögliches effektives
Rezeptivitätsfeld: 77 Pixel hier fast erreicht

10 Schichten

theoretisch mögl. ERF: 27 Pixel

20 Schichten

t. ERF: 47 Pixel

je tiefer PN, desto weniger nach kommt na,
an der Rand - Problem - aus: je tiefer, desto eher
größere Abdeckung ERF

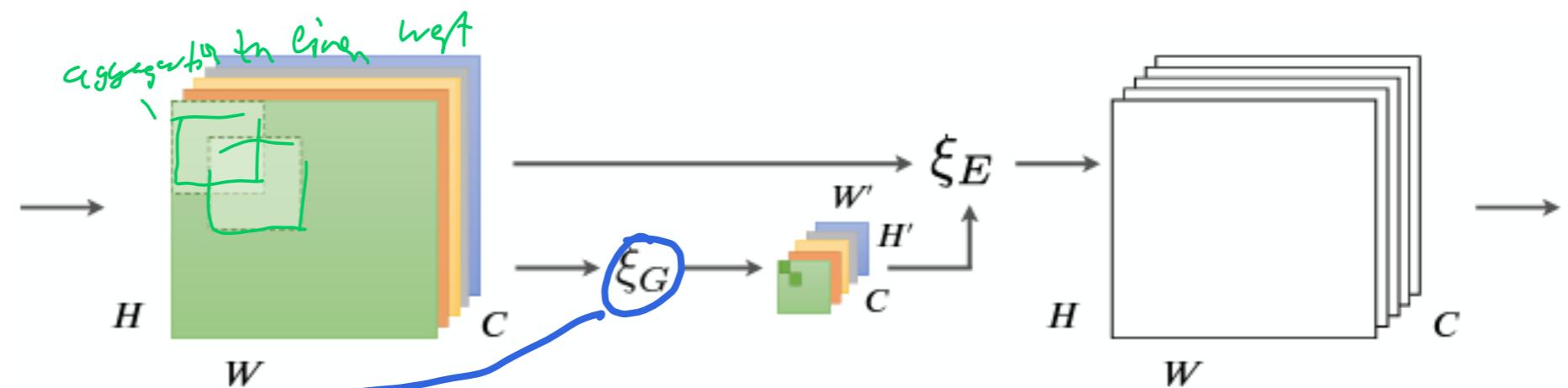
40 Schichten

t. ERF: 87 Pixel

größere Abdeckung ERF
desto weniger nach kommt na,
an der Rand - Problem - aus: je tiefer, desto eher
näher sich ERF einer
Gaußverteilung

Attention in der Bildverarbeitung

Gather-Excite



Gather-Modul: aggregiert räumliche Info von großen räumlichen Nachbarschaften : ξ_G

Excite-Modul: Verteilung dieser Information zurück zu jeder Aktivierung einer Merkmalskarte

Attention in der Bildverarbeitung

Gather-Excite



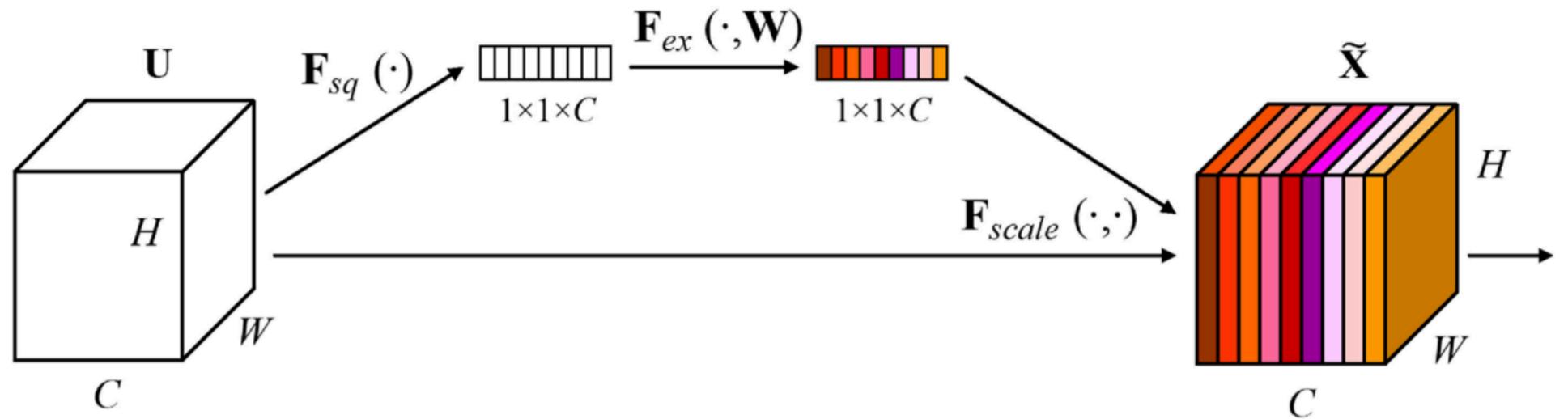
Simple Variante für Gather: Avg-Pooling

Simple Excite Variante: Nearest-Neighbor-Interpolation

Original mit Gewichtung multiplizieren

Attention in der Bildverarbeitung

Squeeze-and-Excitation Network

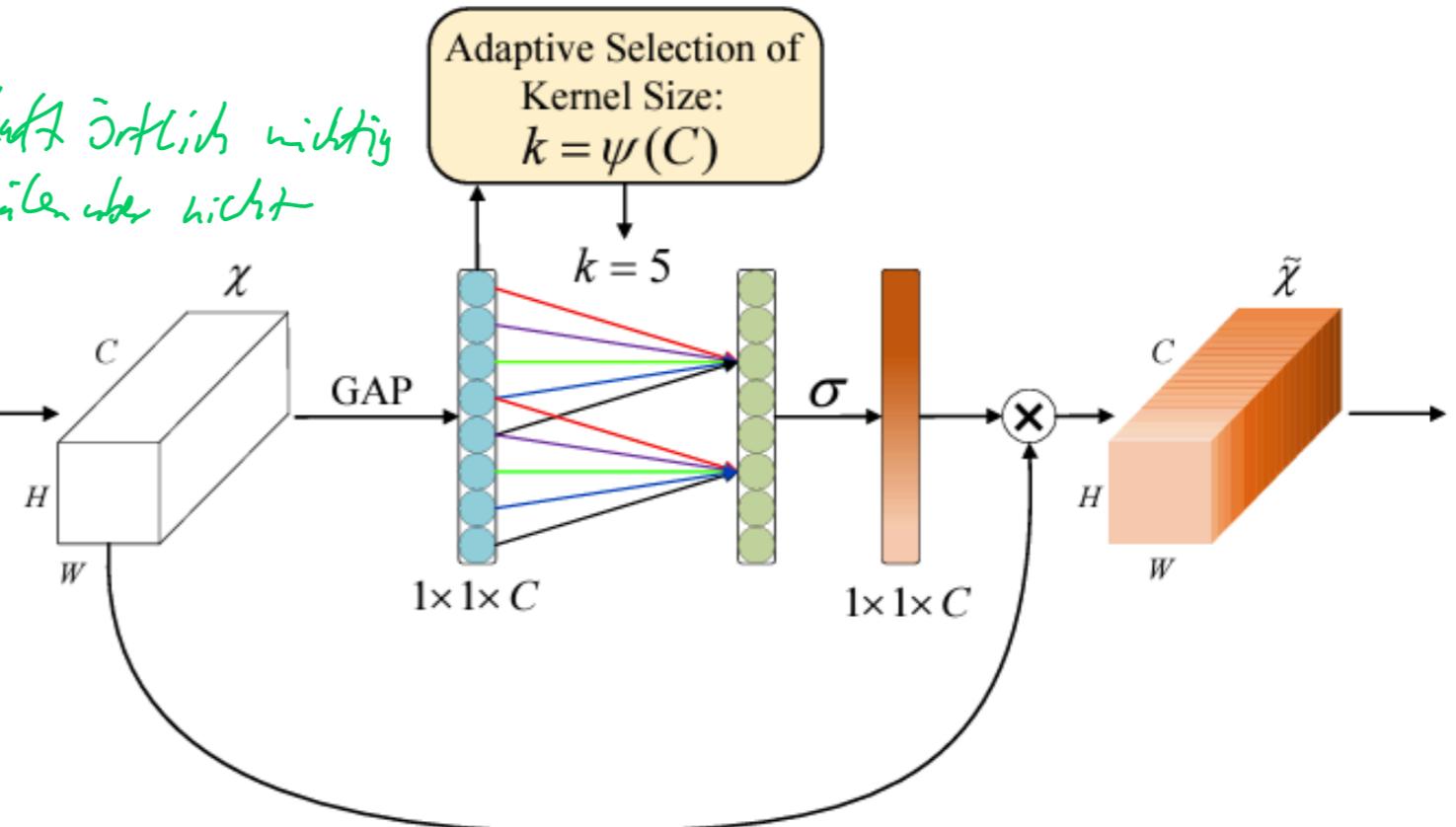


© Hu et al., 2020

Variante: ECA-Net

Warum? Modellierung des Zusammenhangs zwischen benachbarten Kanälen unzureichend bei SE

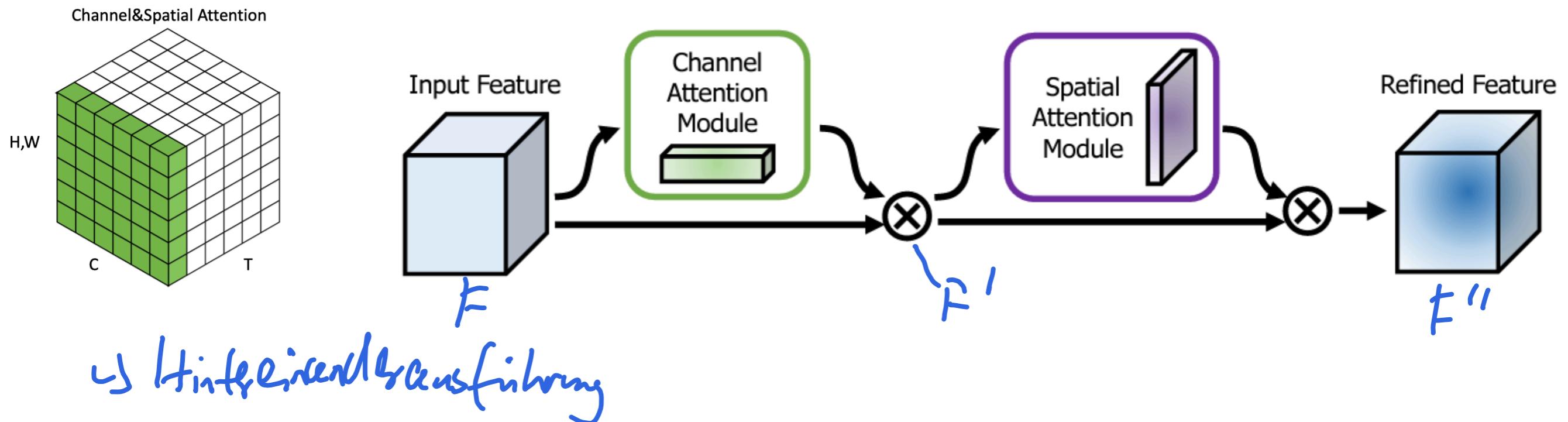
Nachbarschaft örtlich wichtig
„bei Kanälen aber nicht“



© Wang et al., 2020

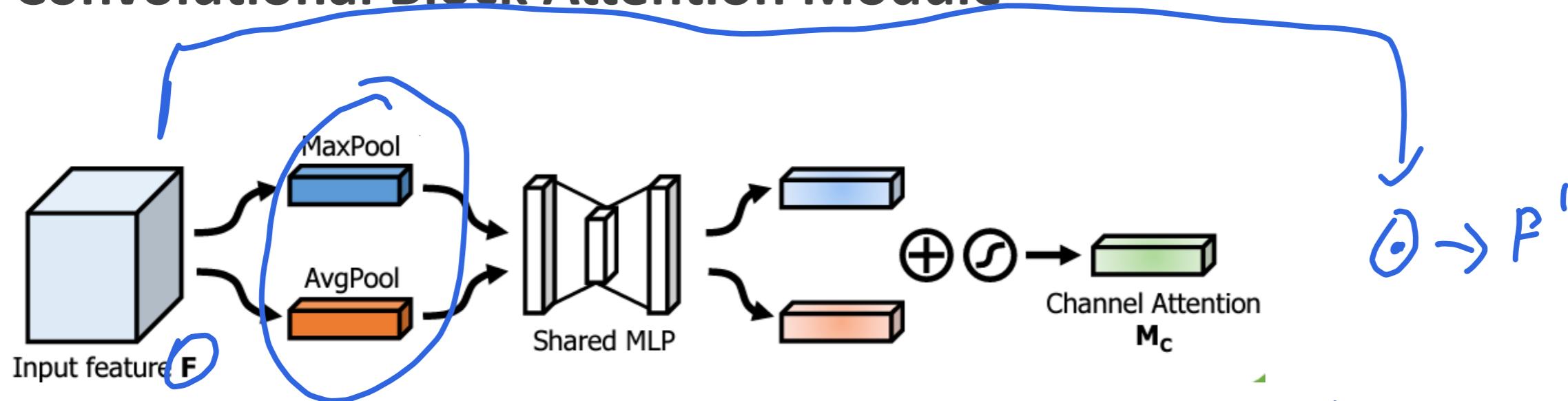
Attention in der Bildverarbeitung

Convolutional Block Attention Module



Attention in der Bildverarbeitung

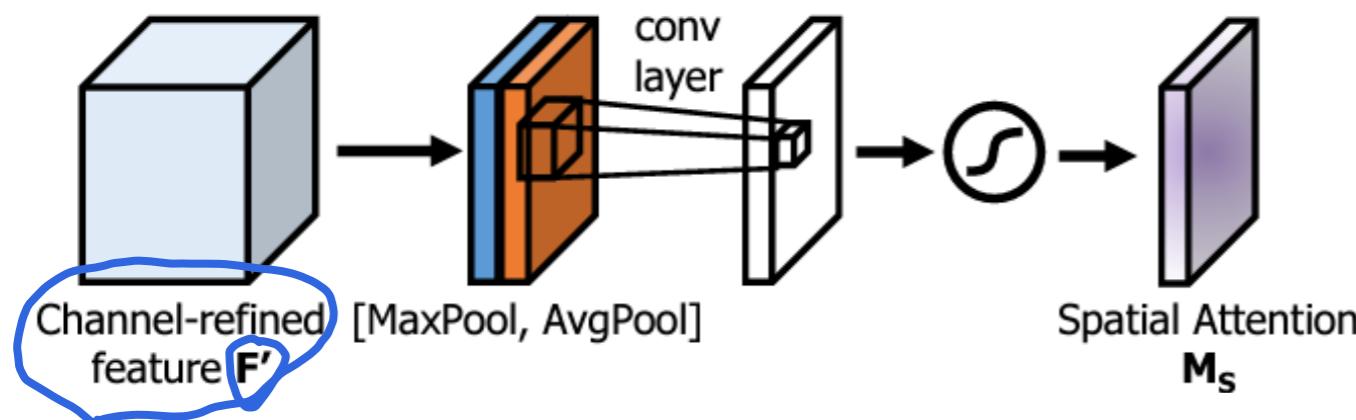
Convolutional Block Attention Module



Channel-Affention: Fokusierung auf das „was“
fast wie SE-Net, aber: Max-Pooling und Avg-Pooling
gefeilte Bottleneck-Architektur, aber getrennte Verarbeitung
Ergebnisse addieren, Sigmoid-Fkt.

Attention in der Bildverarbeitung

Convolutional Block Attention Module

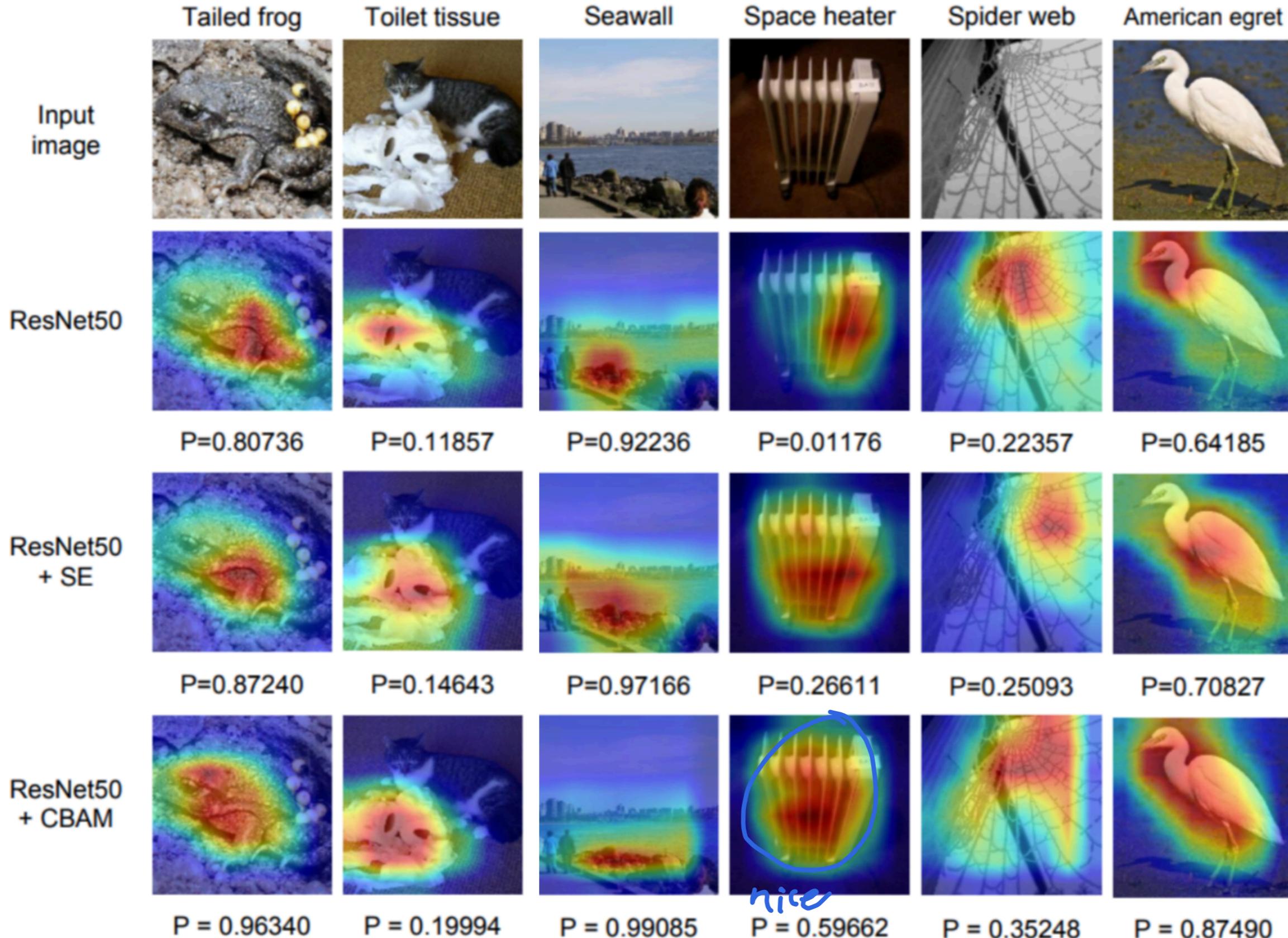


Spatial -Attention : Fokusierung auf das „Wo“

Sowohl Max -Pooling als auch Avg -Pooling
entkoppeln die Achse der Kanäle
+ hochgradig

⇒ Beachte : Input bereits durch Channel -Attention verändert

Attention in der Bildverarbeitung

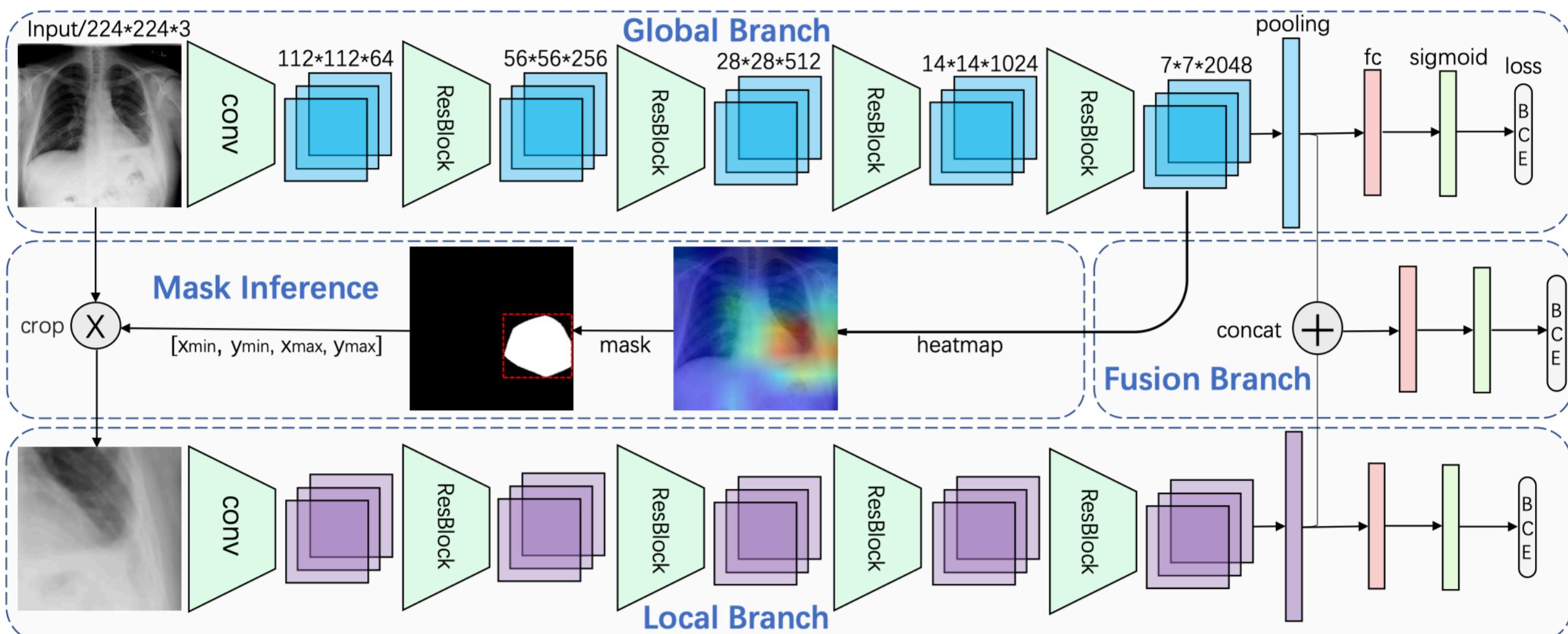


Erkennung
besser

© Ahmadzadeh., 2022

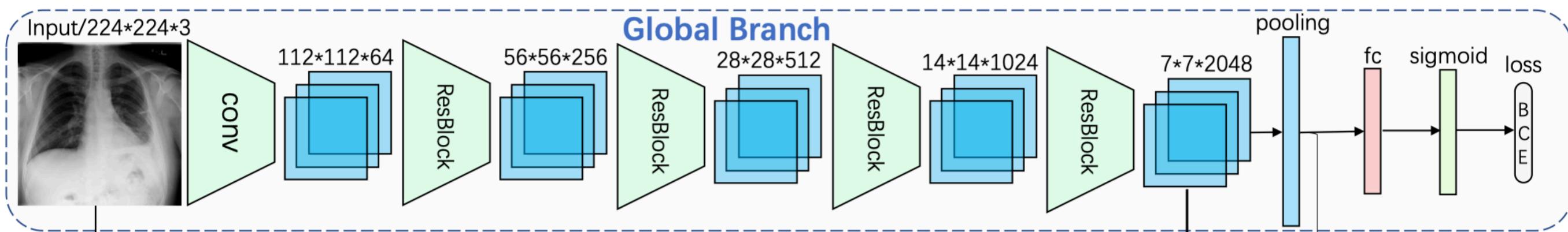
Attention in der Bildverarbeitung

Self-Attention



Attention in der Bildverarbeitung

Self-Attention



↳ verarbeitet gesamtes Bild; bildet Grundlage für Crop mit Heatmap

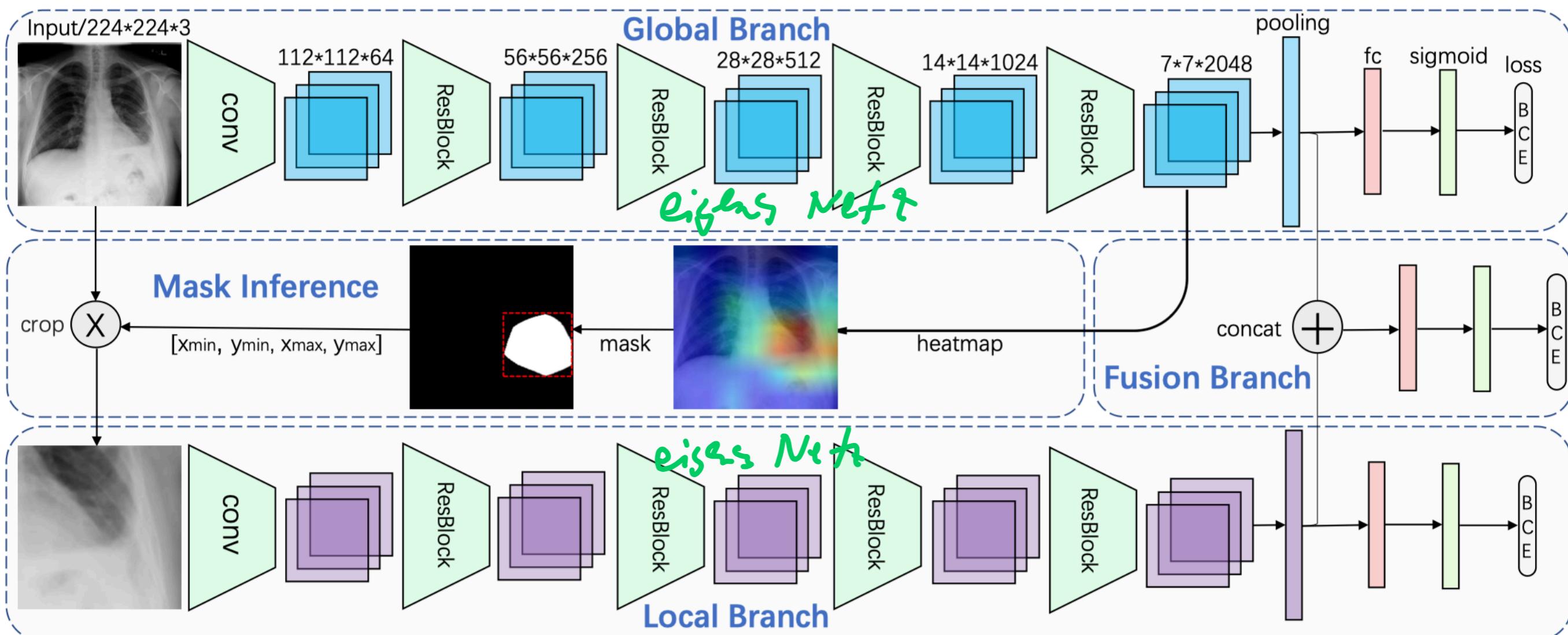
⇒ Heatmap : Mithilfe eines Netw. der Kanäle werden gewichtet an jeder Pixelposition

⇒ Binarisierung ⇒ Regions of Interest

Attention in der Bildverarbeitung

Info kommt nicht von außen hin, sondern von innen (von Input)

Self-Attention



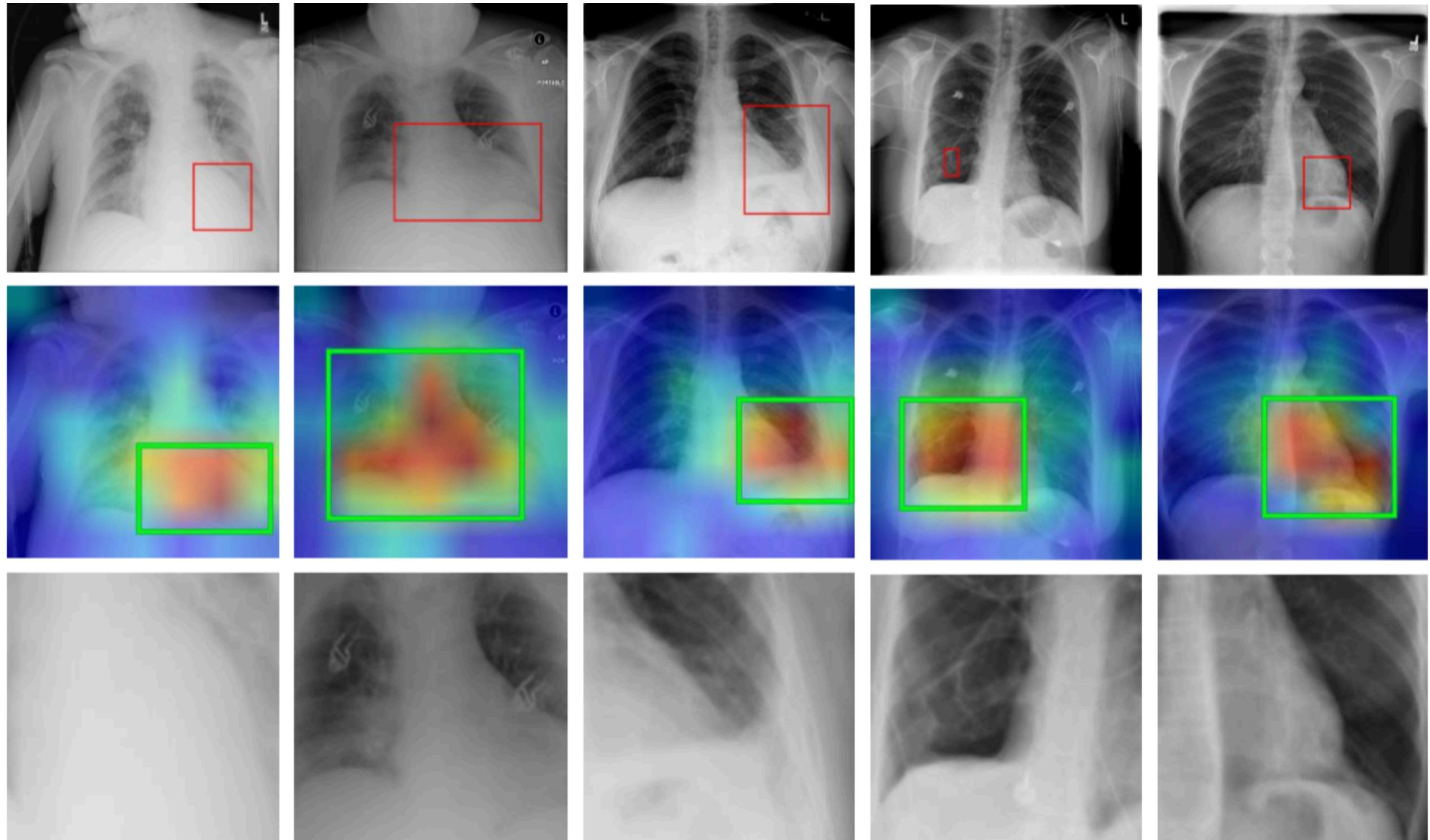
3 unabhängige Bereiche: global, local, fusion

↳ separate Optimierung von lokalem und globalem Branch

↳ kein Gesamttraining möglich wegen Masken

Attention in der Bildverarbeitung

Self-Attention



© Guan et al., 2018