Kompression

Wie hilft Datenkompression bei der Verarbeitung von big data?

Marvin Ahmadi-Moghaddam | Proseminar Effiziente Programmierung

Universität Hamburg | 07.06.2018

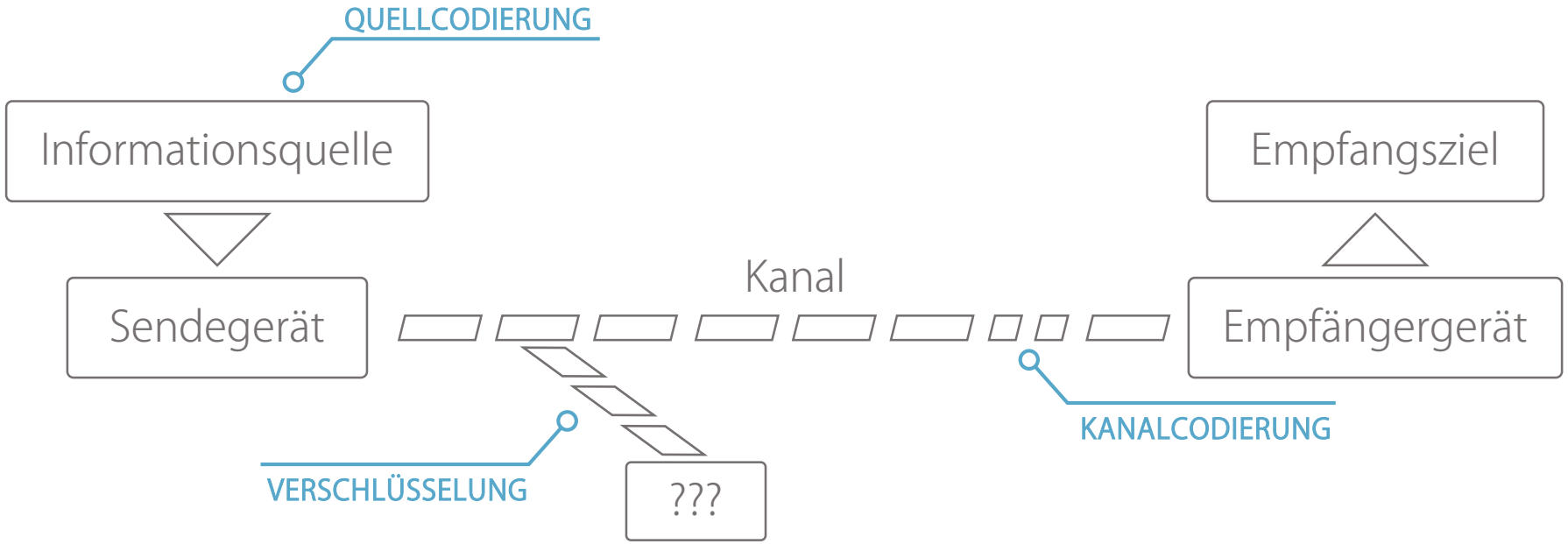
Gliederung

I Was ist Kompression in der Nachrichtentheorie?

II Arten der Kompression und Kompressionsverfahren

III Kompression und Big Data

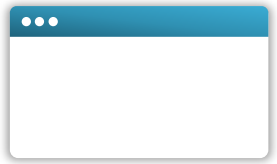
KOMMUNIKATIONSMODELL NACH SHANNON



Verlustfreie & Verlustbehaftete Kompression

Verlustfreie Kompression

Eindeutig korrekte Wiederherstellung



Programme

Übertragung über das Internet

Codierungsverfahren

Prinzip: Informationen anders repräsentieren und Redundanz vermindern

Repräsentation:

5 / V	01 / [Niedrige Spannung] [Hohe Spannung]

Codierungsverfahren

Blockcode: Feste Länge

0000 1100 0001 1111 1011 1000 0001

0000 0001 1011
1100 1111 1000

28 Zeichen im Bitstring

4 Zeichen pro Wort

7 Wörter im Bitstring

Codierungsverfahren

Präfixcode: Variable Länge

0 0 0 0 1 1 0 0 0 0 0 1 1 1 1 1 0 1 1 1 0 0 0 0 0 1

0 0 0 0 | 0 0 1 | 1 0
0 0 0 1 | 1 1 | 0 1

28 Zeichen im Bitstring

2 bis 4 Zeichen pro Wort

11 Wörter im Bitstring

Präfix-Eigenschaft: Eindeutig dekodierbar

Codierungsverfahren

Lauflängencodierung: Redundanz vermindern

0000 1100 0001 1111 1011 1000 0001

40 21 50 61 1031 60 11

28 Zeichen im Bitstring

16 Zeichen lauflängenkodiert

Codierungsverfahren

Entropiekodierung: Häufige Wörter kurz kodiert

a) Shannon-Fano

b) Huffman

c) Arithmetisch

Codierungsverfahren

Huffman:

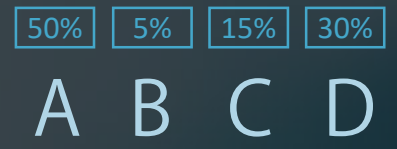
A: 50% B: 5% C: 15% D: 30%

1.) Wörter aufsteigend notieren

2.) Wörter der geringsten Wahrscheinlichkeit zu neuem Wort zusammenfassen

3.) Wiederholen bis 2 Wörter übrig bleiben

=> Kleinstmögliche mittlere Codewortlänge



Codierungsverfahren

Huffman:

A: 50% B: 5% C: 15% D: 30%

1.) Wörter aufsteigend notieren

2.) Wörter der geringsten Wahrscheinlichkeit zu neuem Wort zusammenfassen

3.) Wiederholen bis 2 Wörter übrig bleiben

=> Kleinstmögliche mittlere Codewortlänge

Codierungsverfahren

Huffman:

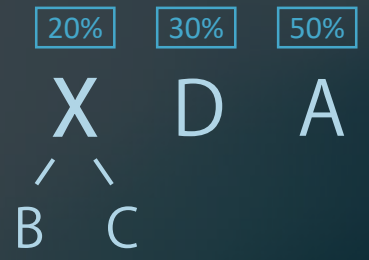
A: 50% B: 5% C: 15% D: 30%

1.) Wörter aufsteigend notieren

2.) Wörter der geringsten Wahrscheinlichkeit zu neuem Wort zusammenfassen

3.) Wiederholen bis 2 Wörter übrig bleiben

=> Kleinstmögliche mittlere Codewortlänge



Codierungsverfahren

Huffman:

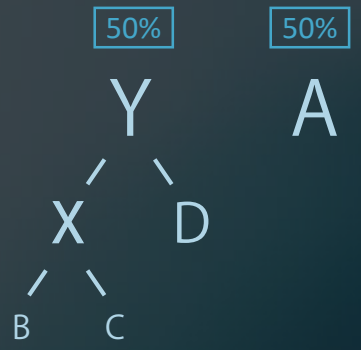
A: 50% B: 5% C: 15% D: 30%

1.) Wörter aufsteigend notieren

2.) Wörter der geringsten Wahrscheinlichkeit zu neuem Wort zusammenfassen

3.) Wiederholen bis 2 Wörter übrig bleiben

=> Kleinstmögliche mittlere Codewortlänge



Codierungsverfahren

Huffman:

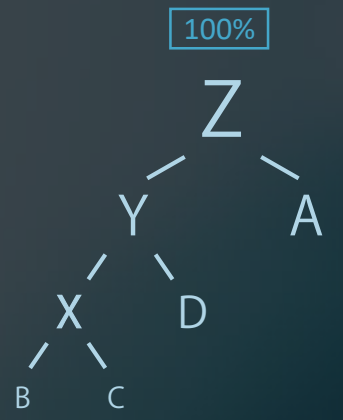
A: 50% B: 5% C: 15% D: 30%

1.) Wörter aufsteigend notieren

2.) Wörter der geringsten Wahrscheinlichkeit zu neuem Wort zusammenfassen

3.) Wiederholen bis 2 Wörter übrig bleiben

=> Kleinstmögliche mittlere Codewortlänge



Codierungsverfahren

Huffman:

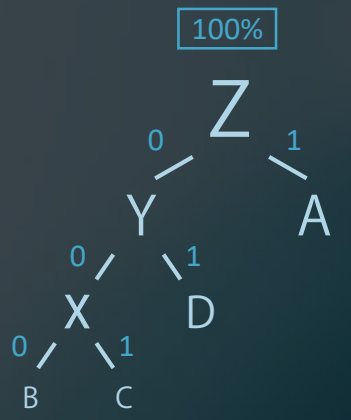
A: 50% B: 5% C: 15% D: 30%

1.) Wörter aufsteigend notieren

2.) Wörter der geringsten Wahrscheinlichkeit zu neuem Wort zusammenfassen

3.) Wiederholen bis 2 Wörter übrig bleiben

=> Kleinstmögliche mittlere Codewortlänge



Kompressionsverfahren

Dictionary-Based

- Teile eines Strings tauchen im Text mehrmals auf
- Dynamisch vs. statisch

LZ77

Abraham Lempel, Jacob Ziv
1977

KOMPRESSIONSVERFAHREN

SLIDING WINDOW (LZ77)

(| |)

OFFSET LÄNGE ZEICHEN

Kompressionsverfahren

Sliding Window (LZ77)

(4 | 1 | B)

Offset Länge Zeichen

Kompressionsverfahren

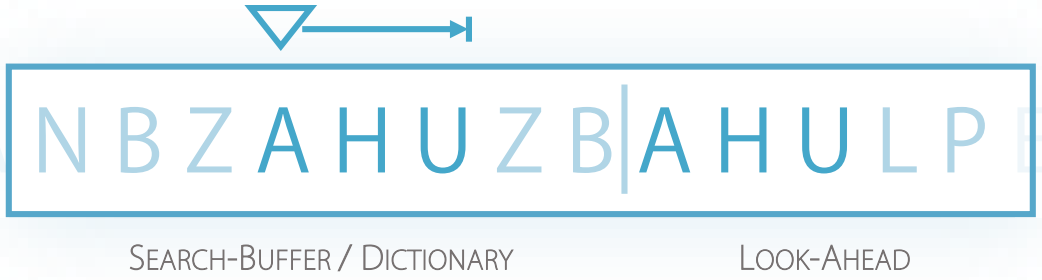
Sliding Window (LZ77)

KOMPRESSIONSVERFAHREN

SLIDING WINDOW (LZ77)

Kompressionsverfahren

Sliding Window (LZ77)



(5 | 3 | L)

Offset Länge Zeichen

Kompressionsverfahren

Sliding Window (LZ77)

Kompressionsverfahren

Sliding Window (LZ77)

(0 | 0 | P)

Offset Länge Zeichen

Kompressionsverfahren

Beispiele aus dem Alltag:

ZIP

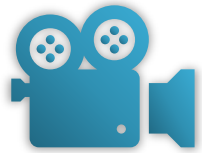
PNG

Verlustbehaftete Kompression

Prinzip: Irrelevante Daten filtern

Bildkompression

Audiokompression



Videokompression

Bildkompression

8 Abstufungen

16 Abstufungen

8-Bit RGB-Modell

256 Abstufungen

256 Abstufungen

256 Abstufungen

16.777.216 Kombinationen

Menschliches Auge

20.000.000
Verschiedene Farben

Bildkompression

Prinzip der natürlichen Bildkompression:

Wenn wir einen Pixel in einem Bild zufällig wählen, ist die Wahrscheinlichkeit hoch, dass benachbarte Pixel die selbe Farbe oder ähnliche Farben haben

[SAL]

Bildkompression

Natürliche Bilder

Y: Luminance -> möglichst geringe Verluste

Cb, Cr: Color -> Verluste dürfen höher sein

Eindeutige Farbabstufungen

Ähnliche Bereiche identifizieren und darauf verweisen

Datenkompression in Big Data

6 Vs von Big Data

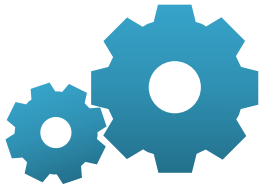
Volume:
Größe, die für vorliegende Computer nicht mehr einfach zu verarbeiten sind

Velocity:
Menge an Strömen, die zu einem Big Data Strom fließen

Variety:
Verschiedene Datenquellen in verschiedenen Formaten

Komplexität von Big Data

1.) Komplexität der **Daten**



2.) Komplexität der **Verarbeitung**

3.) Komplexität der **Systeme**

=> **Aussortieren** und **Zusammenfassen**

Verminderung der Daten

Volume:	Velocity:	Variety:
Größe, die für vorliegende Computer nicht mehr einfach zu verarbeiten sind	Menge an Strömen, die zu einem Big Data Strom fließen	Verschiedene Datenquellen in verschiedenen Formaten
Verminderung der Datenmenge	Zusammenfassen und Sortieren zur Verminderung der Redundanz	Zusammenfassen um Heterogenität zu vermindern

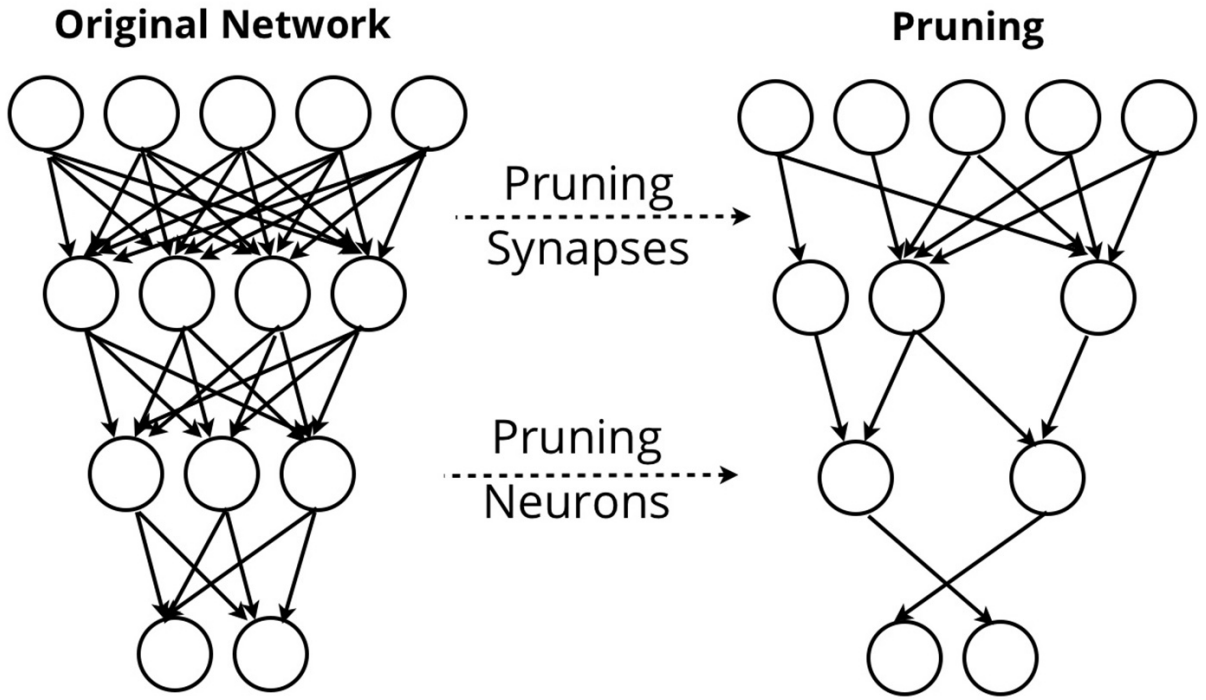
=> Prinzip der Minimierung von Redundanz

Deep Learning Beispiel

1.) Entfernen der überflüssigen Verbindungen

2.) Quantisierung und Gewichtung

3.) Huffman-Kodierung

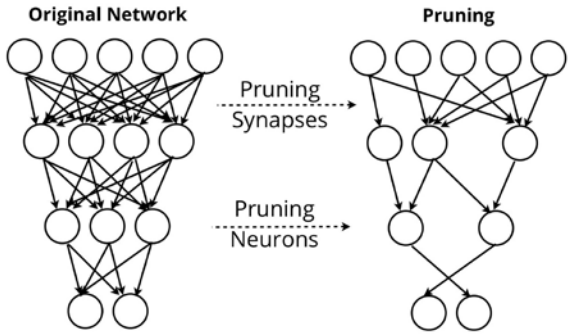


Deep Learning Beispiel

1.) Entfernen der überflüssigen Verbindungen

2.) Quantisierung und Gewichtung

3.) Huffman-Kodierung



Dateigröße vermindern

35 bis 50-mal

Echtzeitübertragung möglich

Häufiges aktualisieren auf mobielen Geräten

Fazit

Kompression

Wie hilft Datenkompression bei der Verarbeitung von big data?

Quellen

Lorica (2016) Compressed representations in the age of big data O'REILLY.
Abgerufen am 05.06.2018 von: https://www.oreilly.com/ideas/compressed-representations-in-the-age-of-big-data

ur Rehman, Liew, Abbas, Jayaraman, Wah, Khan (2016) Big Data Reduction Methods: A Survey Springer Berlin Heidelberg.
Abgerufen am 05.06.2018 von https://link.springer.com/article/10.1007/s41019-016-0022-0

The History of Data Compression (11.11.2014) Commander
Abgerufen am 05.06.2018 von http://blog.commander.com/the-history-of-data-compression/

Shannon (1948) A Mathematical Theory of Communication The Bell System Technical Journal.
Abgerufen am 05.06.2018 von https://link.springer.com/article/10.1007/s41019-016-0022-0

The History of Data Compression (11.11.2014) Commander
Abgerufen am 05.06.2018 von http://blog.commander.com/the-history-of-data-compression/

History of Lossless Data Compression Algorithms (20.08.2014) Engineering and Technology History Wiki
Abgerufen am 05.06.2018 von http://ethw.org/History_of_Lossless_Data_Compression_Algorithms

Wie viele Farben sieht der Mensch? (15.05.2014) Schweizer Radio und Fernsehen
Abgerufen am 05.06.2018 von https://www.srf.ch/sendungen/einstein/fuenfmalklug/wie-viele-farben-sieht-der-mensch

Datenkompression (02.05.2018) Wikipedia
Abgerufen am 05.06.2018 von https://de.wikipedia.org/wiki/Datenkompression

Salomon (2008) A Concise Introduction to Data Compression Springer-Verlag London