Vergleich verlustfreier Datenkompressionsverfahren auf Bilddaten

Nick Schreiber
Technische Hochschule Rosenheim
Master Informatik, Seminar theoretische Informatik
Email: nick.schreiber@stud.th-rosenheim.de

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

I. Einleitung

Datenkompression beschreibt ein Verfahren, das zum Ziel hat, eine Nachricht ohne relevanten Informationsverlust zu verkleinern. Als Nachricht ist jede Art von digitalen Daten gemeint, z.B. Text, Bild, Audio, etc.. Daten können komprimiert werden, indem Redundanz entfernt oder eine Kodierung angewendet wird. Daher wird Datenkompression oft als Kodierung bezeichnet. Kodierung ist ein allgemeiner Begriff, der jede spezielle Darstellung von Daten nach einem bestimmten Schema umfasst. [1]

Es gibt zwei Arten der Datenkompression: die verlustbehaftete und die verlustfreie Kompression. Bei der verlustbehafteten Datenkompression kann eine bestimmte Menge an Information durch die Kompression verloren gehen, was in Kauf genommen wird, da dadurch die Datenmenge erheblich reduziert werden kann oder weil die verlorene Informationen für die Anwendung kaum relevant sind. Das wird auch als Irrelevanzreduktion bezeichnet [2, S. 5]. Ein Beispiel für Irrelevanzreduktion kann bei Audiosignalen beobachtet werden. Der menschliche Hörfrequenzbereich liegt zwischen 20 Hz und 20 kHz [3]. Daher ist es nicht sinnvoll, Frequenzen, die weit außerhalb des hörbaren Bereichs liegen, in Audiodateien zu speichern. Bei der verlustfreien Datenkompression wird die Integrität der Daten bewahrt. Das bedeutet, dass sämtliche Informationen in den komprimierten Daten enthalten sind und die Originaldaten vollständig rekonstruierbar sind. In dieser Arbeit wird nur die verlustfreie Datenkompression untersucht, da Irrelevanzreduktion nicht direkt zum Themengebiet der Datenkompression gehört.

Die Datenkompression von Bildern wird aus verschiedenen Gründen eingesetzt. Speichernutzung: Unkomprimierte Bilddaten können beträchtlich mehr Speicherplatz beanspruchen. Übertragungseffizienz: Bei der Übertragung von Bildern über Netzwerke oder das Internet spielt die Übertragungseffizienz eine entscheidende Rolle. Wenn ein Bild über einen Kanal mit begrenzter Bandbreite gesendet wird, kann es effizienter sein, das Bild zu komprimieren, es zu übertragen und dann beim Empfänger zu dekomprimieren. Dadurch wird die Übertragungszeit verkürzt und das Bild kann schneller bereitgestellt werden. Dies führt zu einer höheren Übertragungsrate und einer reduzierten Bandbreitennutzung.

II. Zielsetzung der Arbeit

Ziel der Arbeit ist es zu untersuchen, ob und warum bestimmte verlustfreie Datenkompressionsverfahren für Bilddaten besser geeignet sind als andere. Dazu werden die theoretischen Aspekte der Kompressionsalgorithmen untersucht. Außerdem wird untersucht, wie Bilddaten aufgebaut sind und welche Besonderheiten in der Datenstruktur für die Datenkompression genutzt werden können. Die Arbeit beinhaltet auch einen praktischen Teil. Verschiedene Algorithmen zur verlustfreien Datenkompression wurden manuell implementiert und an unterschiedlichen Bilddaten getestet. So konnten konkrete Ergebnisse über die Leistungsfähigkeit der Algorithmen gewonnen werden. Die verglichenen Algorithmen sind Run Length Encoding (RLE), Huffman Encoding, Lempel-Ziv 1977 (LZ77), PNG Algorithmus und verschiedene Kombinationen der Algorithmen. Die Ergebnisse werden interpretiert und mit den theoretischen Erwartungswerten verglichen.

III. Grundlagen zur Datenkompression

In der Informatik gehört die Datenkompression zum Teilgebiet der Informationstheorie. Um zu verstehen, wie verlustfreie Datenkompression funktioniert, muss man einige theoretische Grundlagen kennen.

A. Information

Claude Shannon, der Erfinder der Informationstheorie, definiert den Begriff Information als Maß für den Informationsgehlt. Information ist ein Maß der Unsicherheit, das durch das Eintreten eines bestimmten Ereignisses oder das Empfangen einer Nachricht verringert wird. [4] Information ist die Mindestanzahl von Bits, die zur Codierung einer Nachricht verwendet werden müssen. [5] Die grundlegende Idee zu Information besteht darin, dass Informationen umso wertvoller sind, je unerwarteter oder unwahrscheinlicher sie sind.

B. Entropie

Die Quantifizierung des Informationsgehalts erfolgt durch die Entropie (H). Formal drückt die Entropie die durchschnittliche Menge an Bits aus, die benötigt werden, um eine Information zu kodieren. [4] Die Entropie berücksichtigt

die Wahrscheinlichkeiten verschiedener möglicher Ereignisse und erreicht ein Maximum, wenn alle Ereignisse gleich wahrscheinlich sind, was auf maximale Unsicherheit hinweist.

$$H(X) = -\sum_{i=1}^{n} P(x_i) \cdot \log_2(P(x_i)) \tag{1}$$

Formel 1 definiert die Entropie mathematisch. Hierbei steht $H(X)$ für die Entropie der Menge X. $P(x_i)$ steht für die Wahrscheinlichkeit des Auftretens des Ereignisses x_i. Die Summe wird über alle möglichen Ereignisse x_i in X gebildet.

Diese Formel beschreibt die durchschnittliche Menge an Bits, die benötigt werden, um eine Nachricht aus X zu kodieren. Wenn die Entropie hoch ist, ist die Unsicherheit groß, und es werden mehr Bits benötigt, um die Informationen zu repräsentieren. Wenn die Entropie niedrig ist, gibt es weniger Unsicherheit, und somit werden weniger Bits benötigt.

Man kann nun eine direkte Verbindung zwischen Entropie und Kompression herzustellen. Niedrige Entropie bedeutet, dass eine Datenmenge strukturiert ist, bzw. Muster aufweist. Das heißt, dass in den Daten wenig unsicherheit ist und die Daten Redundanz enthalten. Das bedeutet, niedrige Entropie sagt, dass die Daten komprimiert werden können.

C. Redundanz und Mutual Information

Redundanz beschreibt Informationen die in Daten mehrfach vorhanden sind. [6] Einfach gesagt kann man Redundanz als überflüssige Information betrachten. Eine hohe Redundanz sagt aus, dass sich wiederholende oder vorhersehbare Muster innerhalb der Daten befinden.

Um eine Formel für die Redundanz aufzustellen benötigt man die mittlere Codewortlänge. Die mittlere Codewortlänge gibt den durchschnittlichen Bedarf an Bits pro Symbol in einer Nachricht an. Sei X ein Alphabet und $x \in X$. $C(x)$ bezeichnet das zu x gehörende Codewort. $l(x)$ bezeichnet die Länge von $C(x)$. Die mittlere Codewortlänge L(C) einer Nachricht C(x) mit der Wahrscheinlichkeitsverteilung p(x) ist in Formel 2 definiert.

$$L(C) = \sum_{i}^{|X|} p(x_i) \cdot l(x_i) \tag{2}$$

Mit der mittleren Codewortlänge lässt sich nun die Redundanz des Codes, bzw. der Nachricht berechnen. Die Formel 3 definiert die Redundanz einer Nachricht.

$$R_{\text{Code}} = L(C) - H(X) \tag{3}$$

Die Redundanz wird berechnet, indem von der tatsächlichen durchschnittlichen Anzahl an Bits pro Symbol die theoretisch minimale Anzahl an Bits pro Symbol abgezogen werden. Die theoretisch minimale Anzahl an Bits pro Symbol entspricht der enthaltenen Information und ist gleich der Entropie der Nachricht. Daraus ergibt sich, dass die Redundanz ≥ 0 sein muss.

Mutual Information ist ein quantitatives Maß für die gegenseitige Abhängigkeit von zwei Variablen. [4] Es misst, wie sehr die Kenntnis einer Variablen die Unsicherheit über die andere Variable reduziert. Dieses Konzept ist entscheidend, um die Struktur von Daten zu verstehen und voneinander abhängige Informationen zu erkennen.

Wenn die Mutual Information zwischen zwei Variablen hoch ist, bedeutet dies, dass das Wissen über eine Variable bedeutende Informationen über die andere Variable liefert. Hohe Mutual Information sagt dementsprechend aus, dass zwei Variablen stark voneinander Abhängig sind. Das Wissen über den Wert einer Variable trägt bereits wesentlich zur Vorhersage oder zum Verständnis der anderen Variable bei.

Geringe Mutual Informaion sagt aus, dass die beiden Variablen weniger gemeinsame Information teilen. Das Wissen über den Wert einer Variable trägt nicht so stark zur Vorhersage oder zum Verständnis der anderen Variable bei. Das bedeutet eine schwächere Statistische Abhängig der Variablen.

Durch das erkennen von Mutual Information kann gezeigt werden, dass Muster und/ oder Wiederholungen und dementsprechend Redundanz in den Daten enthalten ist. Redundanz spielt im Bezug auf Kompression eine wichtige Rolle. Kompression funktioniert durch das identifizieren und eliminieren redundanter Elemente, um den Informationsgehalt zu maximieren und die Effizienz von Datenrepräsentationen zu steigern.

IV. Informationstheorie

Unterscheidung zwischen Daten und Information: Die Unterscheidung zwischen Daten und Information liegt in der Verarbeitung und Bedeutung. Daten sind rohe Fakten oder Symbole, die an sich keine spezifische Bedeutung haben. Informationen entstehen durch die Interpretation, Organisation und Strukturierung von Daten, wodurch ein sinnvoller Kontext geschaffen wird. Daten werden zu Informationen, wenn sie für einen bestimmten Zweck verwendet werden können. Im Kontext der Datenkompression ist es wichtig zu verstehen, dass nicht alle Daten gleichermaßen informativ sind, und effektive Kompressionsalgorithmen sollten redundante oder weniger informative Teile der Daten eliminieren, um die wesentliche Information zu bewahren.

A. Quellencodierungstheorem/ Source Coding Theorem

Erklärung der Theorie, dass Daten nicht bis ins Unendliche komprimiert werden können. Darlegung der Grenzen und Limitationen der Datenkompression. Schubfachprinzip keine unendliche kompression: Betonung der Bedeutung von Entropie in diesem Kontext.

B. Kolmogorov Komplexität

Die Kolmogorov-Komplexität dient als Maß für die algorithmische Komplexität von Informationen. Diese Komplexität wird durch die kürzeste Beschreibungslänge eines Algorithmus definiert, der eine bestimmte Menge an Daten erzeugt. Eine enge Verbindung besteht zur Vorstellung der optimalen, universellen Datenkompression – je einfacher die Beschreibung, desto effizienter kann die Datenmenge komprimiert werden.

Die Kolmogorov-Komplexität ermöglicht es, den Informationsgehalt eines Datensatzes in Bezug auf die kürzeste mögliche algorithmische Beschreibung zu verstehen. Dieses Konzept hat weitreichende praktische Anwendungen, von der Identifikation von wiederholten Mustern bis hin zur Bewertung von Algorithmeneffizienz. Allerdings stehen der Anwendung auch Herausforderungen gegenüber, wie der Unberechenbarkeit des absoluten Komplexitätsmaßes und der Schwierigkeit, universelle Algorithmen zu finden, die für alle Daten gleichermaßen effizient sind. Das Verständnis der Kolmogorov-Komplexität trägt maßgeblich zur Entwicklung fortgeschrittener Datenkompressionsverfahren bei und bietet Einblicke in die Grenzen der Komprimierbarkeit von Informationen.

C. Anforderungen an Daten, damit diese komprimierbar sind

Strukturierte Daten vs. unstrukturierte Daten. Wiederholungen und Muster in den Daten. Zusammenhang zwischen Datenqualität und Kompressionsmöglichkeiten.

Bsp. unkomprimierbare Daten: gleichverteilte Zufallszahlen, ohne Struktur, ..

V. Introduction

This document is a model and instructions for LaTeX. Please observe the conference page [7] limits.

VI. Ease of Use

A. Maintaining the Integrity of the Specifications

The IEEEtran class file is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

References

[1] F. M. Ingels, *Information and coding theory*. Scranton, Intext Educational Publishers, 1971.

[2] J. S. Peter Maluck, "Quellencodierung, gelenktes entdeckendes lernen," *EducETH*, 2009.

[3] E. Burke, "Schallquellensystem zur untersuchung der wahrnehmung von infraschall in kombination mit hörschall," *Physikalisch-Technische Bindesanstalt*, 2019.

[4] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948. [Online]. Available: http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf

[5] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press, 1949.

[6] B. Friedrichs and B. Friedrichs, "Grundlagen der shannon'schen informationstheorie," *Kanalcodierung: Grundlagen und Anwendungen in modernen Kommunikationssystemen*, pp. 33–68, 1996.

[7] D. Autor, "Titel des artikels," *Name des Journals*, pp. 123–145, 2023.