Machine Learning in Data Compression

KEVIN MALINOWSKI

30.11.2017

Seminar „Effiziente Programmierung“
Arbeitsbereich Wissenschaftliches Rechnen
Fachbereich Informatik
Fakultät für Mathematik, Informatik und Naturwissenschaften
Universität Hamburg

Data Compression 101

- Anzahl an Bits reduzieren
- Lossy / Lossless Compression
- Model and Coder
- Wahrscheinlichkeitsverteilung
- Kürzere Codes für wahrscheinlichere Symbole
- Random Data kann nicht komprimiert werden
- Verstehen -> Compression

Neural Network Data Compressor

- Entwickelt 1996 Schmidhuber Heil
- Dreilagiges Neuronales Netzwerk
- 80 Zeichen Alphabet
- Training/Predictionphasen
- Backpropagation
- unbrauchbar(langsam)

Neural Network Data Compressor

- Mahoney (2000)
- ein bit zur Zeit
- training online
- erstes layer replaced -> Hashfunktion
- Ein Neuron pro Context – Ordnung (1-5)
- 10^5 mal schneller
- Zähler an jedem Gewicht

PAQ

- Mahoney (2002)
- Zähler allein können für Vorhersagen benutzt werden
- Weights beseitigt
- Zähler werden mit Kontext verglichen
- Zähler 1 Byte

Secondary Symbol Estimation

- PAQ2 (Mai 2003)
- Inputs: Vorhersage+Context
- Output: Verbesserte Vorhersage
- Quantisiert zu 64 nonlinearen Werten(„level“)

Secondary Symbol Estimation

- PAQ3 (Sep 2003)
- 32 Quantisierungslevel
- Interpolation zwischen den nähesten Werten
- Anpassung beim Update zur Fehlerreduzierung

Adaptive Linear Mixing

- PAQ4 (Nov 2003)
- 19 Modelle machen unabhängige Vorhersagen
- Verbunden durch gewichtete Summierung
- Gewichte werden angepasst um genauere Modelle zu bevorzugen
- Neue Daten werden gegenüber Alten bevorzugt
- Überhang halbieren

Input	n0	n1	p(1)
0000000000	10	0	0/10
00000000001	6	1	1/7
000000000011	4	2	2/6
0000000000111	3	3	3/6
00000000001111	2	4	4/6
000000000011111	2	5	5/7
0000000000111111	2	6	6/8

Logistic Mixing

- PAQ7 (Dez 2005)
- Transformation der Vorhersagen in den Logistischen Bereich

$$p = \text{squash}(\Sigma_i\, w_i\, \text{stretch}(p_i))$$

$$\text{stretch}(p) = \ln(p/(1 - p))$$
$$\text{squash}(x) = \text{stretch}^{-1}(x) = 1/(1 + e^{-x})$$

$$w_i := w_i + \lambda\, (y - p)\, \text{stretch}(p_i)$$

Dynamic Markov Compression

- PAQ8L (März 2006)
- Tabelle aus variablen Kontextlängen der Bitketten
- Jeder Eintrag hat Zeiger nächstmögliche Folge
- Folgengrenze 1 Byte -> Links Abschnitt
- Cloning um besseren Kontext zu erlangen
- Muss Oft genug vorkommen
- Genügend andere Einträge müssen zum alten Eintrag zeigen
- Bei voller Tabelle wird gesamtes Modell reinitialized
- PAQ8L erhöht Thresholds

Dynamic Markov Compression

State	n_0	n_1	$next_0$	$next_1$
A = 11111	4		B	
B = 110	3	7	E	F

State	n_0	n_1	$next_0$	$next_1$
A = 11111	4		C	
B = 110	1.8	4.2	E	F
C = 111110	1.2	2.8	E	F

Benchmark

Compressor		Calgary	Seconds	Memory	Date	Author	Major changes
P5		992,902	6.1*	256 KB	2000	Mahoney	64K x 1 neural network
P6		841,717	7.4*	16 MB	2000	Mahoney	1M neurons
P12		831,341	7.5*	16 MB	2000	Mahoney	Word context model
PAQ1		716,704	13*	48 MB	2002	Mahoney	Linear mixing with fixed weights
PAQ2		702,382	18*	48 MB	May 2003	Osnach	SSE
PAQ3		696,616	15*	48 MB	Sep 2003	Mahoney	Interpolated SSE
PAQ3N		684,580	30*	80 MB	Oct 2003	Osnach	Sparse models
PAQ4		672,134	43*	84 MB	Nov 2003	Mahoney	Adaptive mixer weights, record models
PAQ5		661,811	70*	186 MB	Dec 2003	Mahoney	Models for text, audio, images, runs, 2 mixers
PAQ6 -6		648,892	99*	202 MB	Jan 2004	Mahoney	Models for PIC, x86
PAQAR 4.0 -6		604,254	408*	230 MB	Jul 2004	Rhatushnyak	Many mixers and SSE chains
PAQ7 -5		611,684	142*	525 MB	Dec 2005	Mahoney	Logistic mixing, image models
PAQ8A -4		610,624	152*	115 MB	Jan 2006	Mahoney	E8E9 preprocessor
PASQDA 4.4 -7	D	571,011	283*	470 MB	Jan 2006	Skibinski	PAQ7 + external dictionary
PAQAR 4.5 -5	D	570,374	299*	191 MB	Feb 2006	Rhatushnyak	PAQAR + external dictionary
PAQ8F -6		605,650	161*	435 MB	Feb 2006	Mahoney	Bytewise indirect model, memory tuning
PAQ8L -6		595,586	368	435 MB	Mar 2007	Mahoney	DMC model
PAQ8PX_V67 -6		598,969	469	421 MB	Jan 2010	Ondrus	Improved JPEG, TIFF, BMP, WAV models

Quellen

Data Compression Explained http://mattmahoney.net/dc/dce.html

http://webhome.cs.uvic.ca/~nigelh/Publications/DMC.pdf

http://mattmahoney.net/dc/mmahoney00.pdf

http://www.mattmahoney.net/dc/mmahoney00.pdf

http://mattmahoney.net/dc/dce.html