Word2Bits - Quantized Word Vectors

Max Lam

Problem

Word vectors take lots of storage and memory

500 dimensions * 4 bytes * 4M words = 8GB

Approach: Word2Bits

Extend Word2Vec to train high quality <u>quantized</u> word vectors

(e.g: each parameter is either -1 or +1)

<u>Loss Function</u>

$$J(u_o, \hat{v}_c) = -\,log(\sigma(\hat{u}_o^T \hat{v}_c)) - \sum_{i=1}^{k} log(\sigma(-\hat{u}_i^T \hat{v}_c))$$

$$\hat{u}_i = Q_{bitlevel}(u_i)$$

$$\hat{v}_c = \sum_{-w+o \le i \le w+o} Q_{bitlevel}(v_i)$$

<u>Quantization Function</u>

$$Q_1(x) = \begin{cases} \frac{1}{3} & x \ge 0 \\ -\frac{1}{3} & x < 0 \end{cases}$$

$$Q_2(x) = \begin{cases} \frac{3}{4} & x > \frac{1}{2} \\ \frac{1}{4} & 0 \le x \le \frac{1}{2} \\ -\frac{1}{4} & -\frac{1}{2} \le x < 0 \\ -\frac{3}{4} & x < -\frac{1}{2} \end{cases}$$

Intrinsic Experiments: Word Analogy and Similarity

Dataset: English Wikipedia 2017

Word Vector Type	Bits per parameter	Dimension	WordSim Similarity	WordSim Relatedness	MEN	M. Turk	Rare Words	SimLex	Google Add / Mul	MSR Add / Mul
Full Precision	32	200	.740	.567	.716	.635	.403	.317	.706/.702	.447/.447
	32	400	.735	.533	.720	.623	.408	.335	**.722**/.734	**.473**/.486
	32	800	.726	.500	.713	.615	.395	.337	.719/.735	.471/**.489**
	32	1000	.741	.529	.745	.617	.400	.358	.664/.675	.423/.434
Thresholded	T1	200	.692	.480	.668	.575	.347	.288	.371/.369	.186/.182
	T1	400	.677	.446	.686	.581	.369	.321	.533/.540	.286/.292
	T1	800	.728	.494	.692	.576	.383	.338	.599/.609	.333/.346
	T1	1000	.689	.504	.694	.551	.358	.342	.521/.520	.303/.305
Quantized	1	800	.772	.653	.746	.612	.417	.355	.619/.660	.395/.390
	1	1000	.768	**.677**	.756	.638	**.425**	.372	.650/.660	.371/.408
	1	1200	**.781**	.628	.765	**.643**	.415	.379	.659/.692	.391/.429
	2	400	.752	.604	.741	.616	.417	.373	.666/.690	.396/.418
	2	800	.776	.634	**.767**	.642	.390	**.403**	.710/.739	.418/.460
	2	1000	.752	.594	.764	.602	.362	.387	.720/**.750**	.436/.482

- Quantized vectors better on similarity, worse on analogy

Intrinsic Experiments: DrQA SQuAD

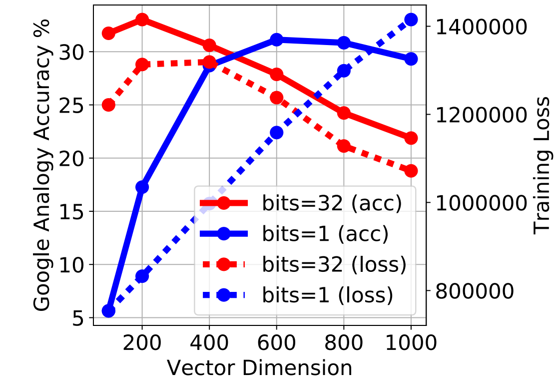
Dataset: English Wikipedia 2017

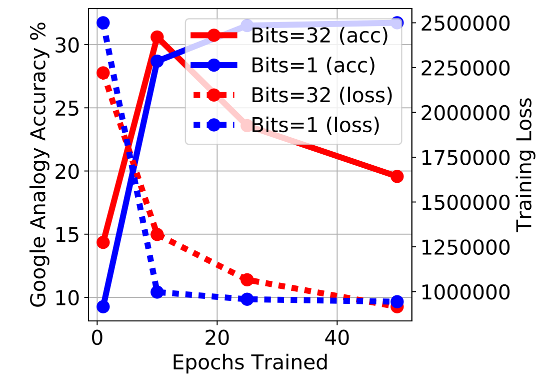
Word Vector Type	Bits per parameter	Dimension	Bytes per word	F1
Full Precision	32	200	800	75.25
	32	400	1600	75.28
	32	800	3200	75.31
	32	1000	4000	9.99
Quantized	1	800	100	76.64
	1	1000	125	76.84
	1	1200	150	76.50
	2	400	100	**77.04**
	2	800	200	76.12
	2	1000	250	75.66

- **Quantized vectors outperform full precision vectors on DRQA**
- **<u>8x-16x</u> less storage/memory than full precision**

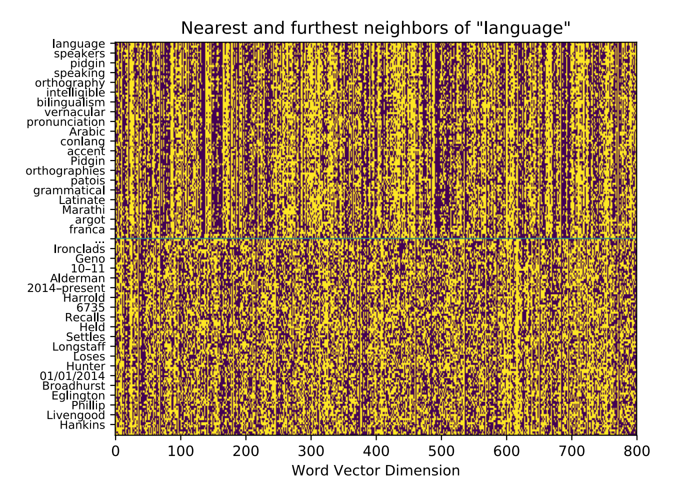
Word2Bits & Regularization

Dataset: 100MB of Wikipedia





Word2Bits Visualization



Pre-trained vectors at: https://github.com/agnusmaximus/Word2Bits/