Compressing Word Embeddings

Martin Andrews
Martin.Andrews@RedCatLabs.com

red cat labs

Summary

Take a standard word embedding, and compress it while maintaining quality

Compression is done in two ways :
- Dense element-wise bit reduction
- Sparse representation and encoding

Contributions

- High compression ratios : 10x
- Dense version gives 'bit budget'
- Sparse / Non-Negative compression

For the sparse embedding :
- Greater interpretability
- GPU-friendly implementation

Open Source Code and Data

Available via GitHub :
- https://github.com/mdda/compressing-word-embeddings

Word Embeddings

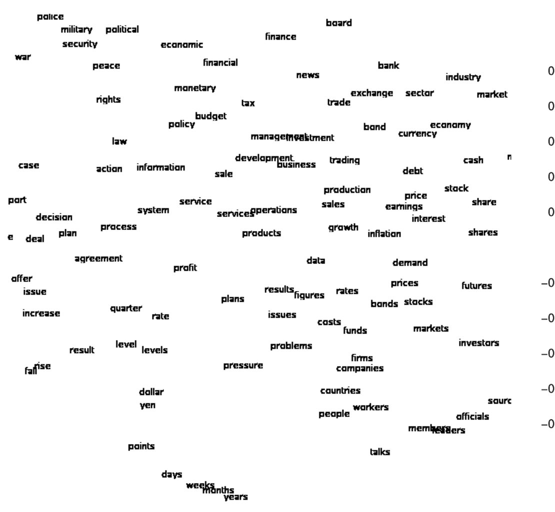
"A word is characterized by the company it keeps" - Firth 1957

Word 'embedding' :
- ~300d vector for each word
- Similar embedding ⇔ similar context
- Learn from ~6bn word corpus

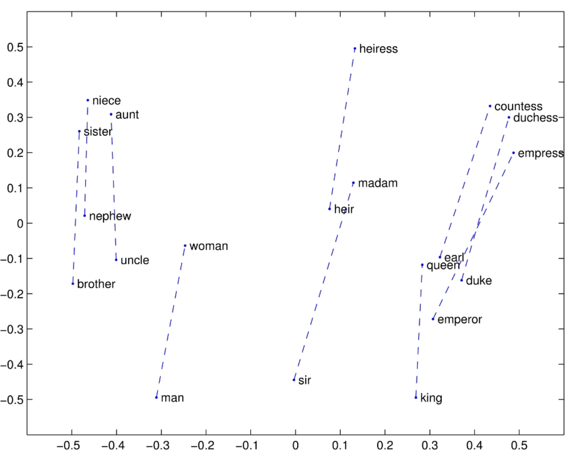
Standard embeddings :
- GloVe, word2vec, etc...

Problems :
- Size
- Interpretability

Testing an Embedding



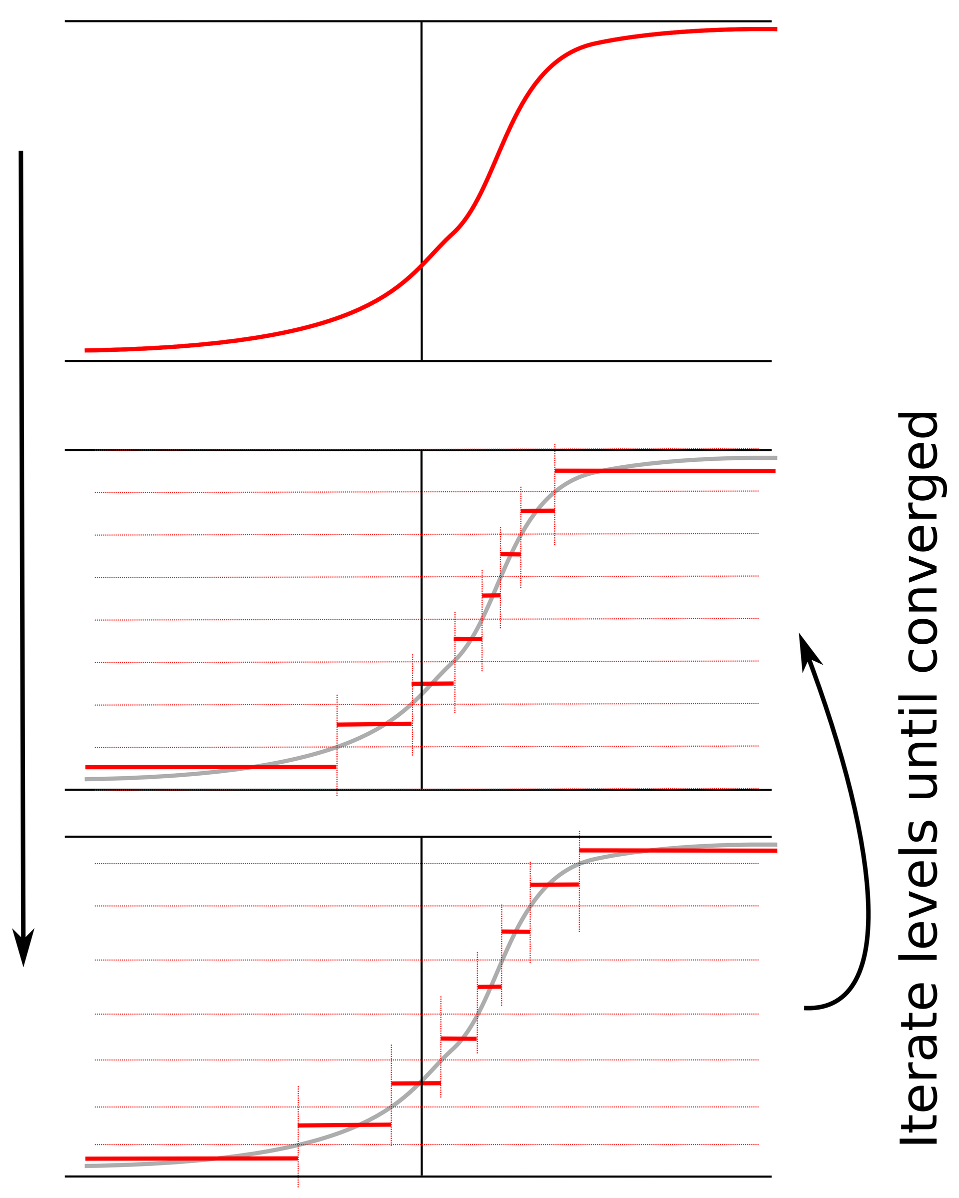
Word Similarity



Word Analogy

Dense Compression

Lloyd's Algorithm to determine "best" quantisation levels for each element



Results

- Word Similarity loss < 1% with 8 levels
- Each element is 3 bits
- 'Bit Budget' = 900 bits for embedding

Sparse Embeddings

Cognitive arguments for sparsity:
- Wider range of features desirable
- Low number of attributes important
- Mainly positive attributes stored

Sparsity also allows for compression:
- Don't store zero elements
- Store addresses
- Use to reconstruct or use 'raw'

Winner-take-all Autoencoders

Desire a sparse representation :
- Add 'λL_1' sparsity-preference?
- No : Impose α% sparsity directly
- Zero all elements outside 'Top-α'

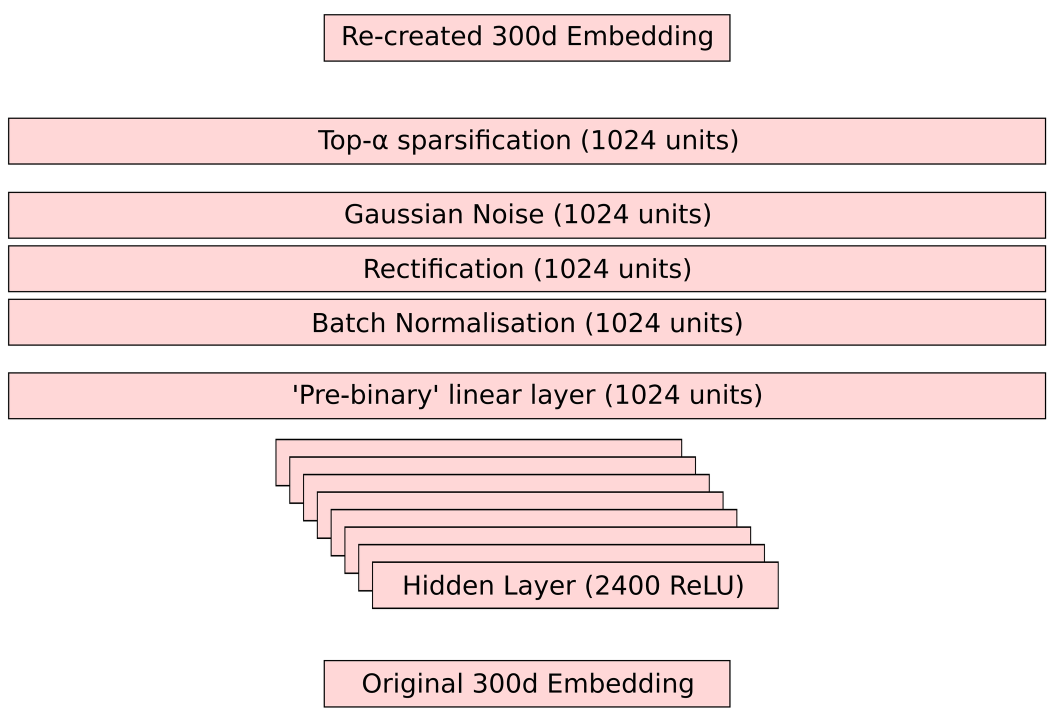
Recasting for GPU

Exact Top-α algorithm requires a sort ...
... which is GPU-unfriendly

Instead, perform a search for Top-α:
- Guess a hurdle, compute approx-α
- Iterate for a fixed number of steps
- Binary section beats other methods
- Overall : 39x speed-up vs CPU/GPU

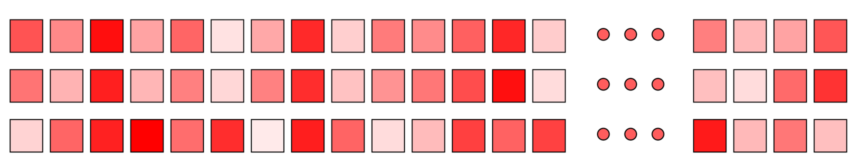
Auto Encoder Scheme

- Match re-created encoding to original
- Dynamic sparsity pressure
- Sparse embedding is a by-product

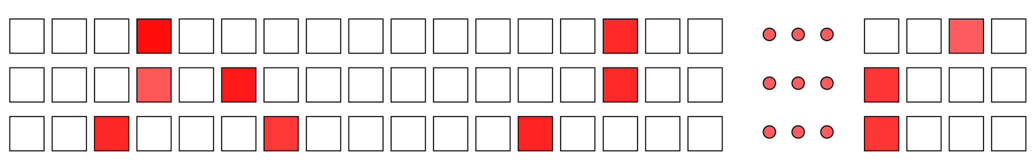


Change of Representation

Dense 300d :

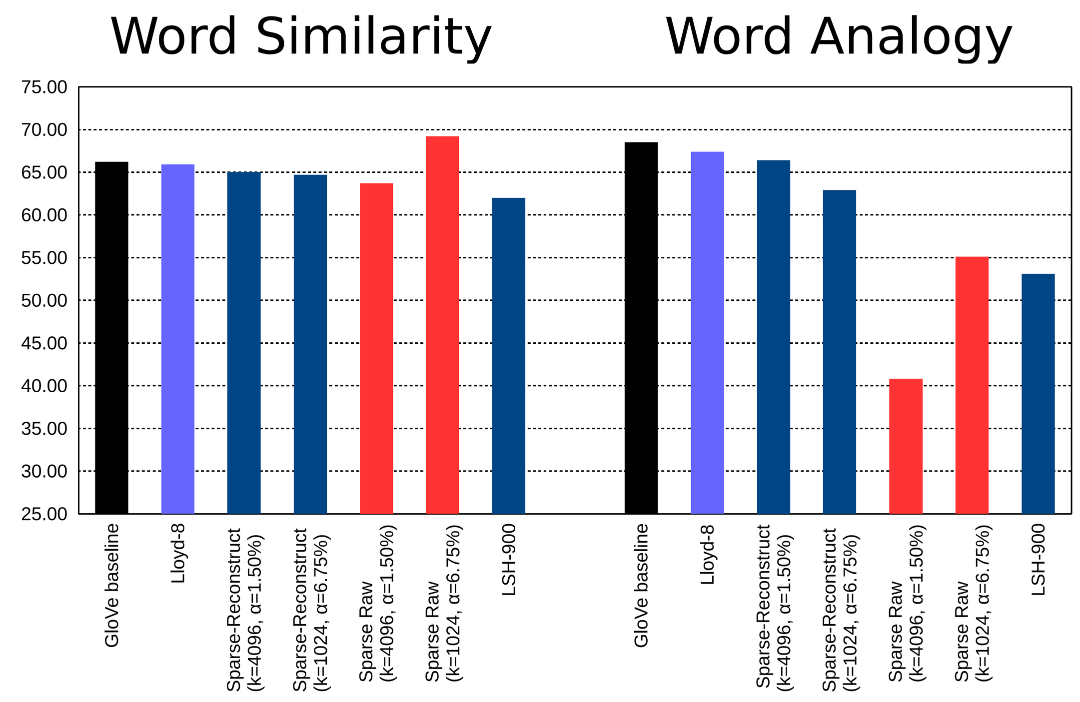


... becomes Sparse 1024d:



- Each sparse element is 10+3bits
- Sort-order can also encode intensity
- Parameters chosen ⇒ 900bits total

Sample Results



Dense vs Sparse

Representation of "Motorbike"
- Top words in each of first 7 dimensions

GloVe baseline (300d)
- lb., four-bladed, propeller, propellers, two-bladed, ...
- passerine, 1975-79, rennae, fyrstenberg, edw, coots, ...
- bancboston, oshiomhole, 30-sept, holmer, smithee, recon, ...
- http://www.nytimes.com, (888), receival, jamiat, shyi, ...
- subjunctive, purley, 11-july, broaddus, muharram, ebit, ...
- proximus, pattani, 31-feb, wgc, 30-nov, crossgen, 2,631, ...
- officership, tvcolumn, integrable, salticidae, o-157, ...

Sparse (k=1024, α=6.75%)
- vehicles, vehicle, cars, scrappage, car, 4x4, armored, ...
- prix, races, race, laps, vettel, rikknen, sprint, ...
- ski, coal, gas, taxicab, nuclear, wine, cellphone, ...
- kool, electrons, pulpit, efta, gallen, gasol, birdman, ...
- eric, anglo, tornadoes, rt, asteroids, dera, rim, ...
- wear, trousers, dresses, jeans, wearing, worn, pants, ...
- stabbed, kercher, 16-year-old, 15-year-old, 18-year-old, ...

Key References

"Learning effective and interpretable semantic models using NNSE" - Murphy et al. (2012)
"A winner-take-all method for training sparse CAE" - Makhzani & Frey (2014)
"Glove: Global vectors for word representation" - Pennington et al (2014)

Contact

Martin.Andrews@RedCatLabs.com
+65 8585 1750
http://RedCatLabs.com/