Low-Rank Tensors for Verbs in Compositional Distributional Semantics

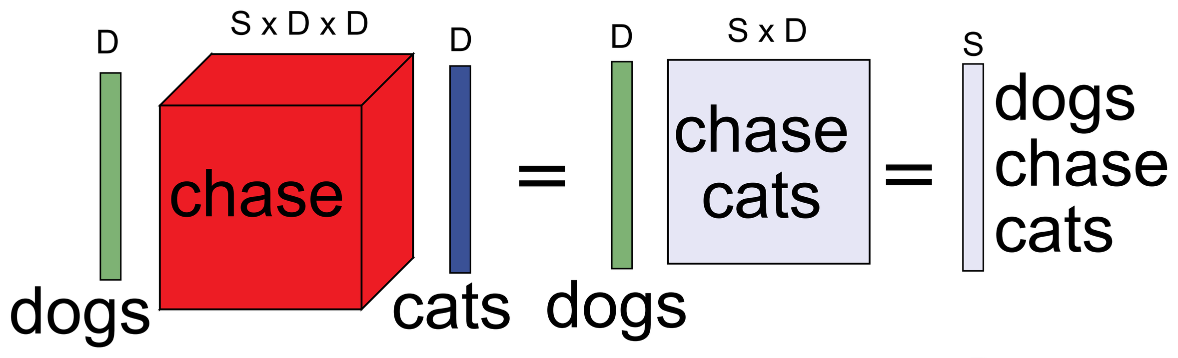
Daniel Fried Tamara Polajnar Stephen Clark
Computer Laboratory, University of Cambridge
{df345, tp366, sc609}@cam.ac.uk

Tensor Models for Verbs

Syntax-driven composition (Coecke et al., 2011)

- Vectors for atomic CCG types (noun, noun phrase, sentence)
- Syntactic functions represented by matrices or higher-order tensors

dogs	chase	cats
NP	$(S\backslash NP)/NP$	NP
	$S\backslash NP$ $>$	
S $<$		



- Third-order tensor $\mathcal{V}$ for each transitive verb. Maps subject $\mathbf{s} \in \mathbb{R}^D$ and object $\mathbf{o} \in \mathbb{R}^D$ to a composed vector $(\mathcal{V}\mathbf{o})\mathbf{s} \in \mathbb{R}^S$:

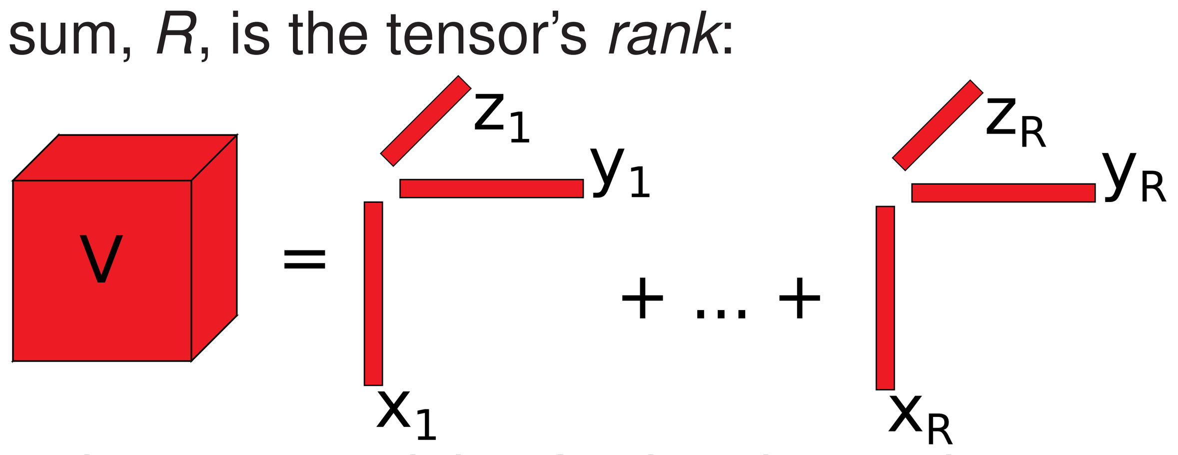
$$[(\mathcal{V}\mathbf{o})\mathbf{s}]_i = \sum_{jk} \mathcal{V}_{ijk}\mathbf{o}_j\mathbf{s}_k$$

- Full tensors require prohibitively many parameters: $S \times D \times D$ for each verb

Low-rank approximations to tensors

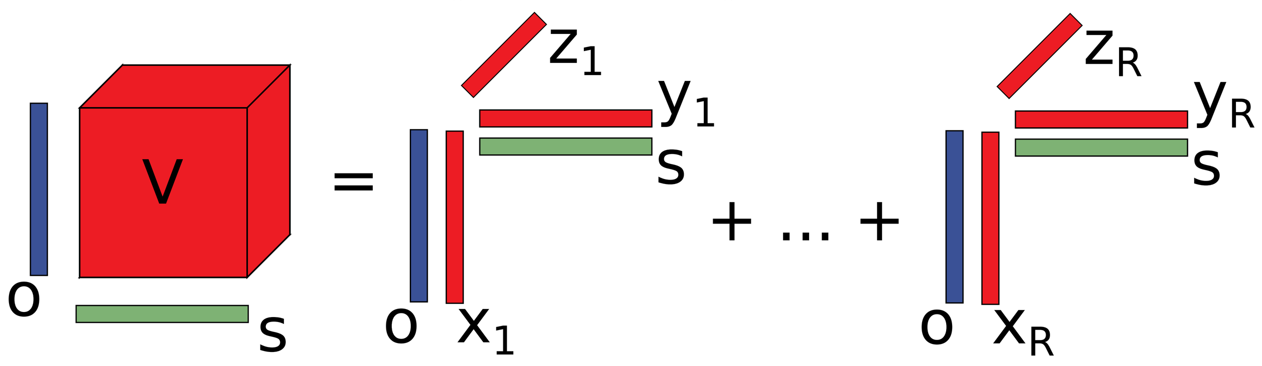
- *CP decomposition* represents a tensor as a sum of vector outer products
- The number of terms in the sum, R, is the tensor's *rank*:

$$\mathcal{V} = \sum_{r=1}^{R} \mathbf{x}_r \otimes \mathbf{y}_r \otimes \mathbf{z}_r$$



- The tensor's action on vectors is a sum, weighted using dot products:

$$(\mathcal{V}\mathbf{o})\mathbf{s} = \sum_{r=1}^{R} (\mathbf{x}_r \cdot \mathbf{o})(\mathbf{y}_r \cdot \mathbf{s})\mathbf{z}_r$$



- By limiting R, we force a low-rank approximation to $\mathcal{V}$ (Lei et al. 2014, 2015)

Learning full and low-rank tensors

- Create distributional vectors for nouns $\mathbf{s}$ and $\mathbf{o}$ and SVO triples, $\mathbf{t}_{sVo}$
- Learn $\mathcal{V}$ by multi-linear regression: minimize squared residual between predicted SVO vector $(\mathcal{V}\mathbf{o})\mathbf{s}$ and actual SVO vector $\mathbf{t}_{sVo}$

$$L(\mathcal{V}) = \frac{1}{M_V} \sum_{i=1}^{M_V} ||(\mathcal{V}\mathbf{o}^{(i)})\mathbf{s}^{(i)} - \mathbf{t}_{sVo}^{(i)}||_2^2$$

- Mini-batched ADADELTA optimization. Optimize $\mathcal{V}$ directly for full tensors, or alternating optimization of $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ for low-rank tensors.

Distributional Vectors

- Corpus: October 2013 Wikipedia download
- SVO triples determined using verb dependencies
- Distributional context: most frequent lemmatized words within sentence boundaries

Count vectors

- Count co-occurrences between nouns (or SVO triples) and context words
- Re-weight and filter rare contexts, reduce to 100 dimensions with SVD

Prediction vectors

- Use skip-gram model (Mikolov et al.) to predict context words from nouns and SVO triples
- Hierarchical sampling, 100-dimensional vectors

Evaluation Tasks

- Compose the learned $\mathcal{V}$, $\mathbf{s}$, $\mathbf{o}$ into vectors for new SVO triples
- Compare to human rankings of SVO triple similarity (Spearman correlation)

Verb disambiguation

- Grefenstette & Sadrzadeh, 2011
- Same subject and object, verb changes to highlight word senses
- *sim*(report draw attention, report attract attention) $>$ *sim*(report draw attention, report depict attention)

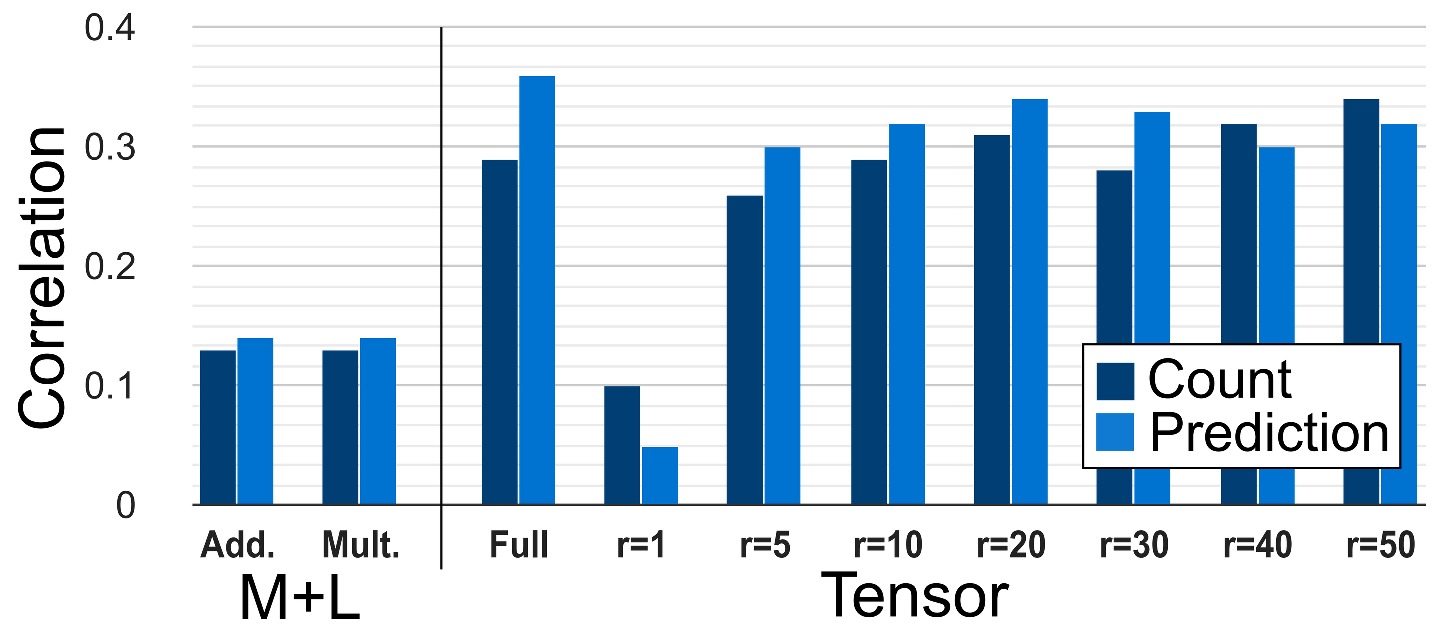
Sentence similarity

- Kartsaklis & Sadrzadeh, 2013
- Subject, object, and verb all vary:
- *sim*(programme offer support, service provide help) $>$ *sim*(author write book, delegate buy land)

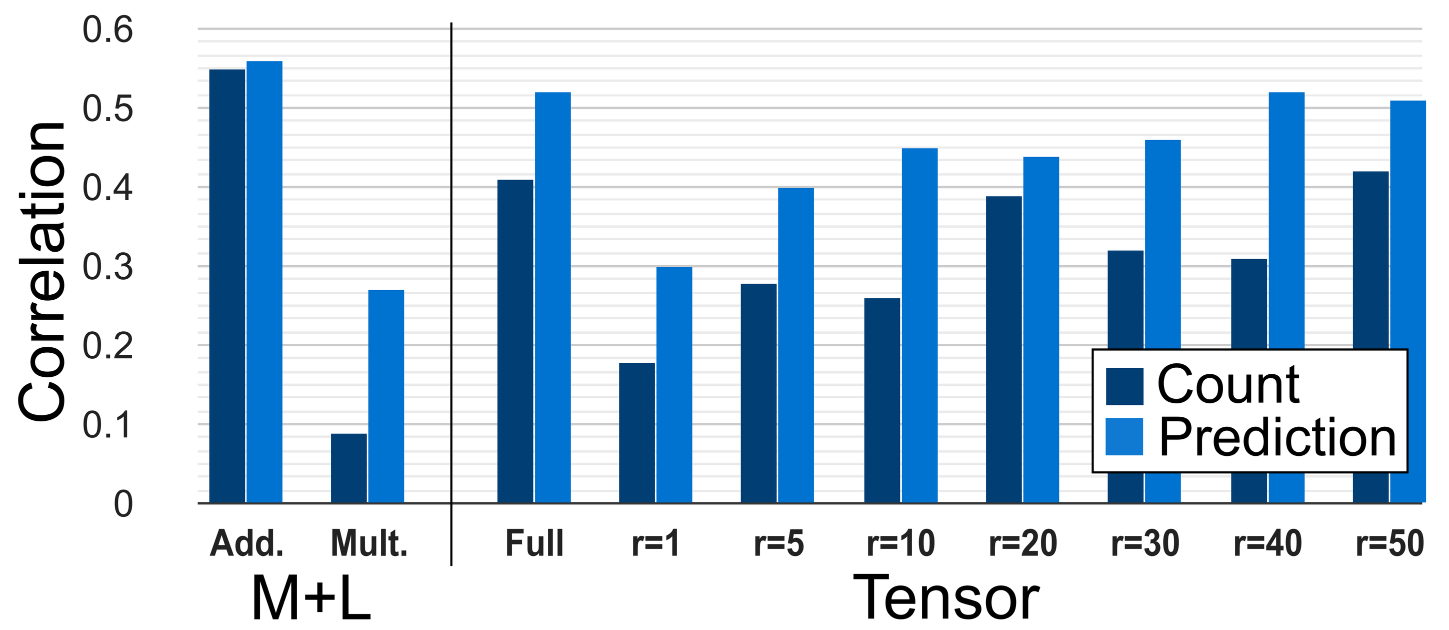
Results

- Low-rank tensor performance generally increases with rank
- Best low-rank comparable to full tensors for prediction vectors; slightly higher for count vectors
- Tensor methods better than element-wise vector addition or multiplication (Mitchell & Lapata, 2008) for verb disambiguation, but worse than addition for sentence similarity

Verb disambiguation



Sentence similarity



- **Large reduction in parameters: $R = 50$ requires 15,000 parameters per verb; full tensor has 1 million**
- Efficiently trainable with backpropagation: never have to store full tensors, and 2x faster to train per verb

UNIVERSITY OF CAMBRIDGE

Daniel Fried is supported by a Churchill Scholarship. Tamara Polajnar is supported by ERC Starting Grant DisCoTex (306920). Stephen Clark is supported by ERC Starting Grant DisCoTex (306920) and EPSRC grant EP/I037512/1.