On Recovering Analogies as Parallel Lines and Contrastive Learning Methods

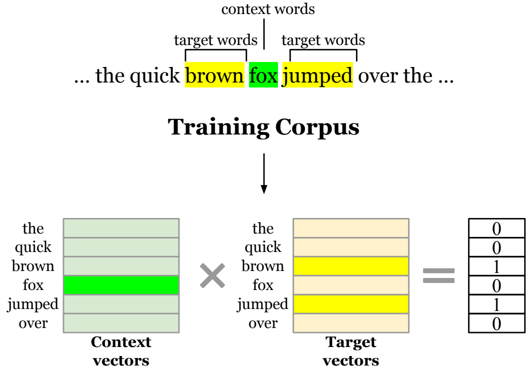
Narutatsu (Edward) Ri[1] Fei-Tzin Lee[1] Nakul Verma[1]

[1] Department of Computer Science
Columbia University

Word Embeddings and Analogies

Popular static word embedding models are based on the *distributional hypothesis*: words that occur in the same contexts tend to have similar meanings [1]

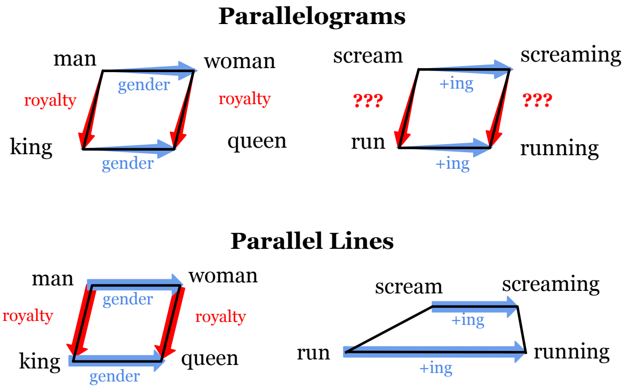
Example: *word2vec*



Two matrices are trained to recover co-occurrence statistics with inner products. Context vectors are used as word embeddings, target vectors are discarded.

Pheonomenon: For all models, analogies are implicitly learned as *some* structure in the embedding space

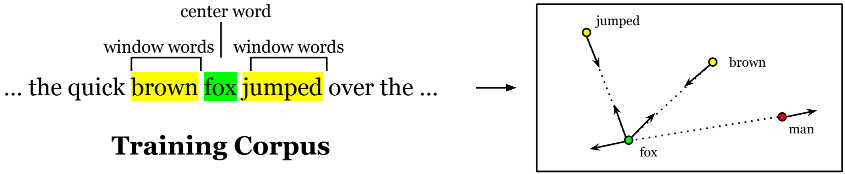
Previous consensus: Parallelograms [2, 3]
Recent works: Parallel lines [4, 5, 6]



Question: How does this happen? What is the core mechanism?
Answer: Unclear, and existing theoretical works are scarce [7]

Our work studies the **underlying machinery for recovering analogies as parallel lines**.

Idea: Pull word vectors that co-occur close together while pushing others away, and keep one vector for each word



Objective:

$$\mathcal{L}_{\mathrm{CWM}}(V) = \sum_{c\in W} \sum_{w\in W} \#(c,w) \cdot \sum_{w' \in D_{c,w}} \Big[m - \underbrace{\hat{v}_c \cdot \hat{v}_w}_{\text{pull}} + \underbrace{\hat{v}_c \cdot \hat{v}_{w'}}_{\text{push}} \Big]_+$$

Explanation:

Difference between $\hat{v}_c \cdot \hat{v}_w$ and $\hat{v}_c \cdot \hat{v}_{w'}$ encourages the *angle* between v_c and v_w to be smaller than between v_c and $v_{w'}$ by at least a margin of m.

Popular Word Embeddings and Push-Pull

Existing methods can be reformulated as push-pull.

word2vec: Vectors for co-occurring words are pulled towards each other, while being pushed away from the mean of all other word vectors:

$$v_c^{\text{new}} = v_c^{\text{old}} + \underbrace{\left(1 - \frac{e^{v_w^\intercal u_{c'}}}{\sum_{w' \in W} e^{v_w^\intercal u_{w'}}}\right) v_w}_{\text{pull}} - \underbrace{\mathbb{E}_{w' \sim W}[v_{w'}]}_{\text{push}} + \text{additional terms}$$

GloVe: Vectors for co-occurring words are pulled towards a common vector, while other words are pushed away from the same vector:

$$\text{pull} \begin{cases} v_c^{\text{new}} &= v_c^{\text{old}} + g(c, c')u_{c'} \\ v_w^{\text{new}} &= v_w^{\text{old}} + g(w, c')u_{c'} \end{cases}$$

$$\text{push} \begin{cases} v_{w'}^{\text{new}} &= v_{w'}^{\text{old}} - g(w', c')u_{c'}, \end{cases}$$

Claim

The word vectors $v_c \in V$ that minimize the global objective is:

$$v_c = \rho_c \left(\sum_{w \in W} \left(\frac{\#(c, w)}{\#(c)} \hat{v}_w \right) - \mathop{\mathbb{E}}_{w' \sim U(W)} \left[\hat{v}_{w'} \right] \right), \tag{1}$$

where $\rho_c \propto \#(c)$.

Theorem

If the word vectors satisfy Eq. (1)*, for any quadruple of words* $a, b, c, d \in W$*, if the co-occurrence statistics satisfy the condition:*

$$\exists \zeta \in \mathbb{R}, \forall w \in W : \left(\frac{\#(a,w)}{\#(a)} - \frac{\#(b,w)}{\#(b)} \right) \Big/ \left(\frac{\#(c,w)}{\#(c)} - \frac{\#(d,w)}{\#(d)} \right) := \zeta, \quad (2)$$

then the corresponding word vectors satisfy the property:

$$\hat{v}_a - \hat{v}_b = \zeta \left(\hat{v}_c - \hat{v}_d \right).$$

Interpretation:

If word co-occurrence statistics follow Theorem 2, then the quadruple will form parallel lines.

Significance:

Given a corpus, one can predict which words will form parallel lines *a priori* to training!

Relation between Co-occurrence and Analogies

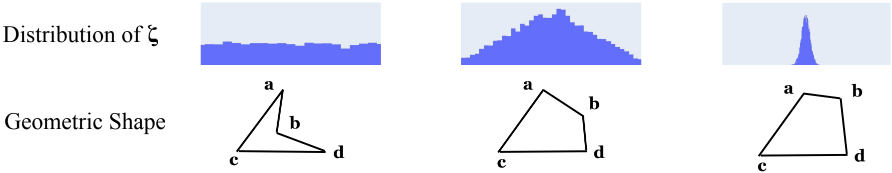
Value of ζ and Geometry

In Eq. (2), a ζ_w can be calculated for each word $w \in W$ for fixed a, b, c, d:

$$\left(\frac{\#(a,w)}{\#(a)} - \frac{\#(b,w)}{\#(b)}\right) \Big/ \left(\frac{\#(c,w)}{\#(c)} - \frac{\#(d,w)}{\#(d)}\right) := \zeta_w$$

$$\Rightarrow \hat{v}_{a,w} - \hat{v}_{b,w} = \zeta_w \left(\hat{v}_{c,w} - \hat{v}_{d,w}\right)$$

Remark 1: The concentration of the the distribution of ζ_w describes how parallel the quadruples' lines will be:



There exists analogies that are vague/ambiguous.

Examples:

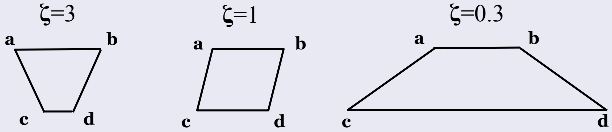
sun : red = sea : blue
sun : yellow = sea : blue
sun : orange = sea : blue
...

run : running = walk : walking
flee : fled = grow : grew
Paris : France = Tokyo : Japan
...

For left pairs, formation of parallelograms/trapezoids for all quadruples is difficult.

Empirically, we want to observe low concentration of ζ_w values for ambiguous analogies, and high concentration for clear analogies.

Remark 2: The value of ζ determines the geometric shape of the quadruple:



$$\hat{v}_a - \hat{v}_b = \zeta\left(\hat{v}_c - \hat{v}_d\right)$$

When $\zeta = 1$: Parallelogram
When $\zeta \neq 1$: Trapezoid

Empirically, for analogy pairs, we want to observe better parallelogram recovery for $\zeta = 1$, and better trapezoid recovery when ζ_w is concentrated.

Results

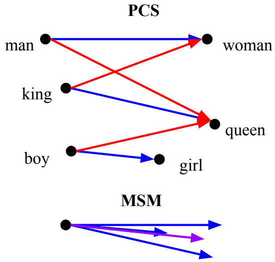
Metrics [8]:

P: True Analogy Pairs
N: Imposter Analogy Pairs

PCS (Pairing Consistency Score):
Measures *relative* offset alignment

MSM (Mean Similarity Measure):
Measures *absolute* offset alignment

There are degenerate configurations which perform well on one but not both.



Performance on BATS Dataset:

	Analogies		Training	
Model	**PCS**	**MSM**	**Time (hrs)**	**Speedup**
CWM	**0.677**	**0.469**	**0.59**	**49**$\times$
SGNS	0.675	0.433	29.27	1$\times$
GloVe	0.667	0.423	30.71	0.91$\times$

CWM performs competitively while achieving dramatic train time speedup (49 times faster than *word2vec*!)

Existence and Effect of ζ and Analogies

Remark 1 Verification:

We extract analogy pairs where ζ_w is concentrated and not concentrated.

Samples where ζ is *highly* concentrated:

- improve : improves = create: creates
- enable : enables = allow : allows
- provide : provides = create : creates
- prevent : prevents = protect : protects
- prevent: preventing = avoid : avoiding
- avoid : avoiding = ensure : ensuring

Samples where ζ is *poorly* concentrated:

- mouse : rodent = beetle : insect
- beetle : insect = squirrel : rodent
- beetle : insect = beaver : rodent
- wall : cement = clothing : fabric
- jewelry : bracelet = poem : haiku
- porcupine : rodent = beetle : insect

Result:

Analogies where relationship is *precise* exhibit high concentration, while bad quality analogies (vague relationship, impossible pairs) exhibit poor concentration

Existence and Effect of ζ and Analogies

Remark 2 Verification:

Extract all quadruples where ζ exists, and separate into when $\zeta \approx 1$ and when $\zeta \not\approx 1$. Take kNN of calculated answer and check whether correct answer is among k nearest neighbors.

Compare parallelogram recovery between all analogies and selected analogies:

Structure	**Subset**	$k = 1$	$k = 5$
Parallel Lines	$\zeta \not\approx 1$	0.80 (619/774)	0.86 (667/774)
Parallelograms	$\zeta \approx 1$	0.65 (137/210)	0.87 (183/210)
	All Analogies	0.21 (12549/59776)	0.27 (16121/59776)

Result:

Trapezoids are very well-recovered for subset of analogies where ζ exists.

Parallelograms are far better recovered when ζ exists and $\zeta = 1$ compared to all analogies in dataset.

Summary

We showed a contrastive learning objective is sufficient in recovering analogies as parallel lines.

- Push-pull method can be mathematically shown to implicitly recover analogies
- Geometry of embeddings can be determined *a priori* training on corpus from co-occurrence statistics
- Analogy pairs tend to follow a specific co-occurrence pattern, while other word pairs do not

Conclusion & Discussion

Future Directions

Full theoretical analysis of contrastive learning approach on sequential data generated with synthetic model

- Issue with natural language: large noise, and analogies are a subjective construct. Can we polish the relationship in Theorem 2 and analyze the optimization procedure of how push-pull exactly leads to the formation of parallel lines?
- The empirical results we show merely indicate sufficiency of push-pull for implicitly encoding analogies as parallel lines. Can we show necessity?
- Sample complexity for recovering analogies: bounds on no. of samples required to learn analogies as parallel lines?

J. Firth.
A synopsis of linguistic theory 1930-1955.
In *Studies in Linguistic Analysis*. Philological Society, Oxford, 1957. reprinted in Palmer, F. (ed. 1968) Selected Papers of J. R. Firth, Longman, Harlow.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean.
Efficient estimation of word representations in vector space.
In *International Conference on Learning Representations*, 2013.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig.
Linguistic regularities in continuous space word representations.
In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

Natalie Schluter.
The word analogy testing caveat.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 242–246, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Tal Linzen.

Issues in evaluating semantic spaces using word analogies.
In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18, Berlin, Germany, August 2016. Association for Computational Linguistics.

Louis Fournier and Ewan Dunbar.
Paraphrases do not explain word analogies, 2021.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski.
A latent variable model approach to pmi-based word embeddings, 2019.

Louis Fournier, Emmanuel Dupoux, and Ewan Dunbar.
Analogies minus analogy test: measuring regularities in word embeddings.
In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 365–375, Online, November 2020. Association for Computational Linguistics.