Efficient Estimation of Word Representations in Vector Space

Ivan N., Ryan L., Chaitanya M.

UC Riverside

January 23, 2024

Background

What is a word vector?

```
vectorize("hello") -> [0, 0.23, 0, .10, 0.5, 0.75, ...]
```

Problem(s):

- Current methods **do not truly capture similarity** between words.
- Computational capabilities only allow for **limited size** of data.
- In-domain data is limited.

Background

Goal(s) of the Paper:

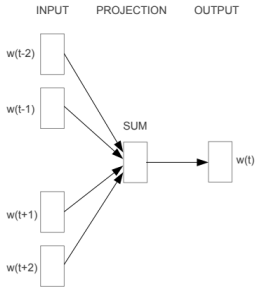
- Capture similarity between words in a better way.
- More efficiently compute word vectors.
- Test performance of this model.

Methodology

Two proposed methods:

1. Continuous Bag-of-Words Model
2. Continuous Skip-gram Model

Continuous Bag-of-Words



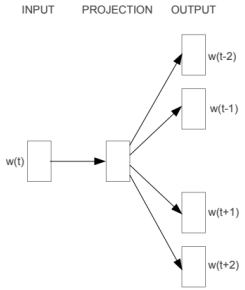
- Complexity:
  $Q = N \times D + D \times log_2(V)$

Methodology

Two proposed methods:

1. Continuous Bag-of-Words Model
2. Continuous Skip-gram Model

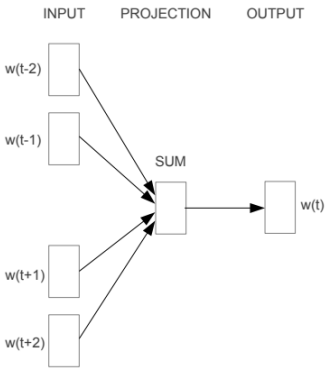
Continuous Skip-Gram

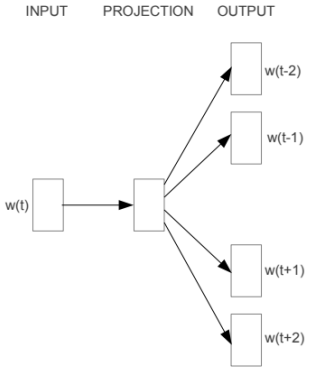


- Complexity:
  $Q = C \times (D + D \times log_2(V))$

Methodology

Both of these models are considered Word2Vec





Results

Summarized Results

- Improved performance over other state-of-the-art methods for capturing word relationship information.
- A much more efficient method for generating word vectors.

Table 7: *Comparison and combination of models on the Microsoft Sentence Completion Challenge.*

Architecture	Accuracy [%]
4-gram [32]	39
Average LSA similarity [32]	49
Log-bilinear model [24]	54.8
RNNLMs [19]	55.4
Skip-gram	48.0
Skip-gram + RNNLMs	**58.9**

Results

Table 1: *Examples of five types of semantic and nine types of syntactic questions in the Semantic-Syntactic Word Relationship test set.*

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Results

Table 3: *Comparison of architectures using models trained on the same data, with 640-dimensional word vectors. The accuracies are reported on our Semantic-Syntactic Word Relationship test set, and on the syntactic relationship test set of [20]*

Model Architecture	Semantic-Syntactic Word Relationship test set		MSR Word Relatedness Test Set [20]
	Semantic Accuracy [%]	Syntactic Accuracy [%]	
RNNLM	9	36	35
NNLM	23	53	47
CBOW	24	64	61
Skip-gram	55	59	56

That's it! Any questions?

References

1. Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
2. T. Mikolov, W.T. Yih, G. Zweig. Linguistic Regularities in Continuous Space Word Representations. NAACL HLT 2013.
3. G. Zweig, C.J.C. Burges. The Microsoft Research Sentence Completion Challenge, Microsoft Research Technical Report MSR-TR-2011-129, 2011.