Incremental Skip-gram Model with Negative Sampling

Nobuhiro Kaji and Hayato Kobayashi
Yahoo Japan Corporation

Summary of This Study

- Existing methods of word embedding (e.g., skip-gram model and GloVe) cannot perform incremental model update when new training data is provided
- We propose an incremental extension of skip-gram model with negative sampling (SGNS) and demonstrate its effectiveness from both theoretical and empirical perspectives

Existing Algorithm for SGNS Training: Two-pass Algorithm

Training data (sequence of words): $w_1, w_2, \cdots, w_i, \cdots, w_n$

Loss function:

$$L(\theta) = -\frac{1}{n}\sum_{i=1}^{n} L_i(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{-c\le j\le c\\ j\ne 0}} \{\log\sigma(\vec{w_i}\cdot\vec{w}_{i+j}) + kE_{v\sim q(v)}[\log\sigma(-\vec{w_i}\cdot\vec{v})]\}$$

Noise distribution: $q(v) \propto (\text{number of word } v \text{ in the training data})^{\alpha}$

```
# Count word frequencies
for i = 1, 2, ⋯, n
    f(w_i) ← f(w_i) + 1

# Compute noise distribution
q(w) ∝ f(w)^α for all w

# Perform SGD update
for i = 1, 2, ⋯, n
    θ ← θ − τ ∂L_i(θ)/∂θ
```

Single-pass Incremental Algorithms

Incremental SGNS

```
for i = 1, 2, ⋯, n
    # Update word frequencies
    f(w_i) ← f(w_i) + 1

    # Compute noise distribution
    q(w) ∝ f(w)^α for all w

    # Perform SGD update
    θ ← θ − τ ∂L_i(θ)/∂θ
```

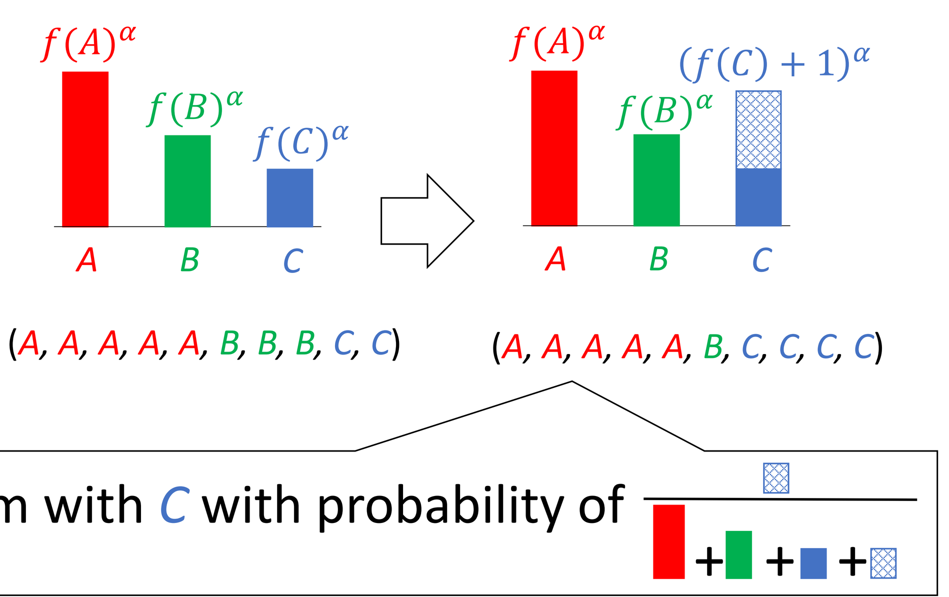
Mini-batch SGNS

```
for t = 1, 2, ⋯, T
    # Update word frequencies
    for i ∈ M_t
        f(w_i) ← f(w_i) + 1

    # Compute noise distribution
    q(w) ∝ f(w)^α for all w

    # Perform SGD update
    for i ∈ M_t
        θ ← θ − τ ∂L_i(θ)/∂θ
```

Efficient Implementation

Use Misra-Gries algorithm (Misra and Gries 82) to maintain dynamic vocabulary

Use weighted reserver sampling (Vitter, 85) to update unigram table for sampling from noise distribution



Theoretical Analysis

Lemma. Let $L(\theta)$ be the loss function of SGNS. Let also $\hat{\theta}$ be the optimal solution of incremental SGNS. Then,

$$\lim_{n\to\infty} \mathrm{E}\left[L(\hat{\theta}) - \min_{\theta} L(\theta)\right] = 0,$$

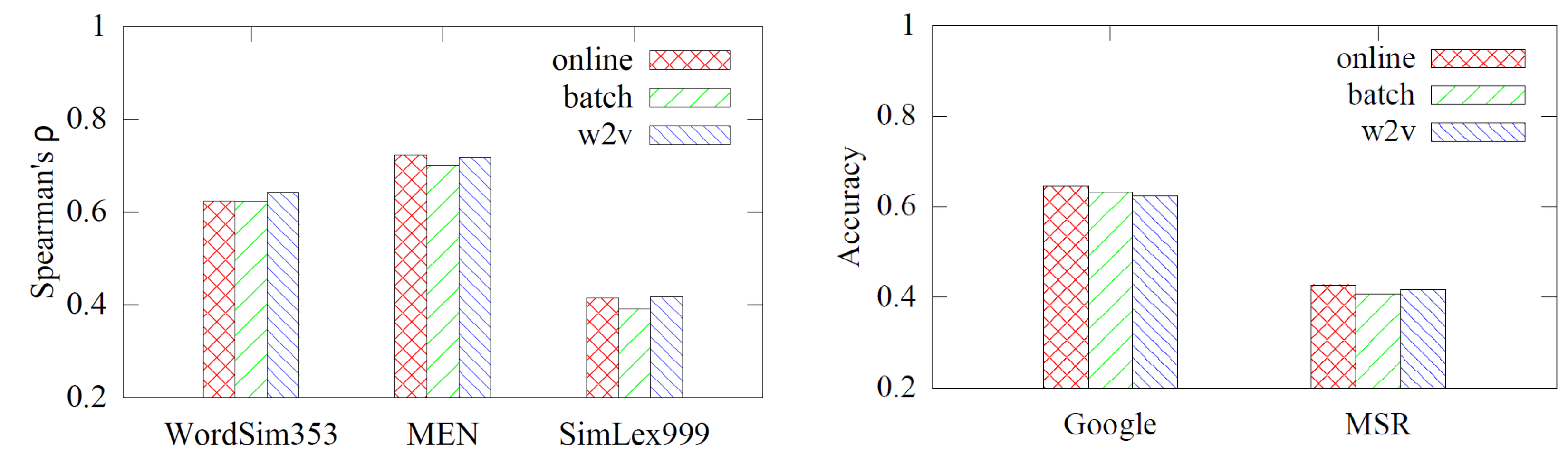
$$\lim_{n\to\infty} \mathrm{V}\left[L(\hat{\theta}) - \min_{\theta} L(\theta)\right] = 0.$$

Theorem. $L(\hat{\theta})$ converges in probability to $\min_{\theta} L(\theta)$ in the limit of $n \to \infty$:

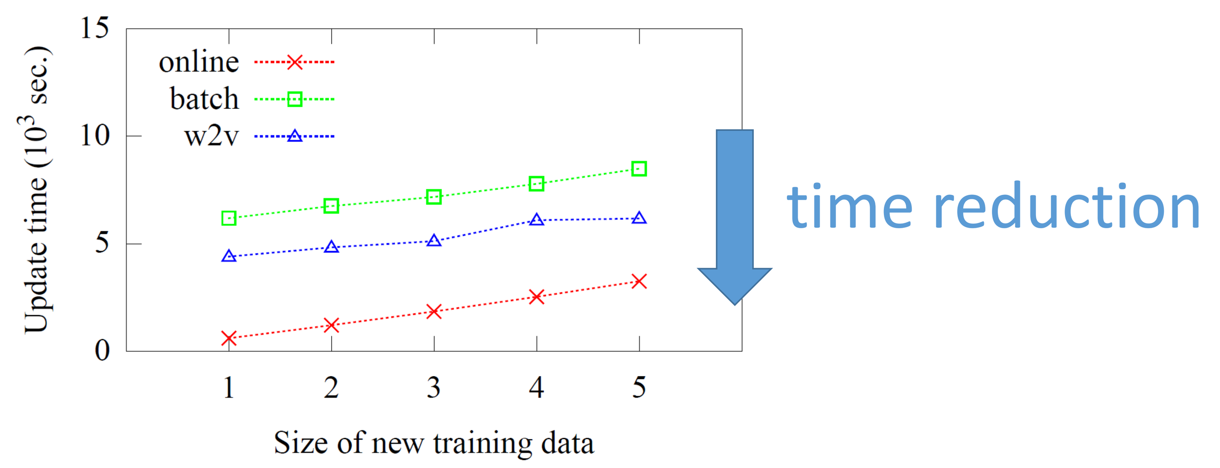
$$\forall\varepsilon > 0, \lim_{n\to\infty} \Pr\left(\left|L(\hat{\theta}) - \min_{\theta} L(\theta)\right| > \varepsilon\right) = 0.$$

Experimental Results

Word embeddings learned by incremental SGNS performed word similarity task (left) and word analogy task (right) comparatively well with the original SGNS



Incremental SGNS achieved up to 90% time reduction when updating old model on additional training data



Conclusion

- SGNS can be trained in a fully online fashion
- Both theory and experiments support the effectiveness of the new incremental algorithm