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): \( w_1, w_2, \cdots, w_l, \cdots, w_n \)

Loss function:
\[
L(\theta) = \frac{1}{n} \sum_{i=1}^{n} L_i(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left[ \log \sigma(w_i \cdot \bar{v}_i) + kE_{v \sim q(v)}[\log \sigma(-w_i \cdot \bar{v})] \right]
\]

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

Single-pass Incremental Algorithms

Incremental SGNS

for \( i = 1, 2, \cdots, n \)
# Count word frequencies
\[
f(w_i) \leftarrow f(w_i) + 1
\]
# Compute noise distribution
\[
q(w) \propto f(w)^\alpha \text{ for all } w
\]
# Perform SGD update
\[
\theta \leftarrow \theta - \tau \frac{\partial L_i(\theta)}{\partial \theta}
\]

Mini-batch SGNS

for \( t = 1, 2, \cdots, T \)
# Count word frequencies
\[
f(w_i) \leftarrow f(w_i) + 1
\]
# Compute noise distribution
\[
q(w) \propto f(w)^\alpha \text{ for all } w
\]
# Perform SGD update
\[
\theta \leftarrow \theta - \tau \frac{\partial L_i(\theta)}{\partial \theta}
\]

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

\[
\text{noise distribution: } q(w) \propto f(w)^\alpha \text{ for all } w
\]

unigram table:
\[
\text{overwrite each item with } C \text{ with probability of } 1 + \epsilon
\]

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

Theoretical Analysis

Lemma. Let \( L(\theta) \) be the loss function of SGNS. Let also \( \hat{\theta} \) be the optimal solution of incremental SGNS. Then,
\[
\lim_{n \to \infty} E \left[ L(\hat{\theta}) - \min_{\theta} L(\theta) \right] = 0,
\]
\[
\lim_{n \to \infty} \Pr \left( L(\hat{\theta}) - \min_{\theta} L(\theta) > \epsilon \right) = 0.
\]

Theorem. \( L(\hat{\theta}) \) converges in probability to \( \min_{\theta} L(\theta) \) in the limit of \( n \to \infty \):
\[
\forall \epsilon > 0, \lim_{n \to \infty} \Pr \left( L(\hat{\theta}) - \min_{\theta} L(\theta) > \epsilon \right) = 0.
\]

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