On SkipGram Word Embedding Models with Negative Sampling: Unified Framework and Impact of Noise Distributions

Ziqiao Wang[1] Yongyi Mao[1] Hongyu Guo[2] Richong Zhang[3]
[1]University of Ottawa [2]National Research Council Canada [3]Beihang University

Introduction

SkipGram word embedding models with negative sampling [1] (SGN) is an elegant family of word embedding models. In this work, we ask the following questions.

- Beyond that particular distribution, if one chooses a different noise distribution, is SGN still theoretically justified?
- Is there a general principle underlying SGN that allows us to build new embedding models?
- If so, how does the noise distribution impact the training of such models and their achievable performances?

Our Contributions

- We formalize a unified framework, referred to as "word-context classification" (WCC), for SGN-like word embedding models.
- We also provide a theoretical analysis that justifies the WCC framework. Consequently, the matrix-factorization result of [2] can also be derived from this analysis as a special case.
- The impact of noise distribution on learning word embeddings in WCC is also studied.

The WCC Framework

Binary classification problem D^+ and D^-:

- Objective: distinguish the word-context pairs drawn from $\mathbb{P}$ from those drawn from $\mathbb{Q}$;
- The classification problem is equivalent to learning the conditional distribution $p_{U|XY}(1|x,y) \coloneqq \sigma(s(x,y))$

WCC: Let $f: X \rightarrow \overline{X}$ and $g: Y \rightarrow \overline{Y}$ be two functions representing the embedding maps for words and contexts respectively. The standard cross-entropy loss for this classification problem is

$$\ell = -\sum_{(x,y)\in D^+} \log \sigma(s(x,y)) - \sum_{(x,y)\in D^-} \log \sigma(-s(x,y))$$

and the solution is

$$(f^*, g^*) \coloneqq \arg\min_{f,g} \ell(f,g)$$

Theorem 1: Suppose that $\widetilde{\mathbb{Q}}$ coverse $\widetilde{\mathbb{P}}$. Then the following holds.

1. The loss ℓ, as a function of s, is convex in s.
2. If f and g are sufficiently expressive, then there is a unique configuration s^* of s that minimizes $\ell(s)$, and the global minimizer s^* of $\ell(s)$ is given by
$$s^*(x,y) = \log\frac{\widetilde{\mathbb{P}}(x,y)}{\widetilde{\mathbb{Q}}(x,y)} + \log\frac{N^+}{N^-}$$
for every $(x,y) \epsilon X \times Y$.

Corollary 1: Let $N^+ = n$ and $N^- = kn$. Then it is possible to construct a distribution $\widehat{\mathbb{P}}$ on $X \times Y$ using f^*, g^*, k and $\mathbb{Q}$ such that for every $(x,y) \epsilon X \times Y$, $\widehat{\mathbb{P}}(x,y)$ converges to $\mathbb{P}(x,y)$ in probability as $n \rightarrow \infty$.

SGN Model

Let $\mathbb{Q}$ factorize in the following form

$$\mathbb{Q}(x,y) = \widetilde{\mathbb{P}}_X(x)\mathbb{Q}_Y(y)$$

Corollary 2: In an unconditional SGN model, the global minimizer of loss function ℓ is given by

$$s^*(x,y) = \bar{x}\cdot\bar{y} = \log\frac{\widetilde{\mathbb{P}}(x,y)}{\widetilde{\mathbb{P}}_X(x)\widetilde{\mathbb{Q}}_Y(y)} - \log k$$

As a special case when $\widetilde{\mathbb{Q}}_Y(y) = \widetilde{\mathbb{P}}_Y(y) \Longrightarrow$ "PMI"!.

Conditional SGN Model

Let $\mathbb{Q}$ factorize in the following form

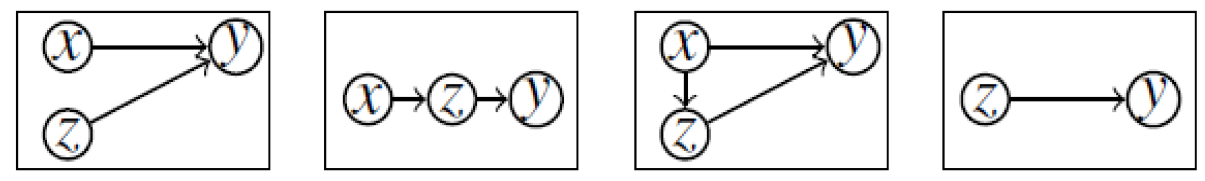
$$\mathbb{Q}(x,y) = \widetilde{\mathbb{P}}_X(x)\mathbb{Q}_{Y|x}(y)$$

Remark 1: Consider some $(x,y) \in \mathrm{Supp}(\widetilde{\mathbb{Q}}) \backslash \mathrm{Supp}(\widetilde{\mathbb{P}})$, namely, (x,y) is "covered" by $\widetilde{\mathbb{Q}}$ but not by $\widetilde{\mathbb{P}}$. Then the gradient is

$$\frac{\partial \ell}{\partial s(x,y)} = \sigma(s(x,y)) \cdot N^- \widetilde{\mathbb{Q}}(x,y)$$

This may result in slow training.

Hypothesis: The best $\widetilde{\mathbb{Q}}$ is the one that barely covers $\widetilde{\mathbb{P}}$, namely, equal to $\widetilde{\mathbb{P}}$.
Under this hypothesis, choose $\mathbb{Q}_{Y|x}$ to closely resemble $\widetilde{\mathbb{P}}_{Y|x} \Longrightarrow$ **GANs** [3]!



(a) caSGN1 (b) caSGN2 (c) caSGN3 (d) aSGN

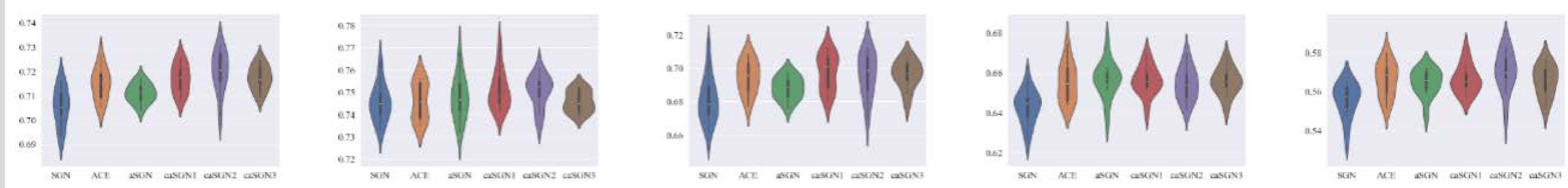
Experiments

Table 1: Spearman's ρ ($*100$) on the word similarity tasks (text8).

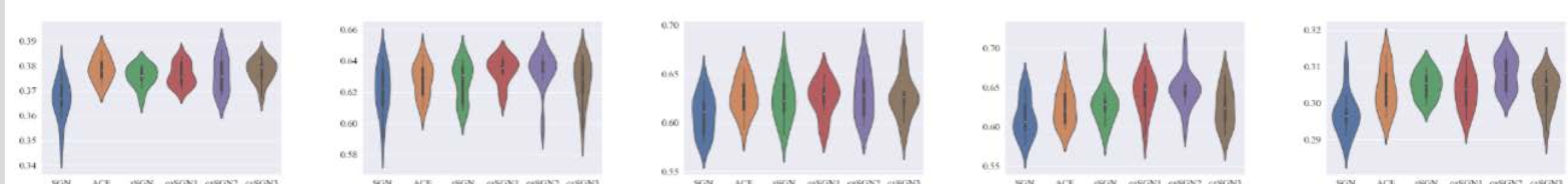
Models	WS-353	WS-SIM	WS-REL	MTurk-287	MTurk-771	RW	MEN	MC	RG	SimLex
SGN	70.58	74.54	68.10	64.29	55.59	36.63	62.16	60.82	60.17	29.69
ACE	71.49	74.61	69.50	65.52	56.63	**37.85**	62.75	62.65	62.39	30.37
aSGN	71.12	74.76	68.82	**65.67**	56.47	37.58	62.63	62.36	62.36	30.49
caSGN1	71.72	**75.11**	**69.77**	65.63	56.63	37.63	**63.40**	62.54	64.18	30.36
caSGN2	**72.02**	75.05	69.64	65.44	**57.02**	37.61	63.36	**62.86**	**64.63**	**30.79**
caSGN3	71.74	74.61	69.63	65.57	56.56	37.78	62.69	62.61	62.52	30.31

Table 2: Spearman's ρ ($*100$) on the word similarity tasks (wiki).

Models	WS-353	WS-SIM	WS-REL	MTurk-287	MTurk-771	RW	MEN	MC	RG	SimLex
SGN	67.49	74.61	61.51	63.00	59.24	39.99	68.73	64.47	69.59	31.37
ACE	71.03	**76.23**	**66.24**	63.05	60.27	40.02	67.55	76.60	70.01	31.10
aSGN	70.66	75.69	65.69	65.14	61.22	39.81	68.99	77.38	73.67	**31.58**
caSGN1	**71.56**	76.05	65.37	63.05	61.63	40.74	69.72	**80.07**	**77.58**	31.27
caSGN2	70.57	73.84	65.83	65.26	**62.28**	41.24	**70.93**	71.57	73.05	30.45
caSGN3	70.27	74.93	65.26	**65.98**	59.52	**41.55**	70.05	75.95	73.52	31.40



(a) WS-353 (b) WS-SIM (c) WS-REL (d) MTurk-287 (e) MTurk-771



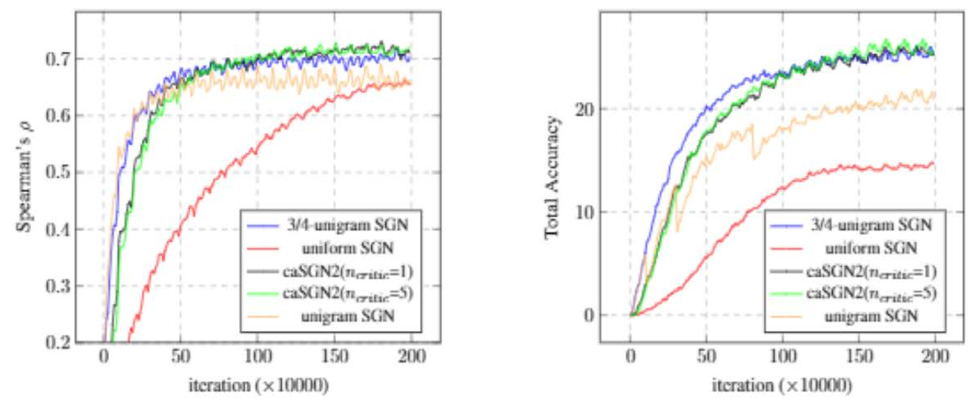
(f) RW (g) MEN (h) MC (i) RG (j) SimLex-999

Table 3: Accuracy on the word analogy task (text8).

Model	Semantic	Syntactic	Total
SGN	20.50	26.77	24.16
ACE	20.43	28.25	25.00
aSGN	20.84	27.86	24.94
caSGN1	21.25	**28.30**	**25.36**
caSGN2	**21.56**	27.79	25.20
caSGN3	20.43	27.76	24.71

Table 4: Accuracy on the word analogy task (wiki).

Model	Semantic	Syntactic	Total
SGN	27.28	35.52	31.77
ACE	27.62	35.30	31.81
aSGN	35.24	38.66	37.10
caSGN1	31.71	38.32	35.31
caSGN2	37.00	**39.96**	38.61
caSGN3	**41.21**	39.24	**40.14**



(a) WS-353 (b) Google Analogy

References

[1] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, 3111–3119.
[2] Levy, O., and Goldberg, Y. 2014. Neural word embedding as implicit matrix factorization. In Advances in neural information processing systems, 2177–2185.
[3] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In Advances in neural information processing systems , 2672–2680.

uOttawa National Research Council Canada