Continuous Word Embedding Fusion via Spectral Decomposition

Tianfan Fu
Georgia Institute of Technology
Atlanta, GA, USA
tfu42@gatech.edu

Cheng Zhang
Microsoft Research Cambridge
Cambridge, CB1 2FB, UK
cheng.zhang@microsoft.com

Stephan Mandt
Los Angeles, CA, USA
stephan.mandt@gmail.com

A Appendix

A.1 Proof of Theorem 1

Proof. First, we define some notations. We denote co-occurrence matrix $\#(w,c)$ for S_0 (before adding the new word) and $\#(w',c')$ for S_1 (after adding the new word). We have $|D| = \sum \#(w)$ and $|D'| = \sum \#(w')$. According to Equation (4) and (6), we have

$$\|S_1 - S_1'\|_F = \|S_0 - R'\|_F. \tag{1}$$

We focus on the element-wise difference between S_0 and R', both of them are sparse matrix. For the (i,j)-th element (which is nonzero), we have

$$\begin{aligned}
&(S_0)_{ij} - (R')_{ij} \\
=&\log \#(w_i,c_j) + \log|D| - \log \#(w) - \log \#(c) \\
&- \Big(\log \#(w_i,c_j) + \log|D'| \\
&- \log \#(w') - \log \#(c')\Big) \\
=&\log|D| - \log|D'| - (\log \#(w) - \log \#(w')) \\
&- ((\log \#(c) - \log \#(c')))
\end{aligned} \tag{2}$$

Now we try to upperbound $\log \#(w') - \log \#(w)$ as

$$\begin{aligned}
&\log \#(w') - \log \#(w) \\
=&\log(1 + \frac{\#(w') - \#(w)}{\#(w)}) \\
\leq&\frac{\#(w') - \#(w)}{\#(w)} \\
\leq&\frac{c_2}{c_3\sqrt{m}},
\end{aligned} \tag{3}$$

where the first inequality follows from the fact that $\log(1+x) \leq x$ for all x and the second inequality uses Assumption 2 and 3. Similarly, we have

$$\begin{aligned}
\log \#(c') - \log \#(c) \leq \frac{c_2}{c_3\sqrt{m}}, \\
\log|D'| - \log|D| \leq \frac{c_2}{c_3\sqrt{m}}.
\end{aligned} \tag{4}$$

Combining the above two equations, we have

$$|(S_0)_{ij} - (R')_{ij}| \leq \frac{3c_2}{c_3\sqrt{m}}.$$

Summing over all nonzero (i,j) pair, we obtain

$$\begin{aligned}
&\|S_1 - S_1'\|_F = \|S_0 - R'\|_F \\
\leq&\sqrt{\max\{\text{nnz}(S_0), \text{nnz}(R')\}\frac{9c_2^2}{c_3^2 m}} \\
\leq&\sqrt{\max\{\text{nnz}(S_0), \text{nnz}(S1)\}\frac{9c_2^2}{c_3^2 m}} \\
\leq&\frac{3\sqrt{c_1}c_2}{c_3}
\end{aligned} \tag{5}$$

Proved.

□

A.2 Nearest Neighbor results

To explore more about our method, we choose some typical domain-specific words and observe their nearest neighbors (in both extended and base vocabulary) after embedding. We choose “ADMM” and “autoencoders” for “NIPS Abstract”. The results are reported in Table 1, 2, 3 and 4. For Economic News, we choose “GDPs” and “reflation”. The results are reported in Table 5, 6, 7 and 8.

A.3 Learning curve of Baseline method

In this section, we provide the learning curve about FOUN and its annealing version in Figure 1. For t-th step, the step size is set to $\epsilon_t = a(t+k)^{-\alpha}$, satisfying Robbins-Monro condition. we compare these two method on several settings and show the case where $a = 1e1$ here. We compare the difference of SGD and Simulated Annealing(SA) on different learning rate for economic NEWS dataset. For each setting we run 10 independent times given the same initial condition and report the average results (solid line) and their 95% confidence interval (dashed line). Red line is the learning curve for SGD while blue line is the curve for SA. For SA method, we shrink the noise by 100 times. We find that the learning curve of simulated annealing has larger variance compared with SGD, validating the fact that the problem is non-convex. Furthermore, we also compare different cooling-down strategies (faster and slower). we find that cooling down slower would produce better results. The result also validates the fact the problem is highly non-convex.

Method	Nearest Neighbor
FOUN	brandreth blanshard wier bainton lockington goodacre estey stoddard eaton gillham golde champney sackheim ritson chapman pinchbeck moncure furr ellerton stookey
SOWE	thresholding nonlinearly drmm bhinflexible nonsmooth svrg biclustering learnable kernelized deconvolutional coreset vqa laplacians hlinear alexnet lstms

Table 1: Nearest Neighbors of “**ADMM**” in extended vocabulary.

Method	Nearest Neighbor
FOUN	brandreth blanshard wier bainton lockington goodacre estey stoddard eaton gillham golde champney sackheim ritson chapman pinchbeck moncure furr ellerton stookey
SOWE	pagerank throughput scalability latency efficiency maneuverability signal-to-noise conductivity reliability behaved glycemic torque bioavailability dof bandwidth

Table 2: Nearest Neighbors of “**ADMM**” in base vocabulary.

Method	Nearest Neighbor
FOUN	kahala sisa maco meridiana telefonica rassa ctia oyj aquiles balabac looc 3com credito disa sunsilk persil irm ipg ganoderma toco teliasonera dmo netapp cooperativa
SOWE	reparameterization kernelized softmax nonstationary nonsmooth coreset ckjinto avector learnable atakes nmbatch sgrrld regularizing tinflexible coevolving drmm

Table 3: Nearest Neighbors of “**autoencoders**” in extended vocabulary.

Method	Nearest Neighbor
FOUN	kahala sisa maco meridiana telefonica rassa ctia oyj aquiles balabac looc 3com credito disa sunsilk persil irm ipg ganoderma toco teliasonera dmo netapp cooperativa caja
SOWE	conclusion mindset probability next realization isomorphicalgebra technique stripped-down locality graph probabilistic nsongs extent idea mentality scaled-down latest formalization dystopian heuristics algorithm

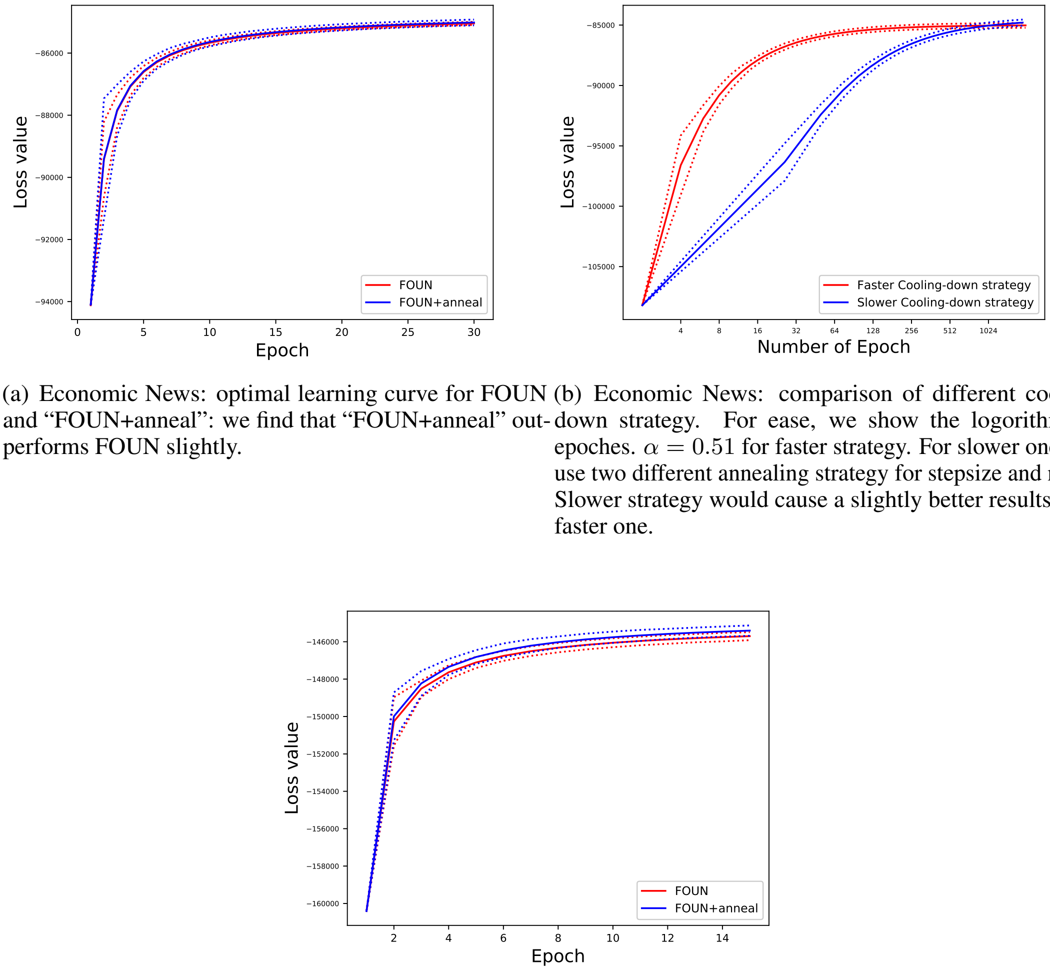
Table 4: Nearest Neighbors of “**autoencoders**” in base vocabulary.

Method	Nearest Neighbor
FOUN	kadai papaver gaura bhang dioscorea boletes zanthoxylum tetraploid jawboning ascomycetes tremella basidiomycetes cordyceps kalaripayattu flavonoid shakha basidiomycota
SOWE	retrenching laggards downsizings nonfinancial strassel itemize rapcan countertrade overinvestment napodano cordant questech gobbling cyclicals mightywords

Table 5: Nearest Neighbors of “**GDPs**” in extended vocabulary.

Method	Nearest Neighbor
FOUN	kadai papaver gaura bhang dioscorea boletes zanthoxylum tetraploid ascomycetes tremella basidiomycetes cordyceps kalaripayattu flavonoid shakha basidiomycota
SOWE	initials copyrights signatory merits offspring attest contents ashamed signatories drillia remarry notable confided divest prefixed schilpp pseudonyms extent restates pronounces

Table 6: Nearest Neighbors of “**GDPs**” in base vocabulary.



(a) Economic News: optimal learning curve for FOUN and "FOUN+anneal": we find that "FOUN+anneal" outperforms FOUN slightly.

(b) Economic News: comparison of different cooling down strategy. For ease, we show the logorithm of epoches. $\alpha = 0.51$ for faster strategy. For slower one, we use two different annealing strategy for stepsize and noise. Slower strategy would cause a slightly better results than faster one.

(c) NIPS: optimal learning curve for FOUN and "FOUN+anneal": we find that "FOUN+anneal" outperforms FOUN slightly.

Figure 1: Learning curve about FOUN and FOUN+anneal. For t-th step, the step size is set to $\epsilon_t = a(t+k)^{-\alpha}$, satisfying Robbins-Monro condition. We compare these two method on several settings and show the case where $a = 5e0$ and $k = 1e3$. When loss value of the current epoch (denoted L_t) is close (smaller than a threshold, $1e2$) to that of previous epoch (L_{t-1}), we claim that the convergence condition is met. It is the near-optimal. We compare the difference of SGD and Simulated Annealing(SA) on different learning rate for economic NEWS dataset. For each setting we run 10 independent times given the same initial condition and report the average results (solid line) and their 95% confidence interval (dashed line). Red line is the learning curve for SGD while blue line is the curve for SA. For SA method, we shrink the noise by 100 times. We find that the learning curve of simulated annealing has larger variance compared with SGD, validating the fact that the problem is non-convex. Furthermore, we also compare different cooling-down strategies (faster and slower). we find that cooling down slower would produce better results. The result also validates the fact the problem is highly nonconvex.

Method	Nearest Neighbor
FOUN	kacher gohde overstocked formspring gibraltarpedia ruess doggystyle mick badfinger sheodred seeyou dreamhost smosh weebl duddy sahaj jtrainor collegehumor monahan
ours	firming shakeout sprinkel tion countertrade sluggishness dovish tiie orszag pessimists tlie jawboning recessionary bankshares flation refunding pretax junkins redemptions

Table 7: Nearest Neighbors of "**reflation**" in extended vocabulary.

Method	Nearest Neighbor
FOUN	kacher gohde formspring gibraltarpedia ruess doggystyle mick badfinger sheodred seeyou dreamhost smosh weebl duddy sahaj jtrainor collegehumor monahan dionyseus
SOWE	globalisation overpopulation warming soros globalization newsround carbon-neutral anata autarky reaganomics plunk splat clurman imbalance anti-semitism

Table 8: Nearest Neighbors of "**reflation**" in base vocabulary.