PROJECT REPORT

COL870

Instructed by:Prof. Anurag Mittal

SPINE: SParse Interpretable Neural Embeddings

Anant Subramanian*, Danish Pruthi*, Harsh Jhamtani*, Taylor Berg-Kirkpatrick, Eduard Hovy

(AAAI 2018)

Submitted by:

Prachi(2021CSY7584)

Nikki Tiwari(2021SIY7560)

Table of Contents-

INTRODUCTION

Distributed representations map words to vectors of real numbers in a continuous space. However, word vectors have dense representations that humans find difficult to interpret. For instance, we are often clueless as to what a "high" value along a given dimension of a vector signifies when compared to a "low" value. So, we get a question-How does one transform word representations to a new space where they are more interpretable?

To address the question, in this paper, we make following contributions:
• A denoising k-sparse autoencoder is employed to obtain SParse Interpretable Neural Embeddings (SPINE), a transformation of input word embeddings. The autoencoder is trained using a novel learning objective and activation function to attain interpretable and efficient representations.
• SPINE is evaluated using a large scale, crowdsourced, intrusion detection test, along with a battery of downstream tasks.

DATASET

- The autoencoder models are pre-trained on GloVe and word2vec embeddings.
- The GloVe vectors were trained on 6 billion tokens from a 2014 dump of Wikipedia and Gigaword5
- Word2vec vectors were trained on around 100 billion words from a part of the Google News dataset.
- Both the GloVe and Word2vec embeddings are 300 dimensions long
- 17k most frequently occurring words are selected out of which 15K and 2K words are used for training and hyperparameter tuning respectively.

METHODOLOGY

Given : $D = [X_1, X_2, X_3, \ldots, X_v]^T \in R_{Vxd}$

Goal : To project these embeddings to a space Rm such that the m-dimensional embeddings in this space are both sparse and non-negative. The required transformation is $R^{VXd} \rightarrow R^{VXm}$.

Method: K- sparse autoencoder is trained to minimize a loss function that concisely captures the required sparsity constraints. Let X_i' be the predicted output for an input embedding Xi ∈ D. That is

$$Z^{(Xi)} = f(X_i W_e + b_e)$$
$$X_i' = Z^{(Xi)} W_o + b_o$$

The set $Z = \{Z^{(X1)}, Z^{(X2)}, \ldots, Z^{(Xm)}\}$ is the set of required sparse embeddings corresponding to each of the input embeddings.

ACTIVATION FUNCTION:

For interpretability -nonnegativity in the output embeddings is a useful property in the context of interpretability.

- Both RELU and sigmoid can be used .
- For sparsity -most of the entries in the embeddings to be zero
- So, sigmoid activation rules out due to its asymptotic nature with respect to 0 activation.
- So, the final activation function is RELU

LOSS FUNCTIONS

In this setting, given *D*, our k-sparse autoencoder is trained to minimize the following loss function.

$$L(D) = RL(D) + \lambda_1 ASL(D) + \lambda_2 PSL(D)$$

Where
RL(D) is the **reconstruction loss** over the data set,
ASL(D) is the **average sparsity loss** over the data set, and
PSL(D) is the **partial sparsity loss** over the data set.
The coefficients λ1 and λ2 determine the relative importance of the two penalty terms.

Reconstruction Loss *(RL)*

RL(D) is the average loss in reconstructing the input representation from the learned representation. If the reconstructed output for an input vector X ∈ R d is Xe ∈ R d , then

$$RL(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{X} \in \mathcal{D}} \left\| \mathbf{X} - \tilde{\mathbf{X}} \right\|_2^2$$

Average Sparsity Loss (ASL)

In order to enforce k-sparse activations in the hidden layer,a modification to the basic autoencoder loss function is described that penalizes any deviation of the observed average activation value from the desired average activation value of a given hidden unit, over a given data set. The loss is as follows:

$$ASL(\mathcal{D}) = \sum_{h \in \mathcal{H}} \left(\max\left(0, \rho_{h,\mathcal{D}} - \rho^*_{h,\mathcal{D}}\right) \right)^2$$

Partial Sparsity Loss (PSL)

It is possible to obtain an ASL value of 0 without actually having k-sparse representations. For example, to obtain an average activation value of 0.5 for a given hidden unit across 4 examples, one feasible solution is to have an

activation value of 0.5 for all the four examples. To obtain activation values that are truly k-sparse, a novel partial sparsity loss term is introduced that penalizes values that are neither close to 0 nor 1, pushing them close to either 0 or 1. The following formulation of PSL is used to do so

$$PSL(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{X} \in \mathcal{D}} \sum_{h \in \mathcal{H}} \left(Z_h^{(\mathbf{X})} \times \left(1 - Z_h^{(\mathbf{X})}\right) \right)$$

This key addition to the loss term facilitates the generation of sparse embeddings with activations close to 0 and 1.

MODEL

A k-sparse autoencoder is an autoencoder for which, with high probability, at most k hidden units are active for any given input. Ng (2011) introduced a mechanism to train k-sparse autoencoders. The underlying idea is to achieve an expected activation value for a hidden unit that is equivalent to k completely activated hidden units.
We have make 2 model for autoencoder

Model 1 (used in paper)

1) Encoder having 1 FC layer which encodes the original embedding of 300 dim into 1000 dim
2) Decoder having 1 FC laye which reconstructs the original 300 dim embedding back from 1000-dim SPINE embedding

Model 2 (modified in the project)

1)Encoder having 2 FC layer with hidden neurons 500,1000 resp, which encodes the original embedding of 300 dim into 500 dim and then to 1000 dim

2)Decoder having 1 FC laye which reconstructs the original 300 dim embedding back from 1000-dim SPINE embedding

DATASETS

Used in paper :

- The autoencoder models are pre-trained on GloVe and word2vec embeddings.
- The GloVe vectors were trained on 6 billion tokens from a 2014 dump of Wikipedia and Gigaword5
- Word2vec vectors were trained on around 100 billion words from a part of the Google News dataset.
- Both the GloVe and Word2vec embeddings are 300 dimensions long.
- 17 K most frequently occurring words are selected out of which 15 K and 2K words are used for training and hyperparameter tuning respectively.

Extra dataset used by in the project :

- Glove 300 dimensions embeddings
- Glove 100 dimensions embeddings
- Word2Vec 300 dimensions embeddings
- Bert 768 dimension embeddings

EXPERIMENTS

The interpretability of the resulting representations against the ones obtained from baseline models is estimated in the following two ways:

- Word Intrusion Detection Test
- Benchmark Downstream Tasks
    - Noun Phrase Bracketing
    - Question Classification (TREC)
    - News Classification
    - Word Similarity Task

Model 1

Word similarity tasks

- WS-353 dataset which contains 353 pairs of English words is used for word similarity task. Each pair of words has been assigned similarity ratings by multiple human annotators.
- Cosine similarity between the embeddings of each pair of words is calculated, and the Spearman's rank correlation coefficient ρ is computed between the human scored list and the predicted similarity list.
- Only those pairs of words are taken where both words are present in the vocabulary.

Word similarity scores for SPINE embeddings of Glove 300 dimensions

Cosine similarity of embeddings Predicted by model	Similarity ratings by multiple Human annotators
4.28829728e-01	6.77000000e+00]
3.12890716e-01	7.35000000e+00
1.00000000e+00	1.00000000e+01
4.54216054e-01	7.46000000e+00
4.41461115e-01	7.62000000e+00
5.85778242e-01	7.58000000e+00
3.81521585e-01	5.77000000e+00
2.47869744e-01	5.96000000e+00
1.02494786e-01	1.31000000e+00
2.27572473e-01	3.04000000e+00
5.51504005e-01	8.13000000e+00]
1.52188939e-01	5.54000000e+00

Word similarity scores for SPINE embeddings of Word2Vec 300 dimensions

Cosine similarity of embeddings Predicted by model	Similarity ratings by multiple Human annotators
0.19508175	6.77
0.76384414	7.35
1	10
0.54014866	7.46
0.38287773	7.62
0.45244305	7.58
0.26086565	5.77
0.38383908	7.5
0.73672559	6.77
0.32459272	7.42

Word similarity scores for SPINE embeddings of Bert 768 dimensions

Cosine similarity of embeddings Predicted by model	Similarity ratings by multiple Human annotators
0.73418906	6.77
0.63358539	7.35
1	10
0.61085131	7.46
0.79036582	7.62

0.73851502	7.58
0.55109324	5.77
0.61850624	7.5
0.40738436	6.77
0.52319311	7.42

Word similarity scores for SPINE embeddings of GloVe 100 dimensions

Cosine similarity of embeddings Predicted by model	Similarity ratings by multiple Human annotators
0.73600211	6.77
0.63899213	7.35
1	10
0.80879407	7.46
0.58838027	7.62
0.83740333	7.58
0.77615633	5.77
0.81898245	7.5
0.89222949	6.77
0.80997182	7.42

Final table for the spearman rank coefficient for different datasets is as follows:

Datasets	**Spearman_rank_coeff**	**sp_rho**
GloVe 300 dimensions	0.6477005798034899	5.527800714848622e-29
Word2Vec 300 dimension	.6373226632772291	1.1060875118763923e-34
GloVe 100 dimensions	0.46291997928143563	0.46291997928143563
Glove with(2 hidden layers)	0.21573009095891893	0.0009423928013857126
Bert embeddings	0.3647800768247168	1.0365912917105751e-08

Sparsity

The dimension of the encoded embedding is 1000 dimension for all the models used in the project with the following sparsity:

Model	**Sparsity**
SPINE on 300-dim Word2Vec	93.63407333333333
SPINE on 300-dim Glove embeddings	87.79586
SPINE on 100-dim Glove embeddings	89.72781333333333
SPINE on 768-dim BERT embeddings	94.25645333333334

Quantative Assesment (Visualization)

Steps followed:

- We have randomly picked some words
- For each picked word

- We have examined top participating dimensions- the dimensions that have highest absolute values for the words .
- for each of selected top dimension we notes words that have highest absolute values for that word

Analysing SPINE embeddings generated from 300-dim GloVe embeddings

1)The word is **intelligence**

Word of interest = intelligence
The contribution of the word 'intelligence' in dimension 575 = 0.857463
Following are the top words in dimension 575 along with their contributions
[(1.0, 'libby'), (1.0, 'leaked'), (1.0, 'confidential'), (0.877596, 'fbi'), (0.860398, 'classified'), (0.859306, 'memo'), (0.858536, 'leak'), (0.857463, 'intelligence'), (0.839776, 'informant'), (0.835968, 'secrets')]
The contribution of the word 'intelligence' in dimension 482 = 0.738917
Following are the top words in dimension 482 along with their contributions
[(1.0, 'warships'), (1.0, 'stealth'), (1.0, 'reconnaissance'), (1.0, 'planes'), (1.0, 'pentagon'), (1.0, 'missile'), (1.0, 'aircraft'), (1.0, 'afghanistan'), (0.996355, 'fighter'), (0.994009, 'drone')]
The contribution of the word 'intelligence' in dimension 272 = 0.733774
Following are the top words in dimension 272 along with their contributions
[(1.0, 'police'), (1.0, 'officers'), (1.0, 'lt.'), (1.0, 'commander'), (0.989962, 'detective'), (0.980658, 'brig.'), (0.9387, 'chief'), (0.90963, 'commandant'), (0.908764, 'kgb'), (0.895963, 'sergeant')]
The contribution of the word 'intelligence' in dimension 648 = 0.599943
Following are the top words in dimension 648 along with their contributions
[(0.944636, 'plotting'), (0.942302, 'mafia'), (0.929367, 'bribe'), (0.928841, 'indicted'), (0.903964, 'bribery'), (0.897165, 'plot'), (0.879948, 'convicted'), (0.848832, 'assassination'), (0.843293, 'conspiracy'), (0.834164, 'jailed')]
The contribution of the word 'intelligence' in dimension 745 = 0.520636
Following are the top words in dimension 745 along with their contributions
[(1.0, 'institute'), (0.943532, 'psychiatry'), (0.913642, 'sciences'), (0.897935, 'physics'), (0.87457, 'laboratory'), (0.861861, 'research'), (0.855285, 'professor'), (0.837385, 'genetics'), (0.824483, 'scientist'), (0.818096, 'physicist')]

2)the word is **`science`**

Word of interest = science
The contribution of the word 'science' in dimension 486 = 0.816299
Following are the top words in dimension 486 along with their contributions
[(1.0, 'university'), (1.0, 'undergraduate'), (1.0, 'sociology'), (1.0, 'professor'), (1.0, 'phd'), (1.0, 'humanities'), (1.0, 'graduate'), (1.0, 'doctorate'), (1.0, 'degree'), (1.0, 'bachelor')]
The contribution of the word 'science' in dimension 635 = 0.668243
Following are the top words in dimension 635 along with their contributions
[(1.0, 'sciences'), (1.0, 'honorary'), (1.0, 'faculty'), (1.0, 'doctorate'), (1.0, 'chemistry'), (1.0, 'bachelor'), (1.0, '1893'), (1.0, '1892'), (1.0, '1891'), (1.0, '1887')]
The contribution of the word 'science' in dimension 745 = 0.612640
Following are the top words in dimension 745 along with their contributions
[(1.0, 'institute'), (0.943532, 'psychiatry'), (0.913642, 'sciences'), (0.897935, 'physics'), (0.87457, 'laboratory'), (0.861861, 'research'), (0.855285, 'professor'), (0.837385, 'genetics'), (0.824483, 'scientist'), (0.818096, 'physicist')]
The contribution of the word 'science' in dimension 518 = 0.574628
Following are the top words in dimension 518 along with their contributions
[(0.989433, 'machinery'), (0.984088, 'plastics'), (0.953205, 'electronics'), (0.936411, 'industries'), (0.90491, 'sectors'), (0.888523, 'appliances'), (0.886905, 'miscellaneous'), (0.865031, 'agriculture'), (0.821493, 'sciences'), (0.820733, 'engineering')]
The contribution of the word 'science' in dimension 939 = 0.539502
Following are the top words in dimension 939 along with their contributions
[(0.919611, 'philosophical'), (0.774294, 'topics'), (0.745954, 'theories'), (0.735126, 'philosophy'), (0.691797, 'explores'), (0.691462, 'theoretical'), (0.68669, 'biblical'), (0.686139, 'writings'), (0.68276, 'mysteries'), (0.664075, 'teachings')]

3)The word of **residents**

Word of interest = residents
The contribution of the word 'residents' in dimension 83 = 1.000000
Following are the top words in dimension 83 along with their contributions
[(1.0, 'young'), (1.0, 'whom'), (1.0, 'themselves'), (1.0, 'students'), (1.0, 'residents'), (1.0, 'parents'), (1.0, 'members'), (1.0, 'friends'), (1.0, 'fans'), (1.0, 'americans')]
The contribution of the word 'residents' in dimension 959 = 0.578546
Following are the top words in dimension 959 along with their contributions
[(1.0, 'wounded'), (1.0, 'total'), (1.0, 'people'), (1.0, 'number'), (1.0, 'nine'), (1.0, 'least'), (1.0, 'killed'), (1.0, 'injured'), (1.0, '70'), (1.0, '100')]
The contribution of the word 'residents' in dimension 862 = 0.557699
Following are the top words in dimension 862 along with their contributions

[(1.0, 'redevelopment'), (1.0, 'borough'), (0.983881, 'zoning'), (0.972571, 'neighborhoods'), (0.967753, 'urban'), (0.943908, 'boroughs'), (0.855015, 'residential'), (0.850603, 'neighborhood'), (0.84964, 'housing'), (0.848653, 'tenants')]
The contribution of the word 'residents' in dimension 363 = 0.405057
Following are the top words in dimension 363 along with their contributions
[(1.0, 'wounding'), (1.0, 'wounded'), (1.0, 'gunmen'), (0.954116, 'convoy'), (0.953344, 'gunfire'), (0.944855, 'killed'), (0.934157, 'grenades'), (0.92345, 'clashes'), (0.921332, 'baghdad'), (0.911051, 'ambushed')]
The contribution of the word 'residents' in dimension 529 = 0.396794
Following are the top words in dimension 529 along with their contributions
[(0.871024, 'tents'), (0.768254, 'rations'), (0.652309, 'homeless'), (0.62604, 'needy'), (0.61288, 'shelters'), (0.56642, 'food'), (0.563817, 'shelter'), (0.563103, 'belongings'), (0.523557, 'survivors'), (0.512051, 'medicines')]

intelligence	'libby', 'leaked', 'confidential', 'fbi' 'classified','memo','leak', 'intelligence', 'informant', 'secrets'
science	'university', 'undergraduate', 'sociology', 'professor', 'phd', 'humanities', 'graduate', 'doctorate', 'degree', 'bachelor'
residents	'young',, 'whom', 'themselves', 'students', 'residents', 'parents', 'members', 'friends', 'fans', 'americans'

Analysis of SPINE embeddings generated from 300-dim word2vec embeddings

1)The word is intelligence

Word of interest = intelligence

Top participating dimension=[126,373,351,250,717]
The contribution of the word 'intelligence' in dimension 126 = 0.259149
Following are the top words in dimension 126 along with their contributions
[(0.692847, 'confidential'), (0.604553, 'secrets'), (0.477451, 'privacy'), (0.475694, 'secrecy'), (0.440114, 'secret'), (0.429814, 'leaked'), (0.41517, 'encryption'), (0.379452, 'hacker'), (0.362801, 'spying'), (0.350274, 'spy')]
The contribution of the word 'intelligence' in dimension 373 = 0.254489
Following are the top words in dimension 373 along with their contributions
[(0.608453, 'bombings'), (0.578556, 'plot'), (0.567095, 'plotting'), (0.541591, 'terrorist'), (0.466947, 'espionage'), (0.461198, 'terror'), (0.460705, 'plotted'), (0.460014, 'assassination'), (0.45651, 'plots'), (0.437507, 'bomb')]
The contribution of the word 'intelligence' in dimension 351 = 0.161416
Following are the top words in dimension 351 along with their contributions
[(0.408696, 'statistical'), (0.404902, 'statistics'), (0.394729, 'measurements'), (0.391682, 'analysis'), (0.390851, 'calculate'), (0.365541, 'calculations'), (0.364624, 'accurate'), (0.357602, 'calculation'), (0.350231, 'data'), (0.342966, 'comparisons')]
The contribution of the word 'intelligence' in dimension 250 = 0.158742
Following are the top words in dimension 250 along with their contributions
[(0.589924, 'obama'), (0.473045, 'cheney'), (0.440059, 'clinton'), (0.376801, 'congressional'), (0.354728, 'iraq'), (0.319052, 'senator'), (0.313614, 'mccain'), (0.312311, 'troop'), (0.307993, 'gop'), (0.307377, 'washington')]
The contribution of the word 'intelligence' in dimension 717 = 0.150282
Following are the top words in dimension 717 along with their contributions
[(0.196892, 'manpower'), (0.188469, 'patrols'), (0.18457, 'coordination'), (0.169818, 'insurgency'), (0.169278, 'cyber'), (0.163525, 'coordinate'), (0.162429, 'coordinated'), (0.158976, 'terrorism'), (0.156738, 'expeditionary'), (0.154828, 'battalions')]

2) the word is science

Word of interest = science
The contribution of the word 'science' in dimension 206 = 0.341808
Following are the top words in dimension 206 along with their contributions
[(1.0, 'doctorate'), (0.925088, 'professor'), (0.868356, 'doctoral'), (0.787881, 'lecturer'), (0.73521, 'sociology'), (0.679905, 'anthropology'), (0.665771, 'professors'), (0.656905, 'dean'), (0.644319, 'undergraduate'), (0.612277, 'dissertation')]
The contribution of the word 'science' in dimension 954 = 0.329959
Following are the top words in dimension 954 along with their contributions

[(0.792442, 'algebra'), (0.644488, 'exam'), (0.638213, 'courses'), (0.626758, 'exams'), (0.605511, 'math'), (0.586214, 'curriculum'), (0.562832, 'mathematics'), (0.542183, 'classes'), (0.539728, 'diploma'), (0.506845, 'teaches')]
The contribution of the word 'science' in dimension 334 = 0.212372
Following are the top words in dimension 334 along with their contributions
[(0.26864, 'sciences'), (0.250766, 'biology'), (0.250609, 'engineering'), (0.240491, 'geology'), (0.220268, 'veterinary'), (0.212372, 'science'), (0.211838, 'agriculture'), (0.211291, 'physics'), (0.209162, 'aerospace'), (0.204978, 'astronomy')]
The contribution of the word 'science' in dimension 863 = 0.190043
Following are the top words in dimension 863 along with their contributions
[(0.938625, 'patents'), (0.844754, 'patent'), (0.594294, 'inventions'), (0.473468, 'pharmaceutical'), (0.451233, 'patented'), (0.444887, 'royalties'), (0.437565, 'inventor'), (0.422462, 'collaborations'), (0.403075, 'innovation'), (0.393665, 'discoveries')]
The contribution of the word 'science' in dimension 138 = 0.173772
Following are the top words in dimension 138 along with their contributions
[(0.400093, 'robots'), (0.365348, 'robotic'), (0.336592, 'algorithms'), (0.31857, 'robot'), (0.308856, 'imaging'), (0.302075, 'computational'), (0.289543, 'computerized'), (0.275038, 'neural'), (0.271736, 'machines'), (0.264544, 'technologies')]

3)the word is residents

Word of interest = residents
The contribution of the word 'residents' in dimension 152 = 0.636624
Following are the top words in dimension 152 along with their contributions
[(0.889577, 'viewers'), (0.826028, 'listeners'), (0.748886, 'travelers'), (0.718482, 'readers'), (0.701529, 'commuters'), (0.692269, 'consumers'), (0.684504, 'fans'), (0.67035, 'audiences'), (0.662655, 'devotees'), (0.656253, 'patrons')]
The contribution of the word 'residents' in dimension 959 = 0.327167
Following are the top words in dimension 959 along with their contributions
[(0.472924, 'climbers'), (0.457502, 'wrestlers'), (0.444738, 'astronauts'), (0.434926, 'cadets'), (0.431367, 'swimmers'), (0.427198, 'skaters'), (0.424545, 'sailors'), (0.419902, 'airmen'), (0.41738, 'pilots'), (0.415792, 'inmates')]
The contribution of the word 'residents' in dimension 128 = 0.281138
Following are the top words in dimension 128 along with their contributions
[(0.414714, 'annexation'), (0.368248, 'townships'), (0.339527, 'annexed'), (0.32571, 'suburb'), (0.314453, 'unincorporated'), (0.293444, 'headquartered'), (0.29279, 'mayor'), (0.281138, 'residents'), (0.272997, 'neighborhoods'), (0.271098, 'suburbs')]
The contribution of the word 'residents' in dimension 842 = 0.239433

Following are the top words in dimension 842 along with their contributions [(0.737633, 'commissioners'), (0.69534, 'township'), (0.617129, 'townships'), (0.579398, 'county'), (0.572287, 'ordinance'), (0.519771, 'annexation'), (0.508809, 'levy'), (0.48494, 'borough'), (0.484603, 'sheriff'), (0.472199, 'municipalities')]
The contribution of the word 'residents' in dimension 325 = 0.140841
Following are the top words in dimension 325 along with their contributions [(0.373727, 'incomes'), (0.363458, 'disabilities'), (0.328084, 'deduction'), (0.313569, 'unemployed'), (0.309681, 'mortgage'), (0.298009, 'homes'), (0.294926, 'homeless'), (0.290278, 'disability'), (0.288704, 'dwellings'), (0.288237, 'housing')]

intelligence	'confidential'), 'secrets'), , 'privacy'), , 'secrecy', 'secret'), 'leaked','encryption', 'hacker', 'spying' 'spy'
science	'doctorate', 'professor', 'doctoral', 'lecturer', 'sociology', (0.679905, 'anthropology', 'professors', 'dean', , 'undergraduate', 'dissertation'
residents	'viewers','listeners', 'travelers', 'readers' 'commuters' 'consumers', 'fans', 'audiences' , 'devotees' ,'patrons,

Analysis of SPINE embeddings generated from 768-dim BERT embeddings

1)the word is intelligence

Word of interest = intelligence
The contribution of the word 'intelligence' in dimension 734 = 0.969242
Following are the top words in dimension 734 along with their contributions [(1.0, 'secretary'), (1.0, 'relations'), (1.0, 'political'), (1.0, 'law'), (1.0, 'jury'), (1.0, 'involved'), (1.0, 'insurance'), (1.0, 'foreign'), (1.0, 'civil'), (1.0, 'administrative')]
The contribution of the word 'intelligence' in dimension 277 = 0.835721
Following are the top words in dimension 277 along with their contributions

[(1.0, 'total'), (1.0, 'television'), (1.0, 'region'), (1.0, 'points'), (1.0, 'october'), (1.0, 'media'), (1.0, 'information'), (1.0, 'history'), (1.0, 'country'), (1.0, 'british')]
The contribution of the word 'intelligence' in dimension 27 = 0.796349
Following are the top words in dimension 27 along with their contributions
[(1.0, 'win'), (1.0, 'television'), (1.0, 'stadium'), (1.0, 'mail'), (1.0, 'lawyer'), (1.0, 'daily'), (1.0, 'annual'), (1.0, 'americans'), (1.0, 'american'), (1.0, '--')]
The contribution of the word 'intelligence' in dimension 877 = 0.687551
Following are the top words in dimension 877 along with their contributions
[(1.0, 'win'), (1.0, 'total'), (1.0, 'team'), (1.0, 'road'), (1.0, 'points'), (1.0, 'lost'), (1.0, 'history'), (1.0, 'germany'), (1.0, 'german'), (1.0, 'american')]
The contribution of the word 'intelligence' in dimension 825 = 0.648800
Following are the top words in dimension 825 along with their contributions
[(1.0, 'tournament'), (1.0, 'staff'), (1.0, 'scheduled'), (1.0, 'recorded'), (1.0, 'polls'), (1.0, 'online'), (1.0, 'interview'), (1.0, 'hotel'), (1.0, 'editor'), (1.0, 'cast')]

2)the word is science

The contribution of the word 'science' in dimension 394 = 0.961116
Following are the top words in dimension 394 along with their contributions
[(1.0, 'travel'), (1.0, 'sciences'), (1.0, 'review'), (1.0, 'publishing'), (1.0, 'proceedings'), (1.0, 'press'), (1.0, 'eds'), (1.0, 'ed'), (1.0, 'cheney'), (1.0, 'articles')]
The contribution of the word 'science' in dimension 989 = 0.954481
Following are the top words in dimension 989 along with their contributions
[(1.0, 'venezuela'), (1.0, 'mining'), (1.0, 'law'), (1.0, 'gm'), (1.0, 'economists'), (1.0, 'economics'), (1.0, 'colombia'), (1.0, 'beijing'), (1.0, 'architecture'), (1.0, 'accounting')]
The contribution of the word 'science' in dimension 999 = 0.711480
Following are the top words in dimension 999 along with their contributions
[(1.0, 'village'), (1.0, 'telephone'), (1.0, 'source'), (1.0, 'potential'), (1.0, 'population'), (1.0, 'personnel'), (1.0, 'navy'), (1.0, 'internet'), (1.0, 'census'), (1.0, 'airport')]
The contribution of the word 'science' in dimension 254 = 0.673572
Following are the top words in dimension 254 along with their contributions
[(1.0, 'species'), (1.0, 'sources'), (1.0, 'match'), (1.0, '9'), (1.0, '6'), (1.0, '19'), (1.0, '18'), (1.0, '16'), (1.0, '11'), (1.0, '10')]
The contribution of the word 'science' in dimension 901 = 0.671730
Following are the top words in dimension 901 along with their contributions
[(1.0, 'three'), (1.0, 'president'), (1.0, 'nations'), (1.0, 'nation'), (1.0, 'left'), (1.0, 'four'), (1.0, 'five'), (1.0, 'february'), (1.0, '15'), (1.0, '14')]

3)the word is residents

The contribution of the word 'residents' in dimension 357 = 0.842695
Following are the top words in dimension 357 along with their contributions
[(1.0, 'senators'), (1.0, 'officials'), (1.0, 'ministers'), (1.0, 'democratic'), (1.0, 'coast'), (1.0, 'christian'), (1.0, 'changes'), (1.0, 'appearances'), (1.0, 'albums'), (1.0, 'adults')]
The contribution of the word 'residents' in dimension 816 = 0.778871
Following are the top words in dimension 816 along with their contributions
[(1.0, 'war'), (1.0, 'run'), (1.0, 'population'), (1.0, 'november'), (1.0, 'german'), (1.0, 'french'), (1.0, 'final'), (1.0, 'enough'), (1.0, 'deal'), (1.0, '6')]
The contribution of the word 'residents' in dimension 326 = 0.629583
Following are the top words in dimension 326 along with their contributions
[(1.0, 'temple'), (1.0, 'ohio'), (1.0, 'lebanese'), (1.0, 'joseph'), (1.0, 'jose'), (1.0, 'joe'), (1.0, 'houston'), (1.0, 'francisco'), (1.0, 'chris'), (1.0, 'bishop')]
The contribution of the word 'residents' in dimension 999 = 0.563764
Following are the top words in dimension 999 along with their contributions
[(1.0, 'village'), (1.0, 'telephone'), (1.0, 'source'), (1.0, 'potential'), (1.0, 'population'), (1.0, 'personnel'), (1.0, 'navy'), (1.0, 'internet'), (1.0, 'census'), (1.0, 'airport')]
The contribution of the word 'residents' in dimension 892 = 0.547316
Following are the top words in dimension 892 along with their contributions
[(1.0, 'roads'), (1.0, 'nights'), (1.0, 'care'), (1.0, 'banks'), (0.99511135, 'watches'), (0.9792694, 'statistics'), (0.9765979, 'fans'), (0.95631623, 'runs'), (0.94974446, 'clothes'), (0.9388703, 'knows')]

intelligence	'Secretary', 'relations', 'political' , 'law' , 'jury' , 'involved' , 'insurance' , 'foreign' , 'civil' , 'administrative'
science	'travel', 'sciences', , 'review', 'publishing', 'proceedings' , 'press', 'eds', 'ed', , 'cheney', 'articles'
residents	'senators', 'officials', 'ministers', 'democratic', 'coast', 'christian', 'changes', 'appearances', 1.0, 'albums', 'adults'

Analysis of SPINE embeddings generated from 100-dim GloVe embeddings

1)The word is intelligence

Word of interest = intelligence
The top participating dimension-[463,869,551,544,992]
The contribution of the word 'intelligence' in dimension 463 = 1.000000
Following are the top words in dimension 463 along with their contributions
[(1.0, 'was'), (1.0, 'to'), (1.0, 'the'), (1.0, 'that'), (1.0, 'on'), (1.0, 'of'), (1.0, 'in'), (1.0, 'for'), (1.0, 'and'), (1.0, 'a')]
The contribution of the word 'intelligence' in dimension 869 = 0.668014
Following are the top words in dimension 869 along with their contributions
[(0.86469805, 'war'), (0.83048975, 'language'), (0.82881796, 'military'), (0.81639874, 'students'), (0.81526506, 'political'), (0.80816257, 'his'), (0.8049718, 'her'), (0.7974087, 'troops'), (0.7967696, 'speaking'), (0.7936704, 'religious')]
The contribution of the word 'intelligence' in dimension 551 = 0.447766
Following are the top words in dimension 551 along with their contributions
[(0.80025774, 'data'), (0.712367, 'equipment'), (0.71176016, 'mechanical'), (0.70573896, 'physical'), (0.6806821, 'computer'), (0.67484933, 'fluid'), (0.6650082, 'breathing'), (0.6643735, 'mechanics'), (0.6584725, 'liquid'), (0.6578232, 'luggage')]
The contribution of the word 'intelligence' in dimension 544 = 0.381709
Following are the top words in dimension 544 along with their contributions
[(0.67144644, 'brain'), (0.62515515, 'bone'), (0.5717817, 'genome'), (0.57112396, 'fractured'), (0.5523084, 'spinal'), (0.5291256, 'dinosaur'), (0.5239038, 'scan'), (0.5227695, 'rna'), (0.5166182, 'muscles'), (0.5143618, 'spectrum')]
The contribution of the word 'intelligence' in dimension 992 = 0.347351
Following are the top words in dimension 992 along with their contributions
[(0.4266755, 'news'), (0.42188734, 'ale'), (0.41153428, 'of'), (0.40809348, 'crude'), (0.39871404, 'presenter'), (0.3908948, 'producer'), (0.3857067, 'afghan'), (0.3853361, 'rivalries'), (0.38137284, 'sports'), (0.37959105, 'refinery')]

2)The word is science

Word of interest = science
Top participating dimension =[463,869,466,551,183]
The contribution of the word 'science' in dimension 463 = 1.000000
Following are the top words in dimension 463 along with their contributions

[(1.0, 'was'), (1.0, 'to'), (1.0, 'the'), (1.0, 'that'), (1.0, 'on'), (1.0, 'of'), (1.0, 'in'), (1.0, 'for'), (1.0, 'and'), (1.0, 'a')]
The contribution of the word 'science' in dimension 869 = 0.569018
Following are the top words in dimension 869 along with their contributions
[(0.86469805, 'war'), (0.83048975, 'language'), (0.82881796, 'military'), (0.81639874, 'students'), (0.81526506, 'political'), (0.80816257, 'his'), (0.8049718, 'her'), (0.7974087, 'troops'), (0.7967696, 'speaking'), (0.7936704, 'religious')]
The contribution of the word 'science' in dimension 466 = 0.461960
Following are the top words in dimension 466 along with their contributions
[(0.62917334, 'matthew'), (0.62547445, 'islam'), (0.6115369, 'prophet'), (0.58658236, 'girls'), (0.57706445, 'conspiracy'), (0.56763554, 'college'), (0.5654101, 'islamic'), (0.54367864, 'shane'), (0.54049116, 'university'), (0.53506833, 'teachings')]
The contribution of the word 'science' in dimension 551 = 0.330509
Following are the top words in dimension 551 along with their contributions
[(0.80025774, 'data'), (0.712367, 'equipment'), (0.71176016, 'mechanical'), (0.70573896, 'physical'), (0.6806821, 'computer'), (0.67484933, 'fluid'), (0.6650082, 'breathing'), (0.6643735, 'mechanics'), (0.6584725, 'liquid'), (0.6578232, 'luggage')]
The contribution of the word 'science' in dimension 183 = 0.317086
Following are the top words in dimension 183 along with their contributions
[(0.5594615, 'federer'), (0.53081805, 'andersen'), (0.51710975, 'wimbledon'), (0.50655586, 'armistice'), (0.50611836, 'finalist'), (0.4777031, 'beatles'), (0.4742257, 'facts'), (0.4720664, 'ulster'), (0.47131974, 'masters'), (0.46970582, 'conceptual')]

3)the word is residents

Word of interest = residents
Top participating dimension=[463,869,57,292,618]
The contribution of the word 'residents' in dimension 463 = 1.000000
Following are the top words in dimension 463 along with their contributions
[(1.0, 'was'), (1.0, 'to'), (1.0, 'the'), (1.0, 'that'), (1.0, 'on'), (1.0, 'of'), (1.0, 'in'), (1.0, 'for'), (1.0, 'and'), (1.0, 'a')]
The contribution of the word 'residents' in dimension 869 = 0.562741
Following are the top words in dimension 869 along with their contributions
[(0.86469805, 'war'), (0.83048975, 'language'), (0.82881796, 'military'), (0.81639874, 'students'), (0.81526506, 'political'), (0.80816257, 'his'), (0.8049718, 'her'), (0.7974087, 'troops'), (0.7967696, 'speaking'), (0.7936704, 'religious')]
The contribution of the word 'residents' in dimension 57 = 0.425558
Following are the top words in dimension 57 along with their contributions

[(0.6802267, 'toll'), (0.63596934, 'released'), (0.6302871, 'unemployment'), (0.6076101, 'grim'), (0.5865692, 'report'), (0.58329666, 'prisons'), (0.5793161, 'reported'), (0.5773655, 'transit'), (0.5720902, 'refugee'), (0.5698954, 'reports')]
The contribution of the word 'residents' in dimension 292 = 0.315702
Following are the top words in dimension 292 along with their contributions
[(0.5362592, 'population'), (0.5191565, 'scoreboard'), (0.49917164, 'populations'), (0.49770728, 'occupancy'), (0.48458424, 'tents'), (0.48447505, 'rainfall'), (0.4806576, 'footage'), (0.47326884, 'sporadic'), (0.47142318, 'incorporates'), (0.46291003, 'densely')]
The contribution of the word 'residents' in dimension 618 = 0.305339
Following are the top words in dimension 618 along with their contributions
[(0.69593966, 'located'), (0.6945901, 'district'), (0.6708719, 'canton'), (0.6563801, 'unincorporated'), (0.6515098, 'oblast'), (0.64047897, 'suburb'), (0.6389987, 'diocese'), (0.6177583, 'uttar'), (0.6140993, 'pradesh'), (0.6064239, 'situated')]

intelligence	'brain', 'bone', 'genome', 'fractured', 'spinal', 'dinosaur', 'scan', 'rna', 'muscles', 'spectrum'
science	'data', 'equipment', 'mechanical', 'physical', 'computer', 'fluid', 'breathing' , 'mechanics', 'liquid', 'luggage'
residents	'war', 'language', 'military', 'students', 'political', 'his', 'her', 'troops', 'speaking', 'religious'

	Initial GloVe vectors	Initial word2vec vectors
mathematics	intelligence, government, foreign, security kashmir, algorithms, heat, computational robes, tito, aviation, backward, dioceses	leukemia, enterprises, wingspan, info, booker ore, greens, badminton, hymns, clay asylum, intercepted, skater, rb, flats
remote	thousands, residents, palestinian, police kashmir, algorithms, heat, computational tamil, guerrilla, spam, rebels, infantry	basilica, sensory, ranger, chapel, memorials microsoft, sr, malaysia, jan, cruisers capt, obey, tents, overdose, cognitive, flats
internet	thousands, residents, palestinian, police intelligence, government, foreign, security nhl, writer, writers, drama, soccer	cardinals, tsar, papal, autobiography, befriends gases, gov, methane, graph, buttons longitude, carr, precipitation, snowfall, homer

	SPINE w/ GloVe	SPINE w/ word2vec
mathematics	sciences, honorary, faculty, chemistry, bachelor university, professor, graduate, degree, bachelor mathematical, equations, theory, quantum	algebra, exam, courses, exams, math theorem, mathematical, mathematician, equations doctorate, professor, doctoral, lecturer, sociology
remote	territory, region, province, divided, district wilderness, ski, camping, mountain, hiking rugged, mountainous, scenic, wooded, terrain	villages, hamlet, villagers, village, huts mountainous, hilly, impoverished, poorest, populated button, buttons, click, password, keyboard
internet	windows, users, user, software, server youtube, myspace, twitter, advertising, ads wireless, telephone, cellular, cable, broadband	hacker, spam, pornographic, cyber, pornography browser, app, downloads, iphone, download cellular, subscriber, verizon, broadband, subscribers

Model 2

Spine embedding generated from 300-dim glove embeddings

- Encoder- 2 FC layer with 500 and 1000 hidden units resp
- decoder- 1 FC layer to reconstruct the originate embedding

Word similarity scores for SPINE embeddings of Glove 300 dimensions

Cosine similarity of embeddings Predicted by model	Similarity ratings by multiple Human annotators
0.9792468	6.77
0.98968876	7.35
1	10
0.98539751	7.46
0.93923787	7.62
0.99885787	7.58
0.99650181	5.77
0.97605417	7.5

0.99720642	6.77
0.99358125	7.42

Spearman_rank_coeff = 353 232 0.21573009095891893
sp_rho = 0.0009423928013857126

Sparsity

The sparsity is **99.45008666666666**

Analysis of SPINE embeddings generated from 300-dim GloVe embeddings

1)the word is intelligence

Top participating dimensions -[299,959,270,825,712]
Word of interest = intelligence
The contribution of the word 'intelligence' in dimension 299 = 0.526745
Following are the top words in dimension 299 along with their contributions
[(0.75183785, 'our'), (0.73675376, 'any'), (0.7353241, 'more'), (0.7280482, 'their'), (0.7262765, 'we'), (0.7250087, 'have'), (0.7135651, 'countries'), (0.71004903, 'be'), (0.7046013, 'not'), (0.7038128, 'are')]
The contribution of the word 'intelligence' in dimension 959 = 0.490835
Following are the top words in dimension 959 along with their contributions
[(0.78015846, 'district'), (0.7382973, 'county'), (0.73435134, 'town'), (0.6977316, 'village'), (0.6976719, 'south'), (0.69195265, 'located'), (0.69189113, 'central'), (0.69155616, 'north'), (0.6889391, 'province'), (0.68886137, 'northern')]
The contribution of the word 'intelligence' in dimension 270 = 0.362930
Following are the top words in dimension 270 along with their contributions
[(0.58780205, 'he'), (0.58105844, 'was'), (0.5774313, 'at'), (0.57718945, 'with'), (0.57685006, 'an'), (0.57182944, 'as'), (0.5694239, 'for'), (0.5682541, 'when'), (0.56812656, 'john'), (0.5680464, 'it')]
The contribution of the word 'intelligence' in dimension 825 = 0.213915
Following are the top words in dimension 825 along with their contributions
[(0.5952785, 'said'), (0.5767116, 'spokesman'), (0.52141607, 'analyst'), (0.50101936, 'maj.'), (0.4862661, 'told'), (0.48422074, 'col.'), (0.47403938, 'dr.'), (0.4713328, 'minister'), (0.4623903, 'director'), (0.46145836, 'economist')]
The contribution of the word 'intelligence' in dimension 712 = 0.047478
Following are the top words in dimension 712 along with their contributions

[(0.610988, 'cents'), (0.60761195, 'billion'), (0.58352315, 'dollars'), (0.55265296, 'pesos'), (0.5440571, 'percent'), (0.5419289, 'points'), (0.5412243, 'million'), (0.52373457, 'yen'), (0.5146968, 'francs'), (0.49386385, 'hk')]

2)the word is science

Top participating dimensions=[959,299,270,825,592]

Word of interest = science
The contribution of the word 'science' in dimension 959 = 0.503708
Following are the top words in dimension 959 along with their contributions
[(0.78015846, 'district'), (0.7382973, 'county'), (0.73435134, 'town'), (0.6977316, 'village'), (0.6976719, 'south'), (0.69195265, 'located'), (0.69189113, 'central'), (0.69155616, 'north'), (0.6889391, 'province'), (0.68886137, 'northern')]
The contribution of the word 'science' in dimension 299 = 0.446613
Following are the top words in dimension 299 along with their contributions
[(0.75183785, 'our'), (0.73675376, 'any'), (0.7353241, 'more'), (0.7280482, 'their'), (0.7262765, 'we'), (0.7250087, 'have'), (0.7135651, 'countries'), (0.71004903, 'be'), (0.7046013, 'not'), (0.7038128, 'are')]
The contribution of the word 'science' in dimension 270 = 0.395903
Following are the top words in dimension 270 along with their contributions
[(0.58780205, 'he'), (0.58105844, 'was'), (0.5774313, 'at'), (0.57718945, 'with'), (0.57685006, 'an'), (0.57182944, 'as'), (0.5694239, 'for'), (0.5682541, 'when'), (0.56812656, 'john'), (0.5680464, 'it')]
The contribution of the word 'science' in dimension 825 = 0.136427
Following are the top words in dimension 825 along with their contributions
[(0.5952785, 'said'), (0.5767116, 'spokesman'), (0.52141607, 'analyst'), (0.50101936, 'maj.'), (0.4862661, 'told'), (0.48422074, 'col.'), (0.47403938, 'dr.'), (0.4713328, 'minister'), (0.4623903, 'director'), (0.46145836, 'economist')]
The contribution of the word 'science' in dimension 592 = 0.112731
Following are the top words in dimension 592 along with their contributions
[(0.8561978, 'genus'), (0.8037467, 'enzyme'), (0.7873987, 'protein'), (0.78626156, 'vector'), (0.77516246, 'snail'), (0.7540237, 'fungi'), (0.75305533, 'proteins'), (0.7450343, 'enzymes'), (0.7378669, 'subspecies'), (0.73251426, 'dorsal')]

3)the word is residents
Top participating dimensions=[959,299,270,825,712]

Word of interest = residents
The contribution of the word 'residents' in dimension 959 = 0.542276

Following are the top words in dimension 959 along with their contributions
[(0.78015846, 'district'), (0.7382973, 'county'), (0.73435134, 'town'), (0.6977316, 'village'), (0.6976719, 'south'), (0.69195265, 'located'), (0.69189113, 'central'), (0.69155616, 'north'), (0.6889391, 'province'), (0.68886137, 'northern')]
The contribution of the word 'residents' in dimension 299 = 0.482383
Following are the top words in dimension 299 along with their contributions
[(0.75183785, 'our'), (0.73675376, 'any'), (0.7353241, 'more'), (0.7280482, 'their'), (0.7262765, 'we'), (0.7250087, 'have'), (0.7135651, 'countries'), (0.71004903, 'be'), (0.7046013, 'not'), (0.7038128, 'are')]
The contribution of the word 'residents' in dimension 270 = 0.389425
Following are the top words in dimension 270 along with their contributions
[(0.58780205, 'he'), (0.58105844, 'was'), (0.5774313, 'at'), (0.57718945, 'with'), (0.57685006, 'an'), (0.57182944, 'as'), (0.5694239, 'for'), (0.5682541, 'when'), (0.56812656, 'john'), (0.5680464, 'it')]
The contribution of the word 'residents' in dimension 825 = 0.093840
Following are the top words in dimension 825 along with their contributions
[(0.5952785, 'said'), (0.5767116, 'spokesman'), (0.52141607, 'analyst'), (0.50101936, 'maj.'), (0.4862661, 'told'), (0.48422074, 'col.'), (0.47403938, 'dr.'), (0.4713328, 'minister'), (0.4623903, 'director'), (0.46145836, 'economist')]
The contribution of the word 'residents' in dimension 712 = 0.085793
Following are the top words in dimension 712 along with their contributions
[(0.610988, 'cents'), (0.60761195, 'billion'), (0.58352315, 'dollars'), (0.55265296, 'pesos'), (0.5440571, 'percent'), (0.5419289, 'points'), (0.5412243, 'million'), (0.52373457, 'yen'), (0.5146968, 'francs'), (0.49386385, 'hk')]

intelligence	'our', 'any', 'more, 'their', 'we', 'have', , 'countries', 'be', 'not', 'are'
science	'district', 'county', 'town', 'village', 'south', ‘‘located', 'central', 'north', 'province', 'northern'
residents	'district', 'county', 'town', 'village', 'south','located', 'central', 'north', 'province', 'northern'

CONCLUSION

- The paper presents a novel mechanism to generate interpretable word embeddings.
- The model shows best results on GloVe embeddings of 300 dimensions.
- The model works best with one hidden layer in the autoencoder. It does not give good results(sparsity increases to 99%) when we added a new hidden layer. Since the loss is propagating twice and ASL and PSL penalised the neurons twice, most of the activations become 0 leading to more sparsity resulting information loss.
- The model does not give very good results on BERT embeddings because unlike GloVe and Word2Vec embeddings, BERT embeddings are contextual and more complex.
- The model does not work well on very dense embeddings like 100 dimensional GloVe due to information loss while transforming from 100 to 1000 dimensions.