Neural Network Architectures for Short Text

Dylan Elliott
November 14, 2017

Rensselaer

Agenda

- Introduction
- Background
- Related Work
- Models Compared
- Experiments Performed
- Experimental Comparison
- Conclusion and Future Work

Introduction

- **Introduction**
- Background
- Related Work
- Models Compared
- Experiments Performed
- Experimental Comparison
- Conclusion and Future Work

1. Motivation
2. Purpose
3. Contributions
4. Experiment Overview

Motivation

- Neural networks are end-to-end trainable.
- Language is a catalyst for learning and discovery.
- Large language barrier between humans and machines!

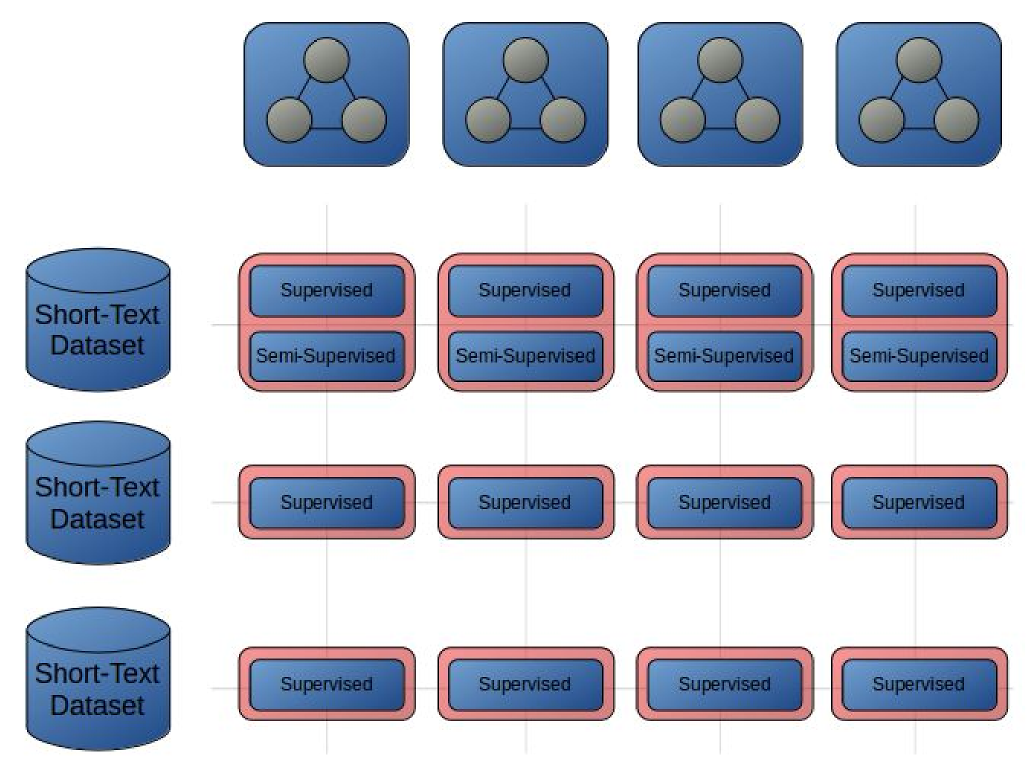
Purpose

- Investigate neural network models used for short text data.
- Gather insight into relative strengths and shortcomings of different models.
- Apply insight to larger integrated systems.

Contributions

- Comparison of four neural network models:
  - Three short-text classification datasets (sample sentence, numerical label)
  - Supervised learning task
  - Semi-Supervised learning task
- Model behavior and analysis:
  - Visualization of learned representations
  - Clustering learned representations
  - Comparison to traditional text representations

Comparison Overview



Background

- Introduction
- **Background**
- Related Work
- Models Compared
- Experiments Performed
- Experimental Comparison
- Conclusion and Future Work

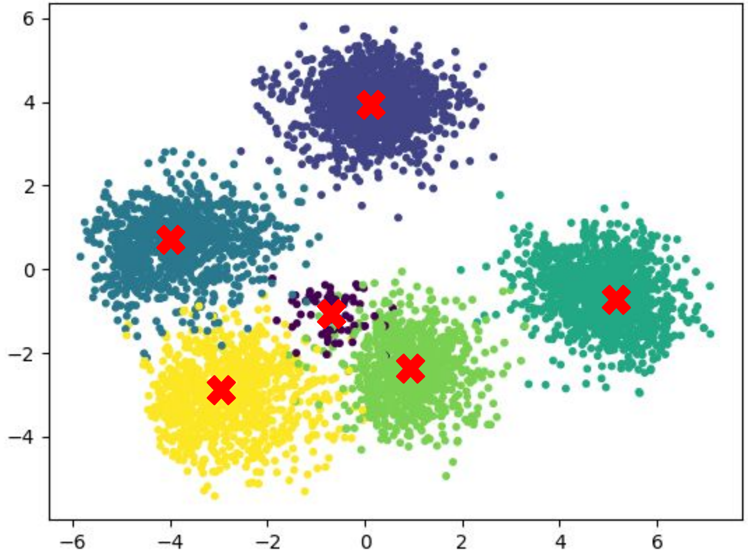
1. NLU
2. Clustering
3. Neural Networks
4. Bag of Words
5. TF-IDF
6. Word Embeddings
7. Features in Text

Natural Language Understanding (NLU)

- Machine comprehension of language as it normally appears to humans.
  - Text
  - Speech
- Challenges:
  - Language is dynamic
  - Hard to define rule-based system
  - Machines accept numerical values

Clustering

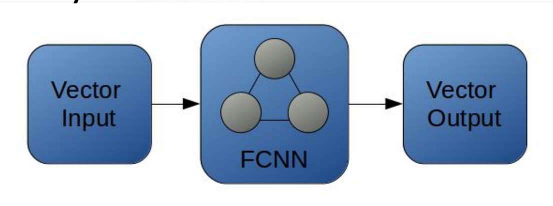
- Search for *K* separable groups in categorical data.
- Unsupervised
- K-Means algorithm:
  - Prior choice of *K*
  - Representative



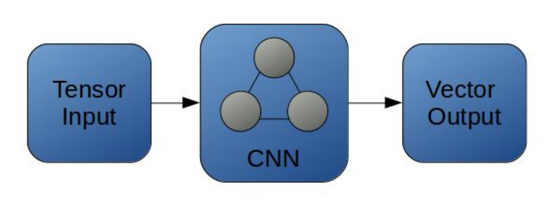
Neural Networks

- Network of linear/nonlinear processing units.
- End-to-end trainable with gradient descent optimization.

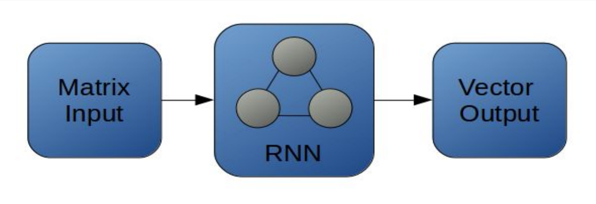
Fully Connected Neural Network



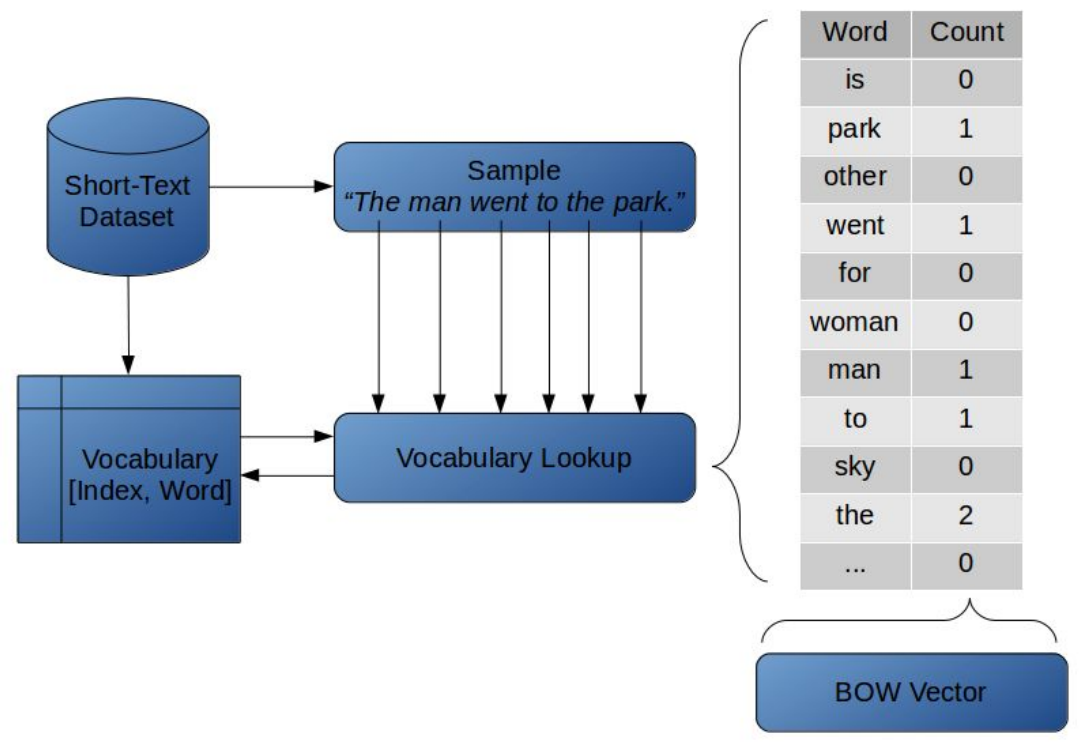
Convolutional Neural Network



Recurrent Neural Network



Text Representation: Bag of Words



Word	Count
is	0
park	1
other	0
went	1
for	0
woman	0
man	1
to	1
sky	0
the	2
...	0

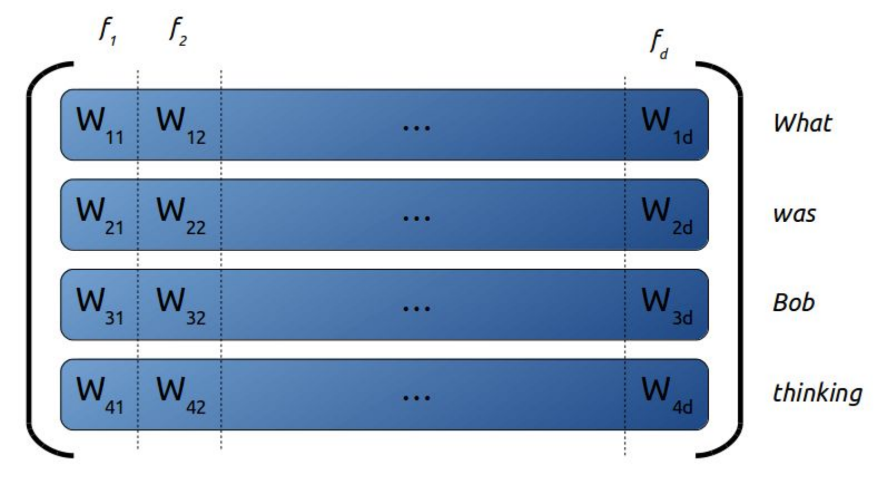
Text Representation: TF-IDF

- Term Frequency: # of occurrences of a word in one text sample.
- Document Frequency: # of occurrences of a word throughout all text samples.

$$w_i = t_i \log \frac{N}{d_i} \qquad \forall i = 1 \ldots |V|$$

Text Representation: Word Vectors

- Words -> word vectors
- Text sequence -> word embedding matrix



Features in Text

- Syntactic Features:
  - Part-of-speech (POS) tags
  - Chunks of POS patterns
- Semantic Features:
  - Topic
  - Named Entities
  - Sentiment
  - Intent
  - Question-Type, Answer-Type and Modifiers

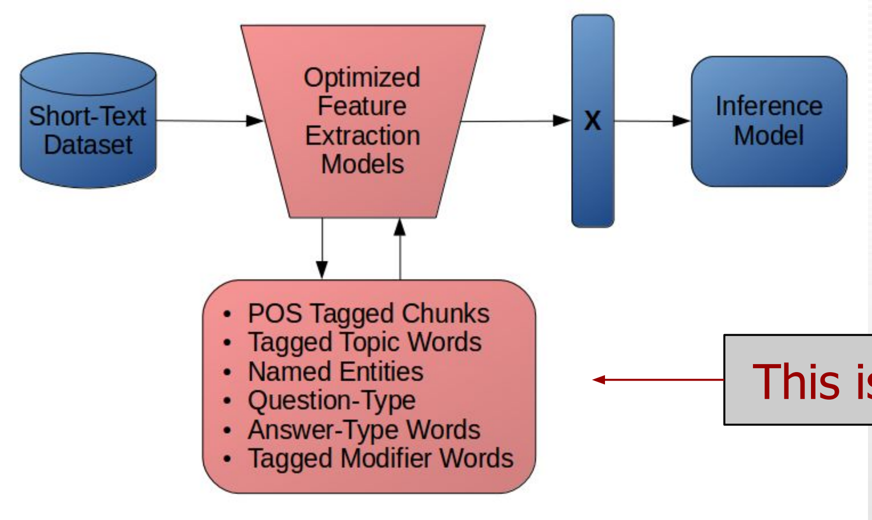
Related Work

- Introduction
- Background
- **Related Work**
- Models Compared
- Experiments Performed
- Experimental Comparison
- Conclusion and Future Work

1. Text Enrichment
2. Neural Networks for Text Data

Text Enrichment

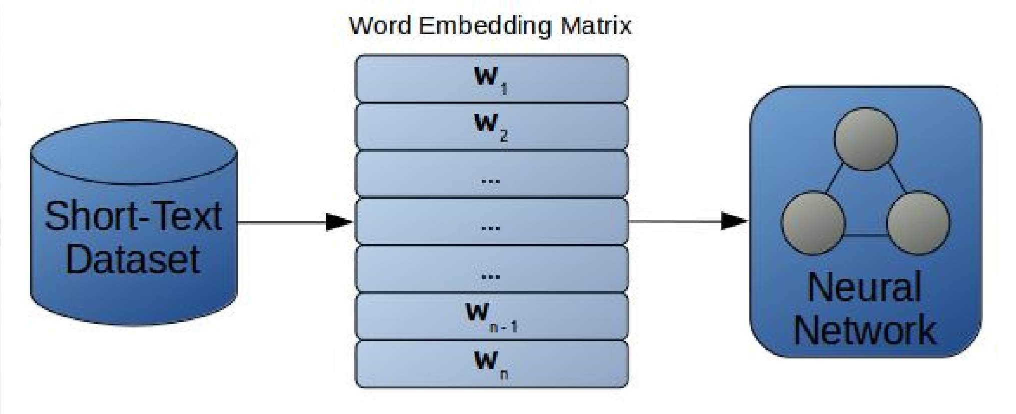
- Convert text to a feature vector **X** through text enrichment.



This is a lot of overhead!

Neural Networks for Text

- Automatically learn features relevant to task.
- Equal or better performance than text enrichment models.

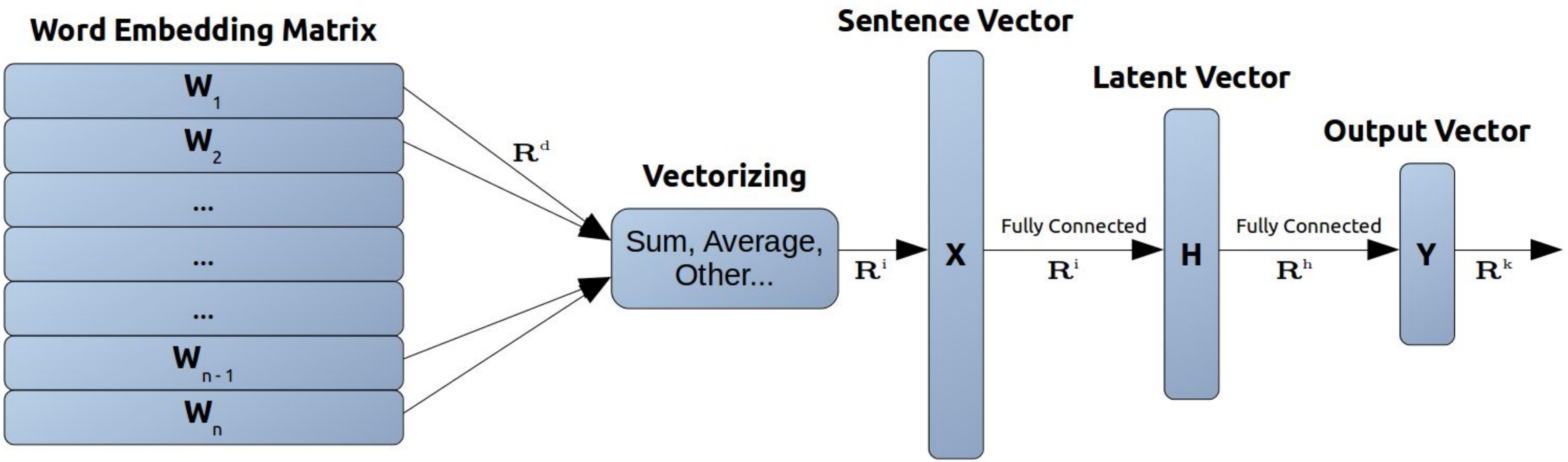


Models Compared

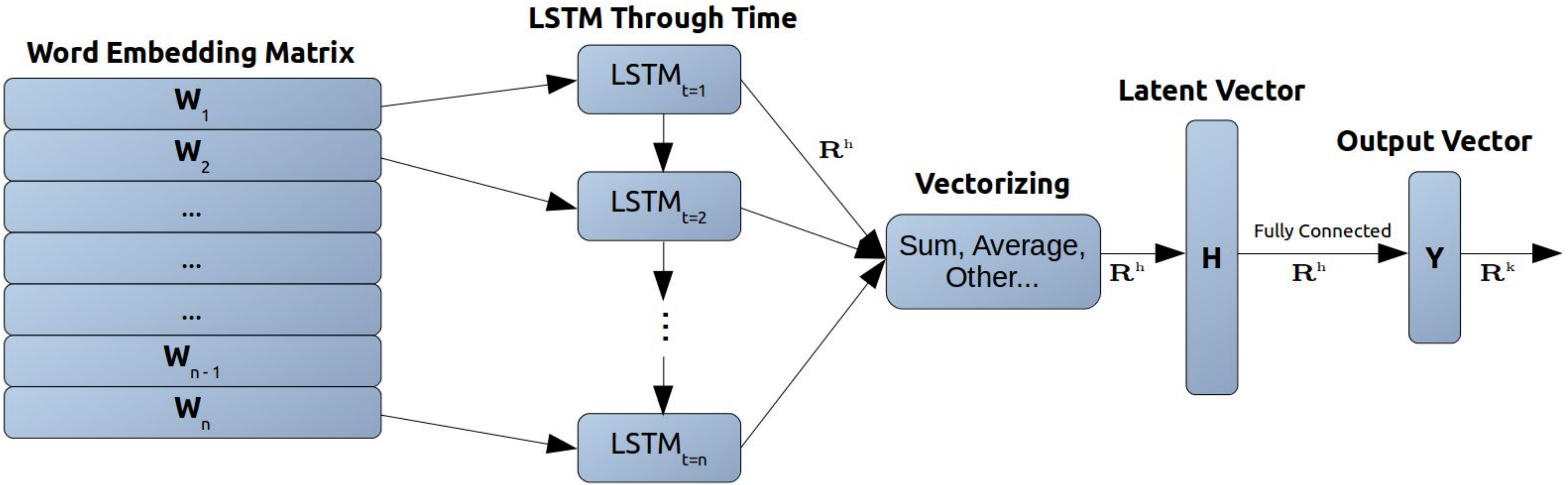
- Introduction
- Background
- Related Work
- **Models Compared**
- Experiments Performed
- Experimental Comparison
- Conclusion and Future Work

1. NBOW
2. LSTM
3. TCNN
4. DCNN

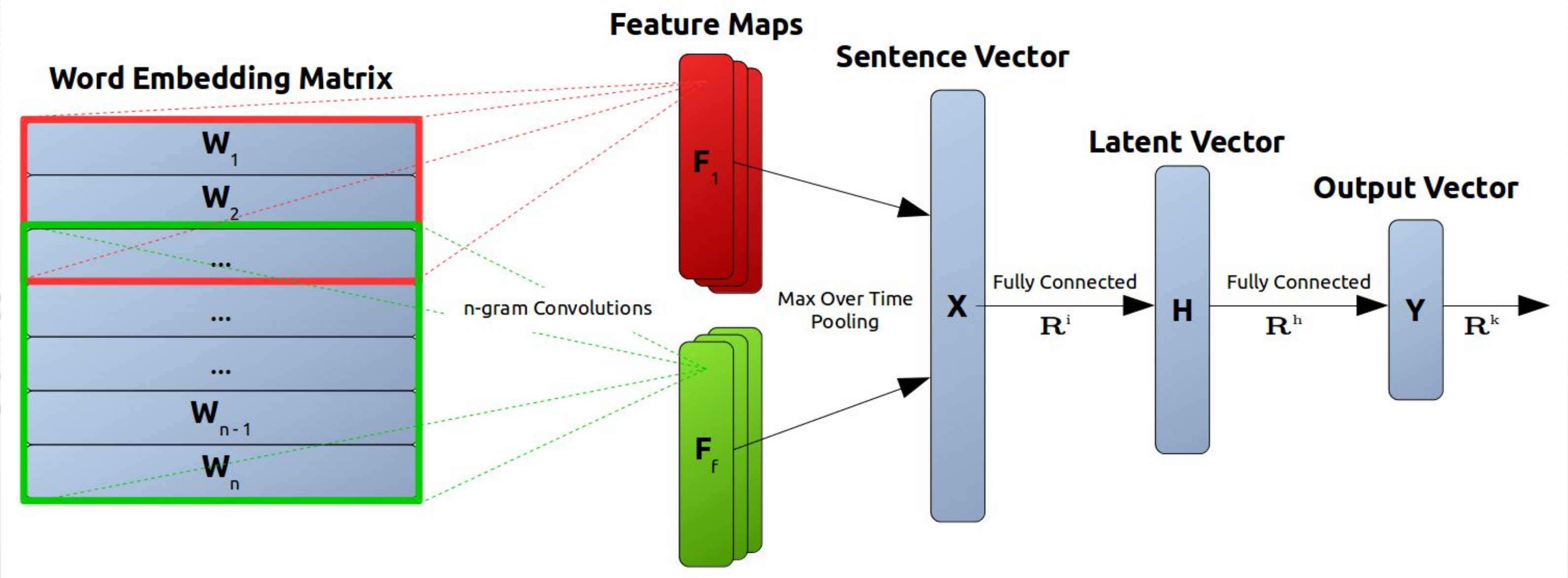
Neural Bag of Words (NBOW) Model



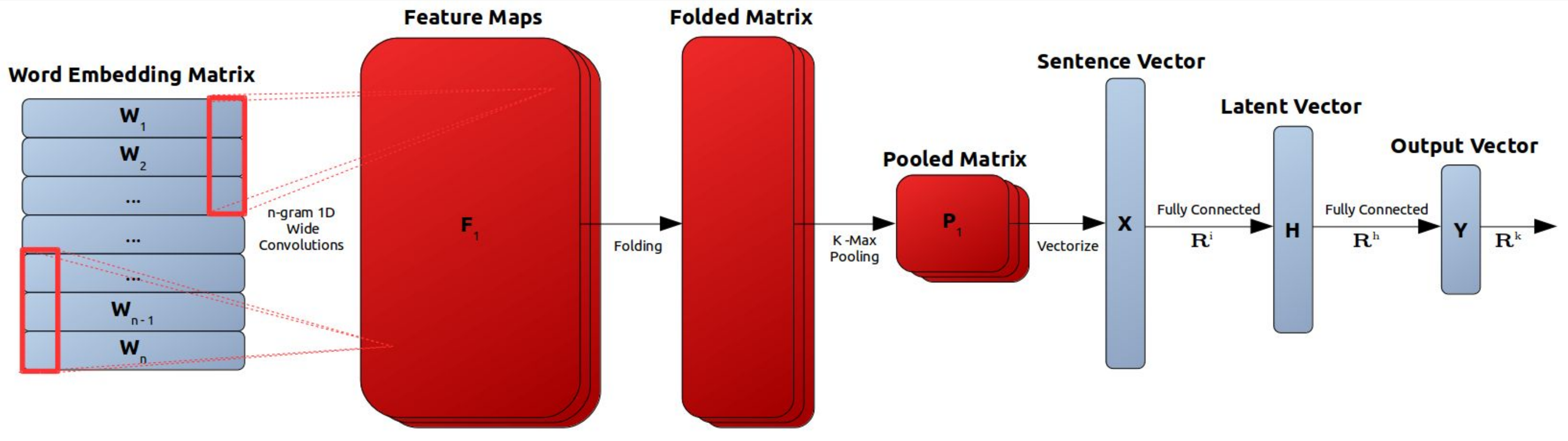
Long Short-Term Memory (LSTM) Model



Max-Over-Time Convolutional Neural Network (TCNN) Model



Dynamic Convolutional Neural Network (DCNN) Model



Experiments Performed

- Introduction
- Background
- Related Work
- Models Compared
- **Experiments Performed**
- Experimental Comparison
- Conclusion and Future Work

1. Supervised Classification Task
2. Semi-Supervised Learning Task
3. Clustering Learned Text Representations

Supervised Classification Task

- Train each model on each dataset with some hyperparameters held constant.
- Softmax output for *K* labels with cross entropy loss.

$$l_i = -\sum_{j}^{k} y_{ij} log(\hat{y}_{ij})$$

Hyperparameter	Value
d	300
η	1×10^{-3}
h	100
$P(keep)$	0.5

Semi-Supervised Learning Task[1]

- Learn latent vector representations from a K-means inspired objective.
- Pre-train model with 10% labeled samples.

$$J = \alpha \sum_{i}^{N} \sum_{j}^{k} r_{ij}\delta_{ij} + (1-\alpha)\sum_{i}^{L}\{\delta_{ig_i} + \sum_{l \neq g_i} \max(m + \delta_{ig_i} - \delta_{il}, 0)\}$$

where,

$$\delta_{ij} = ||f(x_i) - \mu_j||^2$$

Hyperparameter	Value
d	300
h	100
$P(keep)$	0.5

[1] Semi-supervised Clustering for Short Text via Deep Representation Learning (Wang, Mi, Ittycheria - 2017)

Semi-Supervised Learning Task

Parameter	Description
alpha	Weighting to control influence of labeled data. Lower *alpha* = More influence of labeled data.
r_{ij}	Cluster assignments for all samples. 1 if sample i is assigned to cluster j, 0 otherwise. $\mathbf{R}^{N \times K}$ matrix stores all r_{ij} values for a dataset with N samples and K unique labels.
μ_j	h dimensional centroid for cluster j.
g_i	Mapping from truth label i to cluster label g_i

Clustering Learned Text Representations

- Perform K-means on learned latent representations with *K* centroids.
- External cluster evaluation using dataset labels (F-Measure and Adjusted Mutual Info).

$$F = \frac{1}{K}\sum_{i}^{K}\frac{2p_i r_i}{p_i + r_i}$$

$$AMI(\mathcal{C},\mathcal{T}) = \frac{I(\mathcal{C},\mathcal{T}) - E[I(\mathcal{C},\mathcal{T})]}{\max(H(\mathcal{C}),H(\mathcal{T})) - E[I(\mathcal{C},\mathcal{T})]}$$

Experimental Comparison

- Introduction
- Background
- Related Work
- Models Compared
- Experiments Performed
- **Experimental Comparison**
- Conclusion and Future Work

1. Datasets
2. Supervised Classification Results
3. Semi-Supervised Learning Results
4. Clustering Results

Short Text Datasets

Question-Type[2]

- Answer categories for sample questions
- K = 6 classes
- (Abbreviation, Entity, Description, Human, Location, Number)

StackOverflow[3]

- Programming topic categories for StackOverflow queries
- K = 20 classes
- (matlab, bash, apache, excel, etc.)

AG-News[4]

- General news topic categories for news titles
- K = 4 classes
- (World, Sports, Business, Sci/Tech)

[2] Learning Question Classifiers (Li, Roth - 2002)
[3] Short Text Clustering via Convolutional Neural Networks (Xu, Wang, Tian, Zhao, Wang, Hao - 2015)
[4] Character-level Convolutional Networks for Text Classification (Zhang, Zhao, LeCun - 2015)

Dataset Statistics

Dataset	Question-Type	StackOverflow	AG-News
N	5,952	20,000	127,600
N_{train}	5,452	16,000	120,000
N_{test}	500	4,000	7,600
n_{max}	39	36	20
n_{avg}	10	8	6
$\|V\|$	8,983	18,927	50,627

Supervised Classification Results: Testing Set Accuracy

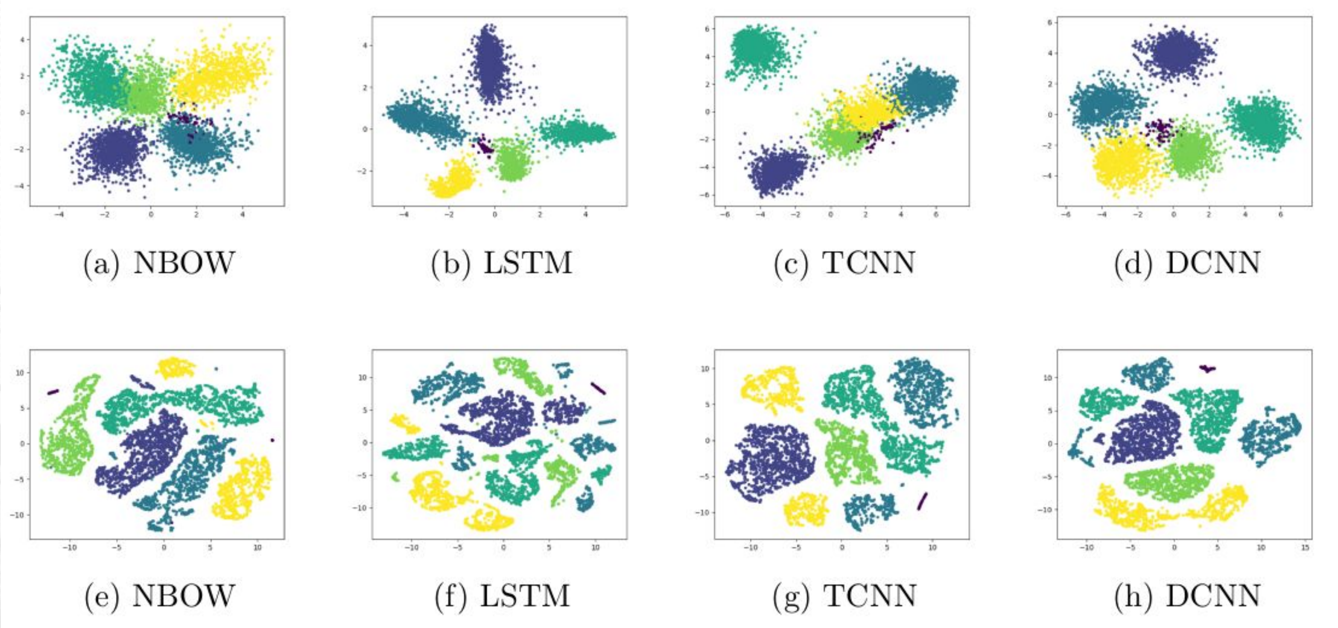
Model/Dataset	Question-Type	StackOverflow	AG-News
NBOW	86.36 ± 0.43	$\mathbf{85.27 \pm 0.05}$	84.62 ± 0.15
LSTM	$\mathbf{87.48 \pm 0.52}$	76.20 ± 0.50	84.26 ± 0.14
TCNN	$\mathbf{88.60 \pm 0.66}$	84.65 ± 0.17	$\mathbf{84.78 \pm 0.29}$
DCNN	86.04 ± 0.50	$\mathbf{85.38 \pm 0.26}$	$\mathbf{85.59 \pm 0.41}$

Supervised Classification Results: Intra/Inter Neighbor Separation

Model/Dataset	Question-Type	StackOverflow	AG-News
NBOW	1.79/4.89/3.10	2.56/6.60/4.04	3.73/10.66/6.93
LSTM	1.36/6.37/5.01	2.45/6.11/3.66	2.22/4.12/1.90
TCNN	2.31/10.33/8.02	3.10/11.74/8.64	4.73/14.44/9.71
DCNN	2.11/8.28/6.17	2.90/9.95/7.05	3.78/13.84/10.06

(intra-neighbor/inter-neighbor/difference margin)

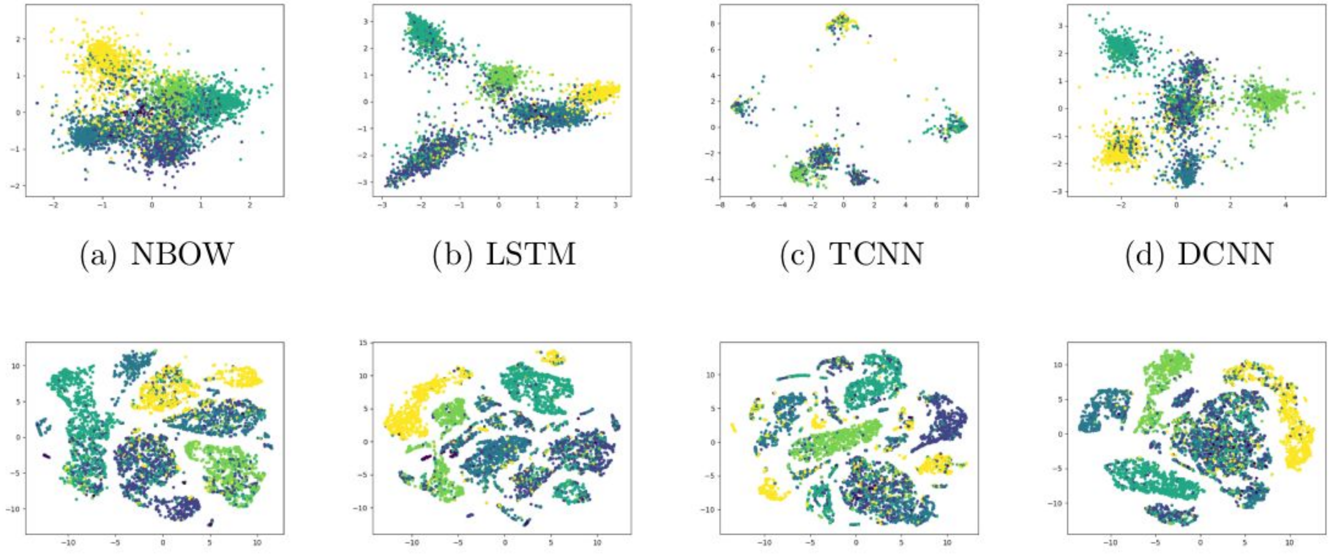
Supervised Classification Results: Question-Type Visualizations



(a) NBOW (b) LSTM (c) TCNN (d) DCNN

(e) NBOW (f) LSTM (g) TCNN (h) DCNN

Semi-Supervised Learning Results: Question Type Visualizations



(a) NBOW (b) LSTM (c) TCNN (d) DCNN

(e) NBOW (f) LSTM (g) TCNN (h) DCNN

Clustering Results: Semi-Supervised Learned Representations on Question-Type

Model	AMI Pre-Train Only	F-Measure Pre-Train Only	AMI Full-Task	F-Measure Full-Task
NBOW	0.425 ± 0.005	0.593 ± 0.014	0.441 ± 0.009	$\mathbf{0.706 \pm 0.025}$
LSTM	0.460 ± 0.007	0.641 ± 0.007	0.492 ± 0.017	$\mathbf{0.699 \pm 0.020}$
TCNN	0.453 ± 0.010	0.632 ± 0.008	0.432 ± 0.010	0.621 ± 0.017
DCNN	0.420 ± 0.012	0.607 ± 0.009	0.433 ± 0.014	0.566 ± 0.025

Representation	AMI	F-Measure
BOW	0.140	0.306
TF-IDF	**0.157**	**0.375**

Conclusion and Future Work

- Introduction
- Background
- Related Work
- Models Compared
- Experiments Performed
- Experimental Comparison
- **Conclusion and Future Work**

1. Observations
2. Insights
3.

Observations

- Classification: at least one CNN-based model with top performance.
- Classification: LSTM seems to struggle with a large amount of labels
- Semi-supervised learning: models that achieve less neighbor separation generally result in in better clustering

Results Suggest…

- Less decisive -> better ability to correct mistakes during further learning.
- Performance is dependent on both model architecture and dataset characteristics.
- Word vector utilization varies.
- LSTM: whole sentence level features.
- CNN: single/multiple word level features.

Future Work: Pre-Trained Word Vectors

- Pre-training word vectors can increase performance.
- Quantify the utilization of word vectors depending on model architecture.

Future Work: IBM HEALS

- Health Empowerment by Analytics, Learning and Semantics.
- Chat-bot health informative framework.
- Subsystem: organize user queries into *intent* groups.
- Use groupings to help construct chat-bot dialog tree.

Future Work: Alternative Models

- Autoencoders with adversarial learning[5]:
  - Latent representations can be constrained by a classification distribution for semi-supervised learning.
- "Siamese" network architectures[6]:
  - Pair-wise similarity learning increases training size and simplifies labeling of data.
  - Features extracted by comparative learning

[5] Adversarial Autoencoders (Makhzani, Shlens, Jaitly, Goodfellow - 2016)
[6] Signature Verification Using a "Siamese" Time-delay Neural Network (Bromley, Guyon, LeCun, Säckinger, Shah - 1994)

Thank You!

- Questions?