From Image to Text Classification: A Novel Approach based on Clustering Word Embeddings

Andrei M. Butnaru, Radu T. Ionescu

butnaruandreimadalin@gmail.com  raducu.ionescu@gmail.com
Department of Computer Science
University of Bucharest
Bucharest, Romania

September 2017
Marseille, France
Table of Contents

- Bag of Visual Words Model
Table of Contents

• Bag of Visual Words Model
• Word Embeddings
Table of Contents

- Bag of Visual Words Model
- Word Embeddings
- Bag of Super Word Embeddings
Table of Contents

• Bag of Visual Words Model
• Word Embeddings
• Bag of Super Word Embeddings
• Training and Classification
Table of Contents

• Bag of Visual Words Model
• Word Embeddings
• Bag of Super Word Embeddings
• Training and Classification
• Polarity Classification Experiments
Table of Contents

• Bag of Visual Words Model
• Word Embeddings
• Bag of Super Word Embeddings
• Training and Classification
• Polarity Classification Experiments
• Text Classification Experiments
Table of Contents

• Bag of Visual Words Model
• Word Embeddings
• Bag of Super Word Embeddings
• Training and Classification
• Polarity Classification Experiments
• Text Classification Experiments
• Conclusions and Future Work
Outline

• Bag of Visual Words Model
• Word Embeddings
• Bag of Super Word Embeddings
• Training and Classification
• Polarity Classification Experiments
• Text Classification Experiments
• Conclusions and Future Work
Bag of Visual Words Model

- A *bag of words* model represents a text as a collection of words.
A bag of words model represents a text as a collection of words.

A bag of visual words uses vector quantized local image descriptors as words.
A bag of words model represents a text as a collection of words.

A bag of visual words uses vector quantized local image descriptors as words.

The bag of visual words model can be described as a histogram of visual words.
Outline

• Bag of Visual Words Model
• Word Embeddings
• Bag of Super Word Embeddings
• Training and Classification
• Polarity Classification Experiments
• Text Classification Experiments
• Conclusions and Future Work
Word Embeddings

• A low-dimensional real valued vector representation of a word
Word Embeddings

• A low-dimensional real valued vector representation of a word
• Semantically related words are in closed vicinity in the generated space
Word Embeddings

- A low-dimensional real valued vector representation of a word
- Semantically related words are in closed vicinity in the generated space
- Documents can be represented as a set of word vectors
Word Embeddings

• A low-dimensional real valued vector representation of a word
• Semantically related words are in closed vicinity in the generated space
• Documents can be represented as a set of word vectors
• Word vectors can be clustered to obtain a histogram representation of a document
Outline

- Bag of Visual Words Model
- Word Embeddings
- Bag of Super Word Embeddings
- Training and Classification
- Polarity Classification Experiments
- Text Classification Experiments
- Conclusions and Future Work
Bag of Super Word Embeddings

- Image descriptors are replaced by word embeddings
Bag of Super Word Embeddings

- Image descriptors are replaced by word embeddings
- Cluster word embeddings to obtain relevant semantic clusters of words
Bag of Super Word Embeddings

- Image descriptors are replaced by word embeddings
- Cluster word embeddings to obtain relevant semantic clusters of words
- The centroid of a semantic cluster can be viewed as a super word vector
Bag of Super Word Embeddings

- Image descriptors are replaced by word embeddings
- Cluster word embeddings to obtain relevant semantic clusters of words
- The centroid of a semantic cluster can be viewed as a super word vector
- A vocabulary is formed based on all super word vectors obtained from a document
Bag of Super Word Embeddings

- Image descriptors are replaced by word embeddings
- Cluster word embeddings to obtain relevant semantic clusters of words
- The centroid of a semantic cluster can be viewed as a super word vector
- A vocabulary is formed based on all super word vectors obtained from a document
- A document is described as a histogram of super word embeddings
Bag of Super Word Embeddings

A. M. Butnaru, R. T. Ionescu

From Image to Text Classification KES 2017
Bag of Super Word Embeddings

• Three major steps:
Bag of Super Word Embeddings

- Three major steps:
  - **Step 1** Build a feature representation
Bag of Super Word Embeddings

• Three major steps:
  • Step 1 Build a feature representation
  • Step 2 Train a kernel method
Bag of Super Word Embeddings

- Three major steps:
  - **Step 1** Build a feature representation
  - **Step 2** Train a kernel method
  - **Step 3** Prediction
Bag of Super Word Embeddings

• Three major steps:
  • **Step 1** Build a feature representation
  • Step 2 Train a kernel method
  • Step 3 Prediction
• Features are represented by the word embeddings extracted from text
BOSWE - Feature Representation

• Features are represented by the word embeddings extracted from text
• Word embeddings are vector quantized in order to generate a vocabulary of super word embeddings
BOSWE - Feature Representation

• Features are represented by the word embeddings extracted from text
• Word embeddings are vector quantized in order to generate a vocabulary of super word embeddings
• Words are assigned to the closest centroid
• Features are represented by the word embeddings extracted from text
• Word embeddings are vector quantized in order to generate a vocabulary of super word embeddings
• Words are assigned to the closest centroid
• The frequencies for each super word embeddings is recorded in a histogram
BOSWE - Feature Representation

- Two alternative pipelines:
  - Pipeline 1: Process the entire collection of documents all at once. Features are represented by the histogram based on the generated vocabulary.
  - Pipeline 2: Group the training documents into classes and process each group separately. Features are represented by the concatenation of all the histograms corresponding to class-specific vocabularies.
• Two alternative pipelines:
  • **Pipeline 1** Process the entire collection of documents all at once
BOSWE - Feature Representation

• Two alternative pipelines:
  • **Pipeline 1** Process the entire collection of documents all at once
  • Features are represented by the histogram based on the generated vocabulary
Two alternative pipelines:

- **Pipeline 1** Process the entire collection of documents all at once
- Features are represented by the histogram based on the generated vocabulary
- **Pipeline 2** Group the training documents into classes and process each group separately.
Two alternative pipelines:

- **Pipeline 1** Process the entire collection of documents all at once
  - Features are represented by the histogram based on the generated vocabulary
- **Pipeline 2** Group the training documents into classes and process each group separately.
  - Features are represented by the concatenation of all the histograms corresponding to class-specific vocabularies
Bag of Super Word Embeddings

• Three major steps:
  • Step 1 Build a feature representation
  • **Step 2** Train a kernel method
  • Step 3 Classification
BOSWE - Kernel Method

- Used kernel methods:
• Used kernel methods:
  • linear kernel
• Used kernel methods:
  • linear kernel
  • intersection kernel
BOSWE - Kernel Method

- Used kernel methods:
  - linear kernel
  - intersection kernel
  - Hellinger’s kernel
BOSWE - Kernel Method

- Used kernel methods:
  - linear kernel
  - intersection kernel
  - Hellinger’s kernel
  - Jensen-Shannon kernel
BOSWE - Kernel Method

- Used kernel methods:
  - linear kernel
  - intersection kernel
  - Hellinger’s kernel
  - Jensen-Shannon kernel
  - PQ kernel [R. T. Ionescu and M. Popescu, ICIAP 2013]
BOSWE - Kernel Method

- Used kernel methods:
  - linear kernel
  - intersection kernel
  - Hellinger’s kernel
  - Jensen-Shannon kernel
  - PQ kernel [R. T. Ionescu and M. Popescu, ICIAP 2013]

- Combine kernel functions with Support Vector Machines
Bag of Super Word Embeddings

• Three major steps:
  • Step 1 Build a feature representation
  • Step 2 Train a kernel method
  • **Step 3 Prediction**
BOSWE - Prediction

- Features are extracted and the vocabulary is formed
BOSWE - Prediction

- Features are extracted and the vocabulary is formed
- The learned SVM weights are multiplied with the test histogram
BOSWE - Prediction

- Features are extracted and the vocabulary is formed
- The learned SVM weights are multiplied with the test histogram
- The system can return:
  - a label (or score)
  - a ranked list of results
BOSWE - Prediction

• Features are extracted and the vocabulary is formed
• The learned SVM weights are multiplied with the test histogram
• The system can return:
  • a label (or score)
BOSWE - Prediction

- Features are extracted and the vocabulary is formed
- The learned SVM weights are multiplied with the test histogram
- The system can return:
  - a label (or score)
  - a ranked list of results
Outline

- Bag of Visual Words Model
- Word Embeddings
- Bag of Super Word Embeddings
- Training and Classification
- Polarity Classification Experiments
- Text Classification Experiments
- Conclusions and Future Work
Polarity Classification Experiments

- Movie Review data set [Pang et al., EMNLP 2002]
Polarity Classification Experiments

• Movie Review data set [Pang et al., EMNLP 2002]
• 2000 movie reviews (1000 positive and 1000 negative)
Polarity Classification Experiments

- Movie Review data set [Pang et al., EMNLP 2002]
- 2000 movie reviews (1000 positive and 1000 negative)
- Baseline: *bag of words*
Polarity Classification Experiments

- Movie Review data set [Pang et al., EMNLP 2002]
- 2000 movie reviews (1000 positive and 1000 negative)
- Baseline: *bag of words*
- Evaluation is performed using a 10-fold cross-validation
Polarity Classification Experiments

Table: Accuracy rates using 10-fold cross-validation on the Moview Review data set

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>Lin. ((L_2))</th>
<th>Hel. ((L_1))</th>
<th>Int ((L_1))</th>
<th>JS ((L_1))</th>
<th>PQ ((L_2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 × 5000 words</td>
<td>84.80%</td>
<td>86.15%</td>
<td>85.40%</td>
<td>85.80%</td>
<td>86.55%</td>
</tr>
<tr>
<td>1 × 10000 words</td>
<td>85.05%</td>
<td>86.45%</td>
<td>85.75%</td>
<td>86.10%</td>
<td>87.15%</td>
</tr>
<tr>
<td>2 × 5000 words</td>
<td>85.75%</td>
<td>87.60%</td>
<td>86.95%</td>
<td>87.35%</td>
<td>88.25%</td>
</tr>
<tr>
<td>2 × 7500 words</td>
<td>87.15%</td>
<td>88.60%</td>
<td>88.15%</td>
<td>87.80%</td>
<td>88.95%</td>
</tr>
</tbody>
</table>
Polarity Classification Experiments

Table: Accuracy rates using 10-fold cross-validation on the Moview Review data set

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>Lin. ($L_2$)</th>
<th>Hel. ($L_1$)</th>
<th>Int ($L_1$)</th>
<th>JS ($L_1$)</th>
<th>PQ ($L_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 × 5000 words</td>
<td>84.80%</td>
<td>86.15%</td>
<td>85.40%</td>
<td>85.80%</td>
<td>86.55%</td>
</tr>
<tr>
<td>1 × 10000 words</td>
<td>85.05%</td>
<td>86.45%</td>
<td>85.75%</td>
<td>86.10%</td>
<td>87.15%</td>
</tr>
<tr>
<td>2 × 5000 words</td>
<td>85.75%</td>
<td>87.60%</td>
<td>86.95%</td>
<td>87.35%</td>
<td>88.25%</td>
</tr>
<tr>
<td>2 × 7500 words</td>
<td>87.15%</td>
<td>88.60%</td>
<td>88.15%</td>
<td>87.80%</td>
<td>88.95%</td>
</tr>
</tbody>
</table>

- Building a vocabulary for each polarity class is better
### Polarity Classification Experiments

Table: Accuracy rates using 10-fold cross-validation on the Moview Review data set with various BOSWE configuration versus two baseline approaches.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline BOW</td>
<td>84.10%</td>
</tr>
<tr>
<td>Pang et al. [EMNLP, 2002]</td>
<td>82.90%</td>
</tr>
<tr>
<td>BOSWE (2 × 7500 words and Hellinger’s kernel)</td>
<td>88.60%</td>
</tr>
<tr>
<td>BOSWE (2 × 7500 words and PQ kernel)</td>
<td>88.95%</td>
</tr>
<tr>
<td>BOSWE (2 × 7500 words and Hellinger’s kernel + PQ kernel)</td>
<td><strong>89.65%</strong></td>
</tr>
</tbody>
</table>
Outline

• Bag of Visual Words Model
• Word Embeddings
• Bag of Super Word Embeddings
• Training and Classification
• Polarity Classification Experiments
• Text Classification Experiments
• Conclusions and Future Work
Text Categorization Experiments

• Reuters-21578 [D. Lewis, SIGIR 1999]
Text Categorization Experiments

- Reuters-21578 [D. Lewis, SIGIR 1999]
- 10787 documents belonging in a total of 90 categories
Text Categorization Experiments

- Reuters-21578 [D. Lewis, SIGIR 1999]
- 10787 documents belonging in a total of 90 categories
- Baseline: *bag of words* adapted for text categorization by topic
Text Categorization Experiments

- Reuters-21578 [D. Lewis, SIGIR 1999]
- 10787 documents belonging in a total of 90 categories
- Baseline: *bag of words* adapted for text categorization by topic
- Evaluation is performed using the $F_1$ micro and macro score
Text Categorization Experiments

Table 1: Micro $F_1$ accuracy rates on the Reuters-21578 test set with different kernels and vocabulary dimension.

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>Lin. ($L_2$)</th>
<th>Hel. ($L_1$)</th>
<th>Int. ($L_1$)</th>
<th>JS ($L_1$)</th>
<th>PQ ($L_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 × 10000 words</td>
<td>86.62%</td>
<td>86.56%</td>
<td>85.28%</td>
<td>86.30%</td>
<td>86.74%</td>
</tr>
<tr>
<td>1 × 20000 words</td>
<td>86.72%</td>
<td>86.61%</td>
<td>85.66%</td>
<td>86.35%</td>
<td>86.80%</td>
</tr>
<tr>
<td>90 × 100 words</td>
<td>86.77%</td>
<td>86.91%</td>
<td>86.25%</td>
<td>86.59%</td>
<td>86.84%</td>
</tr>
<tr>
<td>90 × 200 words</td>
<td>86.83%</td>
<td>87.04%</td>
<td>86.33%</td>
<td>86.74%</td>
<td>87.07%</td>
</tr>
</tbody>
</table>

Table 2: Macro $F_1$ accuracy rates on the Reuters-21578 test set with different kernels and vocabulary dimension.

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>Lin. ($L_2$)</th>
<th>Hel. ($L_1$)</th>
<th>Int. ($L_1$)</th>
<th>JS ($L_1$)</th>
<th>PQ ($L_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 × 10000 words</td>
<td>49.42%</td>
<td>45.21%</td>
<td>41.19%</td>
<td>43.30%</td>
<td>49.31%</td>
</tr>
<tr>
<td>1 × 20000 words</td>
<td>49.58%</td>
<td>45.39%</td>
<td>41.55%</td>
<td>43.54%</td>
<td>49.36%</td>
</tr>
<tr>
<td>90 × 100 words</td>
<td>49.63%</td>
<td>47.71%</td>
<td>42.50%</td>
<td>44.94%</td>
<td>49.49%</td>
</tr>
<tr>
<td>90 × 200 words</td>
<td>49.68%</td>
<td>47.75%</td>
<td>42.64%</td>
<td>45.06%</td>
<td>49.51%</td>
</tr>
</tbody>
</table>
Table: Accuracy rates on the Reutesrs-21578 test set with various BOSWE configurations versus a baseline BOW model.

<table>
<thead>
<tr>
<th>Method</th>
<th>microF&lt;sub&gt;1&lt;/sub&gt;</th>
<th>macroF&lt;sub&gt;1&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline BOW</td>
<td>86.09%</td>
<td>49.45%</td>
</tr>
<tr>
<td>BOSWE (90 × 200 words and linear kernel)</td>
<td>86.83%</td>
<td>49.68%</td>
</tr>
<tr>
<td>BOSWE (90 × 200 words and PQ kernel)</td>
<td>87.07%</td>
<td>49.51%</td>
</tr>
<tr>
<td>BOSWE (90 × 200 words and lin. + PQ kernel)</td>
<td><strong>87.24%</strong></td>
<td><strong>49.72%</strong></td>
</tr>
</tbody>
</table>
Outline

• Bag of Visual Words Model
• Word Embeddings
• Bag of Super Word Embeddings
• Training and Classification
• Polarity Classification Experiments
• Text Classification Experiments
• Conclusions and Future Work
Conclusion and Future Word

• Novel approach for building feature representation for various text classification tasks
Conclusion and Future Word

- Novel approach for building feature representation for various text classification tasks
- Uses word embeddings to form a bag of super word embeddings
Conclusion and Future Word

- Novel approach for building feature representation for various text classification tasks
- Uses word embeddings to form a bag of super word embeddings
- Surpass classical bag of words approach on two different tasks
Conclusion and Future Word

• Novel approach for building feature representation for various text classification tasks
• Uses word embeddings to form a bag of super word embeddings
• Surpass classical bag of words approach on two different tasks
• Replace k-means clustering with alternative approaches
Thank you!

- Thank you!
Thank you!

• Thank you!
• Questions?