Capsule Networks for NLP

Will Merrill
Advanced NLP
10/25/18
Capsule Networks: A Better ConvNet

- Architecture proposed by Hinton as a replacement for ConvNets in computer vision
- Several recent papers applying them to NLP:
  - Zhao et al., 2018
  - Srivastava et al., 2018
  - Xia et al. 2018
- Goals:
  - Understand the architecture
  - Go through recent papers
What’s Wrong with ConvNets?
Convolutional Neural Networks

- Cascade of convolutional layers and max-pooling layers
- Convolutional layer:
  - Slide window over image and apply filter

Max-Pooling

- ConvNets use max-pooling to move from low-level representations to high-level representations

https://computersciencewiki.org/index.php/Max-pooling#/Pooling
Problem #1: Transformational Invariance

- We would like networks to recognize transformations of the same image
- Requires huge datasets of transformed images to learn transformations of high-level features

https://medium.freecodecamp.org/understanding-capsule-networks-ais-alluring-new-architecture-bdb228173ddc
Problem #2: Feature Agreement

- Max-pooling in images loses information about relative position
- More abstractly, lower level features do not need to “agree”

https://medium.freecodecamp.org/understanding-capsule-networks-ais-alluring-new-architecture-bdb228173ddc
Capsule Network Architecture
Motivation

- We can solve problems #1 and #2 by attaching “instantiation parameters” to each filter
  - ConvNet: Is there a house here?
  - CapsNet: Is there a house with width $w$ and rotation $r$ here?
- Each filter at each position has a vector value instead of a scalar
- This vector is called a capsule
Capsules

- The value of capsule $i$ at some position is a vector $u_i$
- $|u_i| \in (0, 1)$ gives the probability of existence of feature $i$
- Direction of $u_i$ encodes the instantiation parameters of feature $i$

https://medium.freecodecamp.org/understanding-capsule-networks-ais-alluring-new-architecture-bdb228173ddc
Capsules (Continued)

https://medium.freecodecamp.org/understanding-capsule-networks-ais-alluring-new-architecture-bdb228173ddc
Capsule Squashing Function

- New squashing function which puts magnitude of vector into (0, 1)
- Referred to in literature as $g(..)$ or squash(..)
- Will be useful later on

$$v_j = \frac{||s_j||^2}{1 + ||s_j||^2} \frac{s_j}{||s_j||}$$

Sabour et al., 2017
Routing by Agreement

- Capture child-parent relationships
- Combine features into higher-level ones only if the lower-level features “agree” locally
- Is this picture a house or a sailboat?

https://medium.freecodecamp.org/understanding-capsule-networks-ais-alluring-new-architecture-bdb228173ddc
Routing: Vote Vectors

- Learned transformation for what information should be “passed up” to the next layer
- Models what information is relevant for abstraction/agreement
- \( \hat{u}_{ji} \) denotes the vote vector from capsule i to capsule j in the next layer

\[
\hat{u}_{j|i} = W^c_{j|i} u_i + \hat{b}_{j|i}
\]

Zhao et al., 2018
Routing: Dynamic Routing Algorithm

- Unsupervised iterative method for computing routing
- No parameters (But depends on vote vectors)
- Used to connect capsule layers
- Compute next layer of capsules \( \{v_j\} \) from vote vectors

**Procedure 1** Routing algorithm.

1: procedure ROUTING\((\hat{u}_{ji}, r, l)\)
2: for all capsule \( i \) in layer \( l \) and capsule \( j \) in layer \((l + 1)\): \( b_{ij} \leftarrow 0 \).
3: for \( r \) iterations do
4: for all capsule \( i \) in layer \( l \): \( c_i \leftarrow \text{softmax}(b_i) \) \( \triangleright \) softmax computes Eq. 3
5: for all capsule \( j \) in layer \((l + 1)\): \( s_j \leftarrow \sum_i c_{ij} \hat{u}_{ji} \)
6: for all capsule \( j \) in layer \((l + 1)\): \( v_j \leftarrow \text{squash}(s_j) \) \( \triangleright \) squash computes Eq. 1
7: for all capsule \( i \) in layer \( l \) and capsule \( j \) in layer \((l + 1)\): \( b_{ij} \leftarrow b_{ij} + \hat{u}_{ji} \cdot v_j \)

return \( v_j \)

Sabour et al., 2017
Types of Capsule Layers

1. **Primary Capsule Layer**: Convolutional output $\rightarrow$ capsules

2. **Convolutional Capsule Layer**: Local capsules $\rightarrow$ capsules

3. **Feedforward Capsule Layer**: All capsules $\rightarrow$ capsules
Primary Capsule Layer

Convolutional output → capsules
Create C capsules from B filters

1. Convolution output with B filters:
   \[ M = [m_1, m_2, ..., m_B] \in \mathbb{R}^{(L-K_1+1)\times B} \]

2. Transform each row of features:
   \[ p_i = g(W^b M_i + b_1) \quad W^b \in \mathbb{R}^{B\times d} \]

3. Collect C d-dimensional capsules:
   \[ P = [p_1, p_2, ..., p_C] \in \mathbb{R}^{(L-K_1+1)\times C\times d} \]

Zhao et al., 2018
Convolutional Capsule Layer

Local capsules in layer #1 $\rightarrow$ capsules in layer #2

- Route a sliding window of capsules in previous layer into capsules in next layer
Feedforward Capsules Layer

All capsules in layer #1 $\rightarrow$ capsules in layer #2

1. Flatten all capsules in layer #1 into a vector
2. Route from this vector of capsules into new capsules
Margin Loss

- Identify each output capsule with a class
- Classification loss for capsules
- Calculate on output of feedforward capsule layer
- Ensures that the capsule vector for the correct class is long ($|v| \approx 1$)

\[
L_k = T_k \max(0, m^+ - ||v_k||)^2 + \lambda (1 - T_k) \max(0, ||v_k|| - m^-)^2
\]

Sabour et al., 2017
Investigating Capsule Networks with Dynamic Routing for Text Classification

Zhao, Ye, Yang, Lei, Zhang, Zhao 2018
Main Ideas

1. Develops capsule network architecture for text classification tasks
2. Achieves state-of-the-art performance on single-class text classification
3. Capsules allow transferring single-class classification knowledge to multi-class task very well
Text Classification

- Read text and classify something about the passage
- Sentiment analysis, toxicity detection, etc.
Multi-Class Text Classification

- Document can be labeled as multiple classes
  - Example: In toxicity detection, Toxic and Threatening
Text Classification Architecture
Architectural Variants

- **Capsule-A**: One capsule network
- **Capsule-B**: Three capsule networks that are averaged at the end
Orphan Category

- Add a capsule that corresponds to no class to the final layer
- Network can send words unimportant to classification to this category
  - Function words like *the*, *a*, *in*, etc.
- More relevant in the NLP domain than in images because images don’t have a “default background”
## Datasets

### Single-Label

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>MR</td>
<td>8.6k</td>
<td>0.9k</td>
<td>1.1k</td>
<td>2</td>
</tr>
<tr>
<td>SST-2</td>
<td>8.6k</td>
<td>0.9k</td>
<td>1.8k</td>
<td>2</td>
</tr>
<tr>
<td>Subj</td>
<td>8.1k</td>
<td>0.9k</td>
<td>1.0k</td>
<td>2</td>
</tr>
<tr>
<td>TREC</td>
<td>5.4k</td>
<td>0.5k</td>
<td>0.5k</td>
<td>6</td>
</tr>
<tr>
<td>CR</td>
<td>3.1k</td>
<td>0.3k</td>
<td>0.4k</td>
<td>2</td>
</tr>
<tr>
<td>AG’s news</td>
<td>108k</td>
<td>12.0k</td>
<td>7.6k</td>
<td>4</td>
</tr>
</tbody>
</table>

Classification Task:
- review classification
- sentiment analysis
- opinion classification
- question categorization
- review classification
- news categorization

### Multi-Label

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reuters-Multi-label</td>
<td>5.8k</td>
<td>0.6k</td>
<td>0.3k</td>
<td>only multi-label data in test</td>
</tr>
<tr>
<td>Reuters-Full</td>
<td>5.8k</td>
<td>0.6k</td>
<td>3.4k</td>
<td>full data in test</td>
</tr>
</tbody>
</table>
## Single-Class Results

<table>
<thead>
<tr>
<th>Model</th>
<th>MR</th>
<th>SST2</th>
<th>Subj</th>
<th>TREC</th>
<th>CR</th>
<th>AG’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>75.9</td>
<td>80.6</td>
<td>89.3</td>
<td>86.8</td>
<td>78.4</td>
<td>86.1</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>79.3</td>
<td>83.2</td>
<td>90.5</td>
<td>89.6</td>
<td>82.1</td>
<td>88.2</td>
</tr>
<tr>
<td>Tree-LSTM</td>
<td>80.7</td>
<td>85.7</td>
<td>91.3</td>
<td>91.8</td>
<td>83.2</td>
<td>90.1</td>
</tr>
<tr>
<td>LR-LSTM</td>
<td>81.5</td>
<td><strong>87.5</strong></td>
<td>89.9</td>
<td>-</td>
<td>82.5</td>
<td>-</td>
</tr>
<tr>
<td>CNN-rand</td>
<td>76.1</td>
<td>82.7</td>
<td>89.6</td>
<td>91.2</td>
<td>79.8</td>
<td>92.2</td>
</tr>
<tr>
<td>CNN-static</td>
<td>81.0</td>
<td>86.8</td>
<td>93.0</td>
<td>92.8</td>
<td>84.7</td>
<td>91.4</td>
</tr>
<tr>
<td>CNN-non-static</td>
<td>81.5</td>
<td>87.2</td>
<td>93.4</td>
<td><strong>93.6</strong></td>
<td>84.3</td>
<td>92.3</td>
</tr>
<tr>
<td>CL-CNN</td>
<td>-</td>
<td>-</td>
<td>88.4</td>
<td>85.7</td>
<td>-</td>
<td>92.3</td>
</tr>
<tr>
<td>VD-CNN</td>
<td>-</td>
<td>-</td>
<td>88.2</td>
<td>85.4</td>
<td>-</td>
<td>91.3</td>
</tr>
<tr>
<td>Capsule-A</td>
<td>81.3</td>
<td>86.4</td>
<td>93.3</td>
<td>91.8</td>
<td>83.8</td>
<td>92.1</td>
</tr>
<tr>
<td>Capsule-B</td>
<td><strong>82.3</strong></td>
<td>86.8</td>
<td><strong>93.8</strong></td>
<td>92.8</td>
<td><strong>85.1</strong></td>
<td><strong>92.6</strong></td>
</tr>
</tbody>
</table>
## Multi-Class Transfer Learning Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Reuters-Multi-label</th>
<th>Reuters-Full</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ER</td>
<td>Precision</td>
</tr>
<tr>
<td>LSTM</td>
<td>23.3</td>
<td>86.7</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>26.4</td>
<td>82.3</td>
</tr>
<tr>
<td>CNN-rand</td>
<td>22.5</td>
<td>88.6</td>
</tr>
<tr>
<td>CNN-static</td>
<td>27.1</td>
<td>91.1</td>
</tr>
<tr>
<td>CNN-non-static</td>
<td>27.4</td>
<td>92.0</td>
</tr>
<tr>
<td>Capsule-A</td>
<td>57.2</td>
<td>88.2</td>
</tr>
<tr>
<td>Capsule-B</td>
<td><strong>60.3</strong></td>
<td><strong>95.4</strong></td>
</tr>
</tbody>
</table>
Interest rates on the London money market were slightly firmer on news U.K. Chancellor of the Exchequer Nigel Lawson had stated target rates for sterling against the dollar and mark, dealers said. They said this had come as a surprise and expected the targets, 2.50 marks and 1.60 dhms, to be promptly tested in the foreign exchange markets. Sterling opened 0.3 points lower in trade weighted terms at 71.3. Dealers noted the chancellor said he would achieve his goals on sterling by a combination of intervention in currency markets and interest rates. Operators feel the foreign exchanges are likely to test sterling on the downside and that this seems to make a fall in U.K. Base lending rates even less likely in the near term, dealers said. The feeling remains in the market, however, that fundamental factors have not really changed and that a rise in U.K. Interest rates is not very likely. The market is expected to continue at around these levels, reflecting the current 10 pct base rate level, for some time. The key three months interbank rate was 1/16 point firmer at 10 9-7/8 pct.
Discussion

● Capsule network performs strongly on single-class text-classification
● Capsule model transfers effectively from single-class to multi-class domain
  ○ Richer representation
  ○ No softmax in last layer
● Useful because multi-class data sets are hard to construct (exponentially larger than single-class data sets)
Identifying Aggression and Toxicity in Comments Using Capsule Networks

Srivastava, Khurana, Tewari 2018
Main Ideas

1. Develop end-to-end capsule model that outperforms state-of-the-art models for toxicity detection
2. Eliminate need for pipelining and preprocessing
3. Performs especially well on code-mixed comments (comments switching between English and Hindi)
Toxicity Detection

- Human moderation of online content is expensive – useful to do algorithmically
- Classify comments as toxic, severe toxic, identity hate, etc.
Challenges in Toxicity Detection

- Out-of-vocabulary words
- Code-mixing of languages
- Class imbalance
Why Capsule Networks?

- Seem to be good at text classification (Zhao et al., 2018)
- Should be better at code-mixing than sequential models (build up local representations)
Architecture

- Very similar to architecture to Zhao et al.
- Feature extraction convolutional layer replaced by LSTM
- Standard softmax layer instead of margin loss
Focal Loss

- Loss function on standard softmax output
- Used to solve the class imbalance problem
- Weights rare classes higher than cross-entropy

\[ FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t), \text{ where } p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{else} \end{cases} \]
Datasets

● Kaggle Toxic Comment Classification
  ○ English
  ○ Classes: Toxic, Severe Toxic, Obscene, Threat, Insult, Identity Hate

● First Shared Task on Aggression Identification (TRAC)
  ○ Mixed English and Hindi
  ○ Classes: Overtly Aggressive, Covertly Aggressive, Non-Aggressive

https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/discussion
## Results

<table>
<thead>
<tr>
<th>Model_Name</th>
<th>Kaggle-toxic comment classification (ROC-AUC)</th>
<th>TRAC - 1 (English-FB) (Weighted F1)</th>
<th>TRAC - 1 (English-TW) (Weighted F1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN-multifilter</td>
<td>95.16</td>
<td>55.43</td>
<td>53.41</td>
</tr>
<tr>
<td>CNN-LSTM</td>
<td>96.85</td>
<td>62.20</td>
<td>47.68</td>
</tr>
<tr>
<td>Bi-directional LSTM with maxpool</td>
<td>97.35</td>
<td>59.79</td>
<td>51.146</td>
</tr>
<tr>
<td>FeedForward Attention Networks</td>
<td>97.42</td>
<td>57.43</td>
<td>55.49</td>
</tr>
<tr>
<td>Hierarchical ConvNets</td>
<td>97.95</td>
<td>51.38</td>
<td>50.43</td>
</tr>
<tr>
<td>Bi-LSTM, Logistic Regression</td>
<td>98.17</td>
<td>57.17</td>
<td>52.1</td>
</tr>
<tr>
<td>Bi-LSTM, xgboosted</td>
<td>98.19</td>
<td>57.33</td>
<td>52.31</td>
</tr>
<tr>
<td>Bi-LSTM with skip connections</td>
<td>98.20</td>
<td>61.78</td>
<td>51.98</td>
</tr>
<tr>
<td>Pre-trained LSTMs</td>
<td>98.25</td>
<td>60.18</td>
<td>58.7</td>
</tr>
<tr>
<td>CapsuleNet without Focal Loss</td>
<td>98.21</td>
<td>62.032</td>
<td>58.600</td>
</tr>
<tr>
<td>CapsuleNet with Focal Loss</td>
<td><strong>98.46</strong></td>
<td><strong>63.43</strong></td>
<td><strong>59.41</strong></td>
</tr>
</tbody>
</table>
Training/Validation Loss

- Training and validation loss stayed much closer for the capsule model
- ⇒ Avoids overfitting

(a) Training and Validation Loss for Kaggle Toxic Comment Classification Dataset
Word Embeddings on Kaggle Corpus

- Three clear clusters:
  - Neutral words
  - Abusive words
  - Toxic words + place names

(b) Clusters for word obtained after training
OOV Embeddings

<table>
<thead>
<tr>
<th>NN to “politics”</th>
<th>NN to “bharat”</th>
</tr>
</thead>
<tbody>
<tr>
<td>politic</td>
<td>bharatiya</td>
</tr>
<tr>
<td>politician</td>
<td>bhar</td>
</tr>
<tr>
<td>politico</td>
<td>mahabharata</td>
</tr>
<tr>
<td>politicize</td>
<td>bharti</td>
</tr>
<tr>
<td>politician</td>
<td>bhashkar</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NN to “kut*e”(Hindi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>chu**ya</td>
</tr>
<tr>
<td>sa*le</td>
</tr>
<tr>
<td>tere</td>
</tr>
<tr>
<td>g**d</td>
</tr>
<tr>
<td>ma*<strong>rc</strong>d</td>
</tr>
</tbody>
</table>

Table 2: Example of handling misspelt words and transliteration. NN : Nearest Neighbour

- Out of vocabulary words randomly initialized
- Converge to accurate vectors
Discussion

- The novel capsule network architecture performed the best on all three datasets
- No data preprocessing done
- Avoids overfitting
- Local representations lead to big gains in mixed-language case
Zero-shot User Intent Detection via Capsule Neural Networks

Xia, Zhang, Yan, Chang, Yu 2018
Main Ideas

1. Capsule networks extract and organize information during supervised intent detection
2. These learned representations can be effectively transferred to the task of zero-shot intent detection
User Intent Detection

- Text classification task for question answering and dialog systems
- Classify which action a user query represents out of a known set of actions
  - GetWeather, PlayMusic
Zero-Shot User Intent Detection

- Training set with known set of intents
  - GetWeather, PlayMusic
- Test set has unseen “emerging” intents
  - AddToPlaylist, RateABook
- Transfer information about known intents to new domain of emerging intents
What Signal is There?

- Embedding of the string name of the unknown and known intents
- Output capsules for known intents
- Can combine these two things to do zero-shot learning
Architecture

Network trained on known intents

Extension for zero-shot inference
SemanticCaps Layer

- Extract features using self-attention LSTM

\[ \begin{align*}
\tilde{h}_t &= \text{LSTM}_{fw}(w_t, \tilde{h}_{t-1}), \\
\tilde{h}_t &= \text{LSTM}_{bw}(w_t, \tilde{h}_{t+1}).
\end{align*} \]

Combine to get \( \mathbf{H} \)

Self-attention weights

\[ \mathbf{A} = \text{softmax} \left( \mathbf{W}_2 \text{tanh} \left( \mathbf{W}_1 \mathbf{H}^T \right) \right) \]

\( \mathbf{M} \) is the extracted features

\[ \begin{align*}
\mathbf{M} &= \mathbf{A} \mathbf{H} \\
(\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_R) &\in \mathbb{R}^{R \times 2D_H}
\end{align*} \]
DetectionCaps Layer

- Standard convolutional capsule layer $\rightarrow$ feedforward capsule layer
Loss During Training

- Normal max-margin loss + regularization
- Regularization incentivizes semantic capsules to capture different features
- Regularization controlled by hyperparameter $\alpha$

Max-margin loss

$$
\mathcal{L} = \sum_{k=1}^{K} \left\{ [y = y_k] \cdot \max(0, m^+ - \|v_k\|)^2 \\
+ \lambda [y \neq y_k] \cdot \max(0, \|v_k\| - m^-)^2 \right\}
$$

Regularization term

$$
+ \alpha \|AA^T - I\|_F^2,
$$
## Intent Detection Results

<table>
<thead>
<tr>
<th>Model</th>
<th>SNIPS-NLU (on 5 existing intents)</th>
<th>CVA (on 80 existing intents)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Precision</td>
</tr>
<tr>
<td>TFIDF-LR</td>
<td>0.9546</td>
<td>0.9551</td>
</tr>
<tr>
<td>TFIDF-SVM</td>
<td>0.9584</td>
<td>0.9586</td>
</tr>
<tr>
<td>CNN</td>
<td>0.9595</td>
<td>0.9596</td>
</tr>
<tr>
<td>RNN</td>
<td>0.9516</td>
<td>0.9522</td>
</tr>
<tr>
<td>GRU</td>
<td>0.9535</td>
<td>0.9535</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.9569</td>
<td>0.9573</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>0.9501</td>
<td>0.9502</td>
</tr>
<tr>
<td>Self-attention Bi-LSTM</td>
<td>0.9524</td>
<td>0.9522</td>
</tr>
<tr>
<td><strong>IntentCapsNet</strong></td>
<td><strong>0.9621</strong></td>
<td><strong>0.9620</strong></td>
</tr>
</tbody>
</table>
Architecture Revisited

- Goal: Use predicted capsules for known intents for zero-shot inference
Generalizing to Emerging Intents

- Build similarity matrix between existing intents and emerging intents based on embeddings for intent names:

\[
q_{lk} = \frac{\exp \{-d(e_{zl}, e_{yk})\}}{\sum_{k=1}^{K} \exp \{-d(e_{zl}, e_{yk})\}},
\]

where

\[
d(e_{zl}, e_{yk}) = (e_{zl} - e_{yk})^T \Sigma^{-1} (e_{zl} - e_{yk})
\]
Classifying Emerging Intents

1. Goal is to get prediction vector for emerging intent $l$
2. Have vote vectors $g_{k,r}$ from known intent classification
3. Represent vote vector for emerging intent as weighted sum of known intents:
   \[ u_{l|r} = \sum_{k=1}^{K} q_{lk} g_{k,r} \]
4. Use dynamic routing to get an activation capsule $n_i$ for each emerging intent
5. Pick the $n_i$ with largest magnitude
Zero-Shot Intent Detection Results

<table>
<thead>
<tr>
<th>Model</th>
<th>SNIPS-NLU (on 2 emerging intents)</th>
<th>CVA (on 20 emerging intents)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Precision</td>
</tr>
<tr>
<td>DeViSE (Frome et al., 2013)</td>
<td>0.7447</td>
<td>0.7448</td>
</tr>
<tr>
<td>CMT (Socher et al., 2013)</td>
<td>0.7396</td>
<td>0.8266</td>
</tr>
<tr>
<td>CDSSM (Chen et al., 2016a)</td>
<td>0.7588</td>
<td>0.7625</td>
</tr>
<tr>
<td>Zero-shot DNN (Kumar et al., 2017)</td>
<td>0.7165</td>
<td>0.7330</td>
</tr>
<tr>
<td>INTENTCapsNet-ZSL w/o Self-attention</td>
<td>0.7587</td>
<td>0.7764</td>
</tr>
<tr>
<td>INTENTCapsNet-ZSL w/o Bi-LSTM</td>
<td>0.7619</td>
<td>0.7631</td>
</tr>
<tr>
<td>INTENTCapsNet-ZSL w/o Regularizer</td>
<td>0.7675</td>
<td>0.7676</td>
</tr>
<tr>
<td>INTENTCapsNet-ZSL</td>
<td>0.7752</td>
<td>0.7762</td>
</tr>
</tbody>
</table>
Discussion

- Representational power of capsule network can be leveraged for zero-shot learning
- Interesting regularizations and architectural extensions for capsule networks
Conclusion

● Capsule representations encode “instantiation parameters” of features
● Papers follow a standard CapsNet architecture for text classification:
  a. Features Extraction (ConvNet or LSTM)
  b. Primary Capsule Layer
  c. Convolutional Capsule Layer
  d. Classification (Margin or softmax)
● Capsule representations can be leveraged for transfer/zero-shot learning
Discussion Questions

1. What is powerful about capsule representations?
2. Are capsule networks good for NLP, or are they just good for vision?
3. Why has NLP capsule research focused on text classification tasks?
4. What are some other NLP tasks that capsule networks could be applied to?
5. What other advanced architectures could be useful in NLP?
Other Papers

Other Materials

- https://medium.com/ai%C2%B3-theory-practice-business/understanding-hintons-capsule-networks-part-i-intuition-b4b559d1159b