SGM: Sequence Generation Model for Multi-Label Classification

Pengcheng Yang (speaker), Xu Sun, Wei Li
Shuming Ma, Wei Wu, Houfeng Wang

Peking University, Beijing, China

24th Aug 2018
Outline

1. Introduction to Multi-Label Classification
2. Proposal: Sequence Generation Model
3. Experiments and Analysis
4. Conclusion
Introduction to Multi-Label Classification
What is Multi-Label Classification?

**Definition:**
- Assign multiple labels to each sample in the dataset.
Introduction to Multi-Label Classification

What is Multi-Label Classification?

1. **Definition:**
   - Assign multiple labels to each sample in the dataset.

2. **Example:**
   
   
   Rearranging the Familiar: Testing Compositional Generalization in Recurrent Networks
   
   João Loula, Marco Baroni, Brenden M. Lake
   
   **Subjects:** Computation and Language (cs.CL); Artificial Intelligence (cs.AI); Machine Learning (cs.LG)

   
   Statistical Model Compression for Small-Footprint Natural Language Understanding
   
   Grant P. Strimel, Kanthashree Mysore Sathyendra, Stanislav Peshterliev
   
   **Subjects:** Computation and Language (cs.CL)

   
   Using Deep Neural Networks to Translate Multi-lingual Threat Intelligence
   
   Priyanka Ranade, Sudip Mittal, Anupam Joshi, Karuna Joshi
   
   **Subjects:** Computation and Language (cs.CL); Cryptography and Security (cs.CR)
What is Multi-Label Classification?

1. **Definition:**
   - Assign multiple labels to each sample in the dataset.

2. **Example:**
     - **Rearranging the Familiar: Testing Compositional Generalization in Recurrent Networks**
     - João Loula, Marco Baroni, Brenden M. Lake
     - Subjects: Computation and Language (cs.CL); Artificial Intelligence (cs.AI); Machine Learning (cs.LG)
     - **Statistical Model Compression for Small-Footprint Natural Language Understanding**
     - Grant P. Strimel, Kanthashree Mysore Sathyendra, Stanislav Peshterliev
     - Comments: Interspeech 2018
     - Subjects: Computation and Language (cs.CL)
   - [arXiv:1807.07517](https://arxiv.org/abs/1807.07517) [pdf, other]
     - **Using Deep Neural Networks to Translate Multi-lingual Threat Intelligence**
     - Priyanka Ranade, Sudip Mittal, Anupam Joshi, Karuna Joshi
     - Subjects: Computation and Language (cs.CL); Cryptography and Security (cs.CR)

3. **Applications:**
   - Text categorization, information retrieval, and so on.
Background

Previous work:

1. Can’t capture label correlations very well or is computationally intractable.
   - **Label correlations:** Some labels are closely correlated.
Background

Previous work:

1. Can’t capture label correlations very well or is computationally intractable.
   - **Label correlations**: Some labels are closely correlated.

2. Ignore differences in the contributions of textual content when predicting different labels.

| Generating descriptions for videos has many applications including human robot interaction. |
| Generating descriptions for videos has many applications including human robot interaction. |
| Many methods for image captioning rely on pre-trained object classifier CNN and Long Short Term Memory recurrent networks. |
| Many methods for image captioning rely on pre-trained object classifier CNN and Long Short Term Memory recurrent networks. |
| How to learn robust visual classifiers from the weak annotations of the sentence descriptions. |
| How to learn robust visual classifiers from the weak annotations of the sentence descriptions. |

(a) Visual analysis when the SGM model predicts “CV”.

(b) Visual analysis when the SGM model predicts “CL”.

**Figure 1**: Visualization of attention.
Proposal: Sequence Generation Model
Transform classification task into generation task.

Key ideas:

- View the text as the source language and the label as target language.
- Base on sequence-to-sequence model.
Transform classification task into generation task.

1. **Key ideas:**
   - View the text as the source language and the label as target language.
   - Base on sequence-to-sequence model.

2. **Advantages:**
   - **Capture label correlations:** Generate labels sequentially, and predict the next label based on its previously generated labels.
   - **Consider differences in contributions of textual content:** Apply the attention mechanism.
Difficulties and solutions:

1. Repeated labels:
   - Use the masked softmax layer to smooth the probability distribution.
Difficulties and solutions:

1. Repeated labels:
   - Use the masked softmax layer to smooth the probability distribution.

2. Exposure bias:
   - Use the adaptive gate to introduce the global information of previous time-steps.
Difficulties and solutions:

1. **Repeated labels:**
   - Use the masked softmax layer to smooth the probability distribution.

2. **Exposure bias:**
   - Use the adaptive gate to introduce the global information of previous time-steps.

3. **Sequence order:**
   - Sort the label sequence of each sample according to the frequency of labels and high-frequency labels are placed in the front.
The proposed model is based on the **Seq2Seq** model, which consists of an encoder and a decoder with global embedding.
Proposal: Masked Softmax Layer

Masked softmax layer: Prevent the decoder from predicting repeated labels.

1. $y_t$ is the probability distribution over the label space $\mathcal{L}$ at time-step $t$.

$$y_t = \text{softmax}(o_t + l_t)$$ (1)

2. $l_t \in \mathbb{R}^L$ is the mask vector.

$$(l_t)_i = \begin{cases} -\infty & \text{if the label } l_i \text{ has been predicted.} \\ 0 & \text{otherwise.} \end{cases}$$ (2)
Proposal: Global Embedding

Global embedding: Introduce the global information of previous time-steps to alleviate the exposure bias.

\[
\bar{e} = \sum_{i=1}^{L} y_{t-1}^{(i)} e_i
\]

\[
g(y_t) = (1 - H) \odot e + H \odot \bar{e}
\]

\[
H = W_1 e + W_2 \bar{e}
\]

1. \(e\) is the embedding vector of the label which has the highest probability under distribution \(y_{t-1}\).
2. \(e_i\) is the embedding vector of the \(i\)-th label.
3. \(W_1, W_2 \in \mathbb{R}^{L \times L}\) are weight matrices.
Experiments and Analysis
Datasets and Evaluation Metrics

1. **Datasets:**
   - RCV1-V2: Reuters Corpus Volume I.

2. **Evaluation metrics:**
   - Hamming loss and micro-\( F_1 \) score are our main evaluation metrics.
   - Micro-precision and micro-recall are also reported to assist the analysis.
### Results

<table>
<thead>
<tr>
<th>Models</th>
<th>HL(-)</th>
<th>P(+)</th>
<th>R(+)</th>
<th>F1(+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BR</td>
<td>0.0086</td>
<td>0.904</td>
<td>0.816</td>
<td>0.858</td>
</tr>
<tr>
<td>CC</td>
<td>0.0087</td>
<td>0.887</td>
<td>0.828</td>
<td>0.857</td>
</tr>
<tr>
<td>LP</td>
<td>0.0087</td>
<td>0.896</td>
<td>0.824</td>
<td>0.858</td>
</tr>
<tr>
<td>CNN</td>
<td>0.0089</td>
<td><strong>0.922</strong></td>
<td>0.798</td>
<td>0.855</td>
</tr>
<tr>
<td>CNN-RNN</td>
<td>0.0085</td>
<td>0.889</td>
<td>0.825</td>
<td>0.856</td>
</tr>
<tr>
<td>SGM</td>
<td>0.0081</td>
<td>0.887</td>
<td>0.850</td>
<td>0.869</td>
</tr>
<tr>
<td>+ GE</td>
<td><strong>0.0075</strong></td>
<td>0.897</td>
<td><strong>0.860</strong></td>
<td><strong>0.878</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Models</th>
<th>HL(-)</th>
<th>P(+)</th>
<th>R(+)</th>
<th>F1(+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BR</td>
<td>0.0316</td>
<td>0.644</td>
<td>0.648</td>
<td>0.646</td>
</tr>
<tr>
<td>CC</td>
<td>0.0306</td>
<td>0.657</td>
<td>0.651</td>
<td>0.654</td>
</tr>
<tr>
<td>LP</td>
<td>0.0312</td>
<td>0.662</td>
<td>0.608</td>
<td>0.634</td>
</tr>
<tr>
<td>CNN</td>
<td>0.0256</td>
<td><strong>0.849</strong></td>
<td>0.545</td>
<td>0.664</td>
</tr>
<tr>
<td>CNN-RNN</td>
<td>0.0278</td>
<td>0.718</td>
<td>0.618</td>
<td>0.664</td>
</tr>
<tr>
<td>SGM</td>
<td>0.0251</td>
<td>0.746</td>
<td>0.659</td>
<td>0.699</td>
</tr>
<tr>
<td>+ GE</td>
<td><strong>0.0245</strong></td>
<td>0.748</td>
<td><strong>0.675</strong></td>
<td><strong>0.710</strong></td>
</tr>
</tbody>
</table>

(a) Performance on RCV1-V2 test set. (b) Performance on AAPD test set.

**Table 1**: Comparison between our methods and all baselines on two datasets. GE denotes the global embedding. HL, P, R, and F1 denote hamming loss, micro-precision, micro-recall, and micro-$F_1$, respectively.
Goal: Explore how the performance of our model is affected by the proportion between two kinds of embeddings.

Settings:

- Adaptive gate:

  \[ g(y_t) = (1 - H) \odot e + H \odot \bar{e} \]  
  \[ H = W_1 e + W_2 \bar{e} \]

- Coefficient averaging:

  \[ g(y_t) = (1 - \lambda) \ast e + \lambda \ast \bar{e} \]
Exploration of Global Embedding

Figure 3: The performance of the SGM when using different $\lambda$. The red dotted line represents the results of using the adaptive gate.
Exploration of Global Embedding

Figure 3: The performance of the SGM when using different $\lambda$. The red dotted line represents the results of using the adaptive gate.

1. The weighted average embedding contains richer information, leading to the improvement in the performance of the model.

2. The adaptive gate can automatically determine the most appropriate $\lambda$ value according to the actual condition.
Experiments and Analysis

Ablation Study

Table 2: Ablation study on the RCV1-V2 test set. GE denotes global embedding. ↓ indicates that the performance of the model is degraded.

<table>
<thead>
<tr>
<th>Models</th>
<th>HL(-)</th>
<th>F1(+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGM</td>
<td>0.0081</td>
<td>0.869</td>
</tr>
<tr>
<td>w/o mask</td>
<td>0.0083(↓ 2.47%)</td>
<td>0.866(↓ 0.35%)</td>
</tr>
<tr>
<td>w/o sorting</td>
<td>0.0084(↓ 3.70%)</td>
<td>0.858(↓ 1.27%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Models</th>
<th>HL(-)</th>
<th>F1(+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGM + GE</td>
<td>0.0075</td>
<td>0.878</td>
</tr>
<tr>
<td>w/o mask</td>
<td>0.0078(↓ 4.00%)</td>
<td>0.873(↓ 0.57%)</td>
</tr>
<tr>
<td>w/o sorting</td>
<td>0.0083(↓ 10.67%)</td>
<td>0.859(↓ 2.16%)</td>
</tr>
</tbody>
</table>

(a) Ablation study for SGM.

(b) Ablation study for SGM with GE.
Experiments and Analysis

Ablation Study

<table>
<thead>
<tr>
<th>Models</th>
<th>HL(-)</th>
<th>F1(+)</th>
<th>Models</th>
<th>HL(-)</th>
<th>F1(+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGM</td>
<td>0.0081</td>
<td>0.869</td>
<td>SGM + GE</td>
<td>0.0075</td>
<td>0.878</td>
</tr>
<tr>
<td>w/o mask</td>
<td>0.0083</td>
<td>0.866</td>
<td>w/o mask</td>
<td>0.0078</td>
<td>0.873</td>
</tr>
<tr>
<td>w/o sorting</td>
<td>0.0084</td>
<td>0.858</td>
<td>w/o sorting</td>
<td>0.0083</td>
<td>0.859</td>
</tr>
</tbody>
</table>

(a) Ablation study for SGM.  
(b) Ablation study for SGM with GE.

Table 2: Ablation study on the RCV1-V2 test set. GE denotes global embedding. ↓ indicates that the performance of the model is degraded.

1. Sorting is important because humans need to **predefine the order** of output labels.
2. The mask module has little impact because **label cardinality is small**.

Pengcheng Yang (speaker), Xu Sun, Wei Li Shuming Ma, Wei Wu, Houfeng Wang (Peking University, Beijing, China)
COLING 2018 Oral Report
24th Aug 2018 17 / 23
Figure 4: The performance of SGM on different subsets of the RCV1-V2 test set. LLS represents the length of label sequence of each sample in the subset.
Figure 4: The performance of SGM on different subsets of the RCV1-V2 test set. LLS represents the length of label sequence of each sample in the subset.

1. The performance of all methods **deteriorates** when LLS increases.
2. The advantages of SGM are more **significant** when LLS is large.
Visualization of Attention

<table>
<thead>
<tr>
<th>(a) Visual analysis when the SGM model predicts “CV”.</th>
<th>(b) Visual analysis when the SGM model predicts “CL”.</th>
</tr>
</thead>
<tbody>
<tr>
<td>● Generating descriptions for videos has many applications including human interaction.</td>
<td>● Generating descriptions for videos has many applications including human interaction.</td>
</tr>
<tr>
<td>● How to learn robust visual classifiers from the weak annotations of the sentence descriptions.</td>
<td>● How to learn robust visual classifiers from the weak annotations of the sentence descriptions.</td>
</tr>
</tbody>
</table>

Figure 5: An example abstract in the AAPD dataset, from which we extract three informative sentences. This abstract is assigned two labels: “CV” and “CL”. They denote computer vision and computational language, respectively.
Experiments and Analysis

Visualization of Attention

- Generating descriptions for videos has many applications including human robot interaction.
- Many methods for image captioning rely on pre-trained object classifier CNN and Long Short Term Memory recurrent networks.
- How to learn robust visual classifiers from the weak annotations of the sentence descriptions.

(a) Visual analysis when the SGM model predicts “CV”.

- Generating descriptions for videos has many applications including human robot interaction.
- Many methods for image captioning rely on pre-trained object classifier CNN and Long Short Term Memory recurrent networks.
- How to learn robust visual classifiers from the weak annotations of the sentence descriptions.

(b) Visual analysis when the SGM model predicts “CL”.

**Figure 5:** An example abstract in the AAPD dataset, from which we extract three informative sentences. This abstract is assigned two labels: “CV” and “CL”. They denote computer vision and computational language, respectively.

- The attention mechanism can select the most informative words automatically when predicting different labels.
Case Study

**Figure 6**: Several examples of generated label sequences on the RCV1-V2 dataset. The red bold labels in each example indicate that they are highly correlated.

<table>
<thead>
<tr>
<th>Reference</th>
<th>BR</th>
<th>SGM</th>
<th>SGM + GE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCAT, <strong>C15, C152</strong>, C41, C411</td>
<td>CCAT, C15, C13</td>
<td><strong>C15, C152</strong></td>
<td><strong>C15, C152</strong></td>
</tr>
<tr>
<td>CCAT, GCAT, ECAT, C31, GDIP, C13, C21, <strong>E51, E512</strong></td>
<td>CCAT, GCAT, GDIP, E51</td>
<td><strong>E51, E512</strong></td>
<td><strong>C15, E51, E512, C31</strong></td>
</tr>
<tr>
<td>GCAT, ECAT, <strong>G15, G154, G151, G155</strong></td>
<td>GCAT, ECAT, GENV, G15</td>
<td><strong>G15, G154, G156</strong></td>
<td><strong>G15, G154, G155</strong></td>
</tr>
</tbody>
</table>
Case Study

**Figure 6:** Several examples of generated label sequences on the RCV1-V2 dataset. The red bold labels in each example indicate that they are highly correlated.

<table>
<thead>
<tr>
<th>Reference</th>
<th>BR</th>
<th>SGM</th>
<th>SGM + GE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCAT, <strong>C15, C152, C41, C411</strong></td>
<td>CCAT, C15, C13</td>
<td><strong>CCAT, C15, C152</strong></td>
<td><strong>CCAT, C15, C152, C41, C411</strong></td>
</tr>
<tr>
<td>CCAT, GCAT, ECAT, C31, GDIP, C13, C21, <strong>E51, E512</strong></td>
<td>CCAT, GCAT, GDIP, E51</td>
<td><strong>CCAT, ECAT, GDIP, E51, E512</strong></td>
<td><strong>CCAT, GCAT, ECAT, C31, GDIP, E51, E512, C312</strong></td>
</tr>
<tr>
<td>GCAT, ECAT, <strong>G15, G154, G151, G155</strong></td>
<td>GCAT, ECAT, GENV, G15</td>
<td><strong>GCAT, ECAT, E21, G15, G154, G156</strong></td>
<td><strong>GCAT, ECAT, E21, G15, G154, G155</strong></td>
</tr>
</tbody>
</table>

1. The proposed SGM can **capture the correlations** between labels.
2. The SGM with global embedding predicts labels **more accurately**.
Conclusion
Conclusion

1. The sequence generation model is able to capture the correlations between labels well.
2. The attention mechanism can select the most informative words automatically when predicting different labels.
3. The global embedding can alleviate exposure bias by introducing the global information of previous time-steps.
If there is any question, please contact Pengcheng Yang (yang_pc@pku.edu.cn)

The code and datasets are available at https://github.com/lancopku/SGM

Thank you!