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?

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

What is Multi-Label Classification?

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

[6] **arXiv:1807.07545 [pdf, other]**

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)

[7] **arXiv:1807.07520 [pdf, ps, other]**

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)**

[8] **arXiv: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)

What is Multi-Label Classification?

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

[6] **arXiv:1807.07545 [pdf, other]**
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)

[7] **arXiv:1807.07520 [pdf, ps, other]**
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)**

[8] **arXiv: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

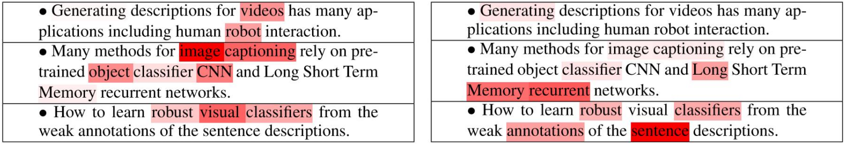
Previous work:

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.



(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

Proposal: Sequence Generation 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.

Proposal: Sequence Generation 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.

Proposal: Sequence Generation Model

Difficulties and solutions:

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

Proposal: Sequence Generation Model

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.

Proposal: Sequence Generation Model

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.

Proposal: Sequence Generation Model

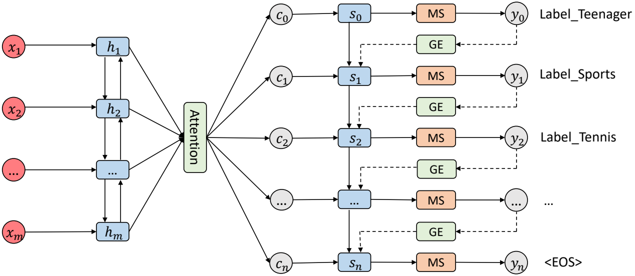


Figure 2: The overview of SGM with global embedding. MS denotes the masked softmax layer. GE denotes the global embedding.

- 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.

$$y_t = softmax(o_t + I_t) \tag{1}$$

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

$$(I_t)_i = \begin{cases} -\infty & \text{if the label } l_i \text{ has been predicted.} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Proposal: Global Embedding

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

$$\bar{e} = \sum_{i=1}^{L} y_{t-1}^{(i)} e_i \tag{3}$$

$$g(y_t) = (1 - H) \odot e + H \odot \bar{e} \tag{4}$$

$$H = \boldsymbol{W}_1 e + \boldsymbol{W}_2 \bar{e} \tag{5}$$

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. $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathbb{R}^{L \times L}$ are weight matrices.

Experiments and Analysis

Datasets and Evaluation Metrics

1. **Datasets:**
   - **RCV1-V2:** Reuters Corpus Volume I.
   - **AAPD:** Arxiv Academic Paper Dataset.
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

Models	HL(-)	P(+)	R(+)	F1(+)
BR	0.0086	0.904	0.816	0.858
CC	0.0087	0.887	0.828	0.857
LP	0.0087	0.896	0.824	0.858
CNN	0.0089	**0.922**	0.798	0.855
CNN-RNN	0.0085	0.889	0.825	0.856
SGM	0.0081	0.887	0.850	0.869
+ GE	**0.0075**	0.897	**0.860**	**0.878**

(a) Performance on RCV1-V2 test set.

Models	HL(-)	P(+)	R(+)	F1(+)
BR	0.0316	0.644	0.648	0.646
CC	0.0306	0.657	0.651	0.654
LP	0.0312	0.662	0.608	0.634
CNN	0.0256	**0.849**	0.545	0.664
CNN-RNN	0.0278	0.718	0.618	0.664
SGM	0.0251	0.746	0.659	0.699
+ GE	**0.0245**	0.748	**0.675**	**0.710**

(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.

Exploration of Global Embedding

1. **Goal:** Explore how the performance of our model is affected by the proportion between two kinds of embeddings.
2. **Settings:**
   - Adaptive gate:

$$g(y_t) = (1 - H) \odot e + H \odot \bar{e} \tag{6}$$
$$H = \boldsymbol{W}_1 e + \boldsymbol{W}_2 \bar{e} \tag{7}$$

   - Coefficient averaging:

$$g(y_t) = (1 - \lambda) * e + \lambda * \bar{e} \tag{8}$$

Exploration of Global Embedding

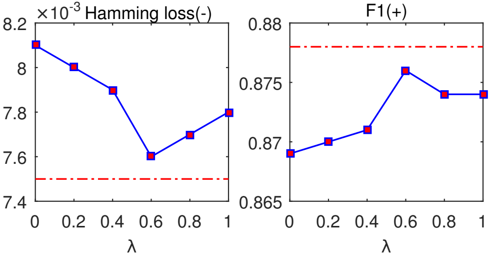


Figure 3: The performance of the SGM when using different λ. The red dotted line represents the results of using the adaptive gate.

Exploration of Global Embedding

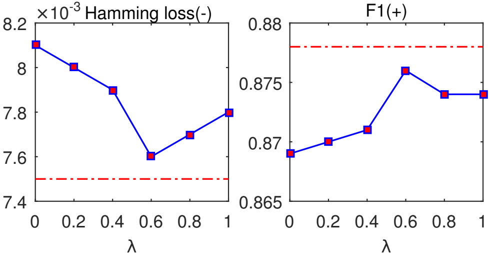


Figure 3: The performance of the SGM when using different λ. The red dotted line represents the results of using the adaptive gate.

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 λ value according to the actual condition.

Ablation Study

Models	HL(-)	F1(+)
SGM	0.0081	0.869
w/o mask	0.0083(↓ 2.47%)	0.866(↓ 0.35%)
w/o sorting	0.0084(↓ 3.70%)	0.858(↓ 1.27%)

(a) Ablation study for SGM.

Models	HL(-)	F1(+)
SGM + GE	0.0075	0.878
w/o mask	0.0078(↓ 4.00%)	0.873(↓ 0.57%)
w/o sorting	0.0083(↓ 10.67%)	0.859(↓ 2.16%)

(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.

Ablation Study

Models	HL(-)	F1(+)
SGM	0.0081	0.869
w/o mask	0.0083(↓ 2.47%)	0.866(↓ 0.35%)
w/o sorting	0.0084(↓ 3.70%)	0.858(↓ 1.27%)

(a) Ablation study for SGM.

Models	HL(-)	F1(+)
SGM + GE	0.0075	0.878
w/o mask	0.0078(↓ 4.00%)	0.873(↓ 0.57%)
w/o sorting	0.0083(↓ 10.67%)	0.859(↓ 2.16%)

(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.

Error Analysis

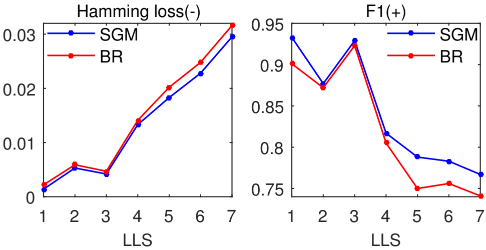


Figure 4: The performance of SGM on different subsets of the RCV1-V2 test set. LLS represents the length of label sequence of each sample in the subset.

Error Analysis

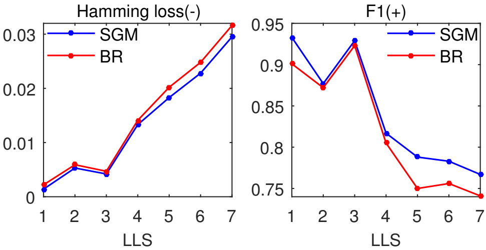


Figure 4: The performance of SGM on different subsets of the RCV1-V2 test set. LLS represents the length of label sequence of each sample in the subset.

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

• 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.

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

Reference	BR	SGM	SGM + GE
CCAT, **C15, C152**, C41, C411	CCAT, C15, C13	CCAT, **C15, C152**	CCAT, **C15, C152**, C41, C411
CCAT, GCAT, ECAT, C31, GDIP, C13, C21, **E51, E512**	CCAT, GCAT, GDIP, E51	CCAT, ECAT, GDIP, **E51, E512**	CCAT, GCAT, ECAT, C31, GDIP, **E51, E512**, C312
GCAT, ECAT, **G15, G154, G151, G155**	GCAT, ECAT, GENV, G15	GCAT, ECAT, E21, **G15, G154, G156**	GCAT, ECAT, E21, **G15, G154, G155**

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.

Case Study

Reference	BR	SGM	SGM + GE
CCAT, **C15, C152**, C41, C411	CCAT, C15, C13	CCAT, **C15, C152**	CCAT, **C15, C152**, C41, C411
CCAT, GCAT, ECAT, C31, GDIP, C13, C21, **E51, E512**	CCAT, GCAT, GDIP, E51	CCAT, ECAT, GDIP, **E51, E512**	CCAT, GCAT, ECAT, C31, GDIP, **E51, E512**, C312
GCAT, ECAT, **G15, G154, G151, G155**	GCAT, ECAT, GENV, G15	GCAT, ECAT, E21, **G15, G154, G156**	GCAT, ECAT, E21, **G15, G154, G155**

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.

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!