# Cross-Modal Commentator: Automatic Machine Commenting Based on Cross-Modal Information

**Pengcheng Yang[1,2]*, Zhihan Zhang[2]*, Fuli Luo[2], Lei Li[2], Chengyang Huang[3], Xu Sun[1,2]**

[1]Deep Learning Lab, Beijing Institute of Big Data Research, Peking University

[2]MOE Key Lab of Computational Linguistics, School of EECS, Peking University

[3]Beijing University of Posts and Telecommunications

# Task: Cross-Modal Automatic Commenting



**Title:**

春意盎然 山西万亩桃花惹人醉

Spring is coming! Thousands of acres are filled with intoxicating peach blossoms in Shanxi.

**Body:**

近日山西平鲁万亩桃花竞相绽放，游人沉醉花丛中，尽情感受春天的气息。

Recently, thousands of acres of peach blossoms are in full bloom at Pinglu, Shanxi Province. Visitors are immersed in the beautiful flowers, enjoying the breath of spring.

**Comments:**

挺漂亮，流连忘返！

Beautiful flowers! I can't move my eyes from them.

没有绿草的衬托，桃花少了一点美感。

Peach blossoms seem to be a little less pretty without any green grass as background.

绿色多点就好了。

It would be better if there is more greenness.

# Data: Cross-Modal Comment Dataset

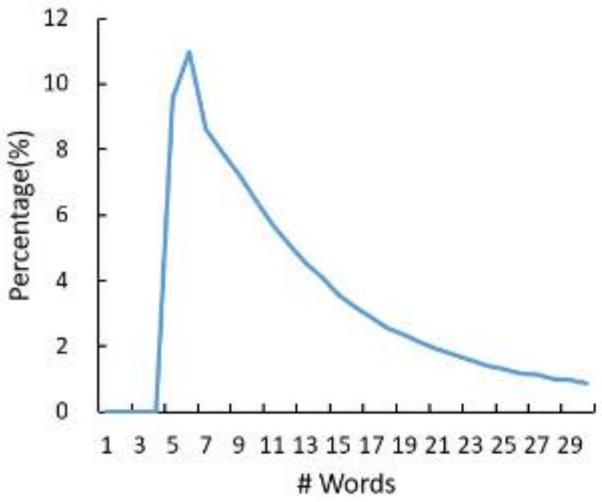
| Statistic | Train | Dev | Test | Total |
|---|---|---|---|---|
| # News | 19,162 | 3,521 | 1,451 | 24,134 |
| # Comments | 746,423 | 131,175 | 53,058 | 930,656 |
| Avg. Images | 5.81 | 5.78 | 5.81 | 5.80 |
| Avg. Body | 54.75 | 54.72 | 55.07 | 54.77 |
| Avg. Comment | 12.19 | 12.21 | 12.18 | 12.19 |

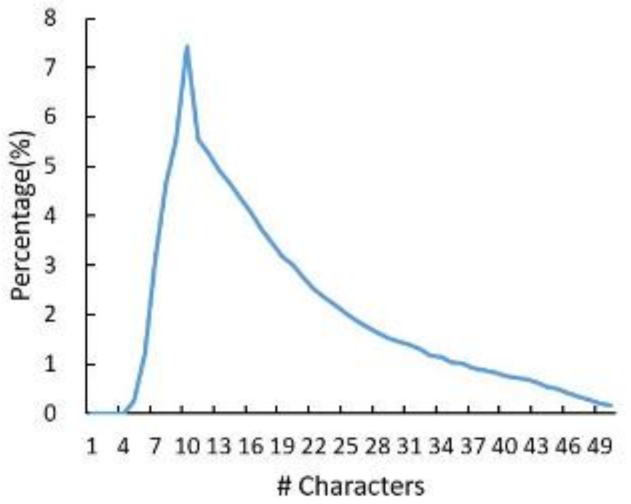
- Large-scale cross-modal dataset
- Collected from Netease News
- 24,134 news articles
- 202,951 news pictures and photos
- 930,656 high-quality comments

# Data: Cross-Modal Comment Dataset

| Evaluation | Flue. | Rele. | Info. | Overall |
|---|---|---|---|---|
| Score | 9.2 | 6.7 | 6.4 | 7.6 |
| Pearson | 0.74 | 0.76 | 0.66 | 0.68 |

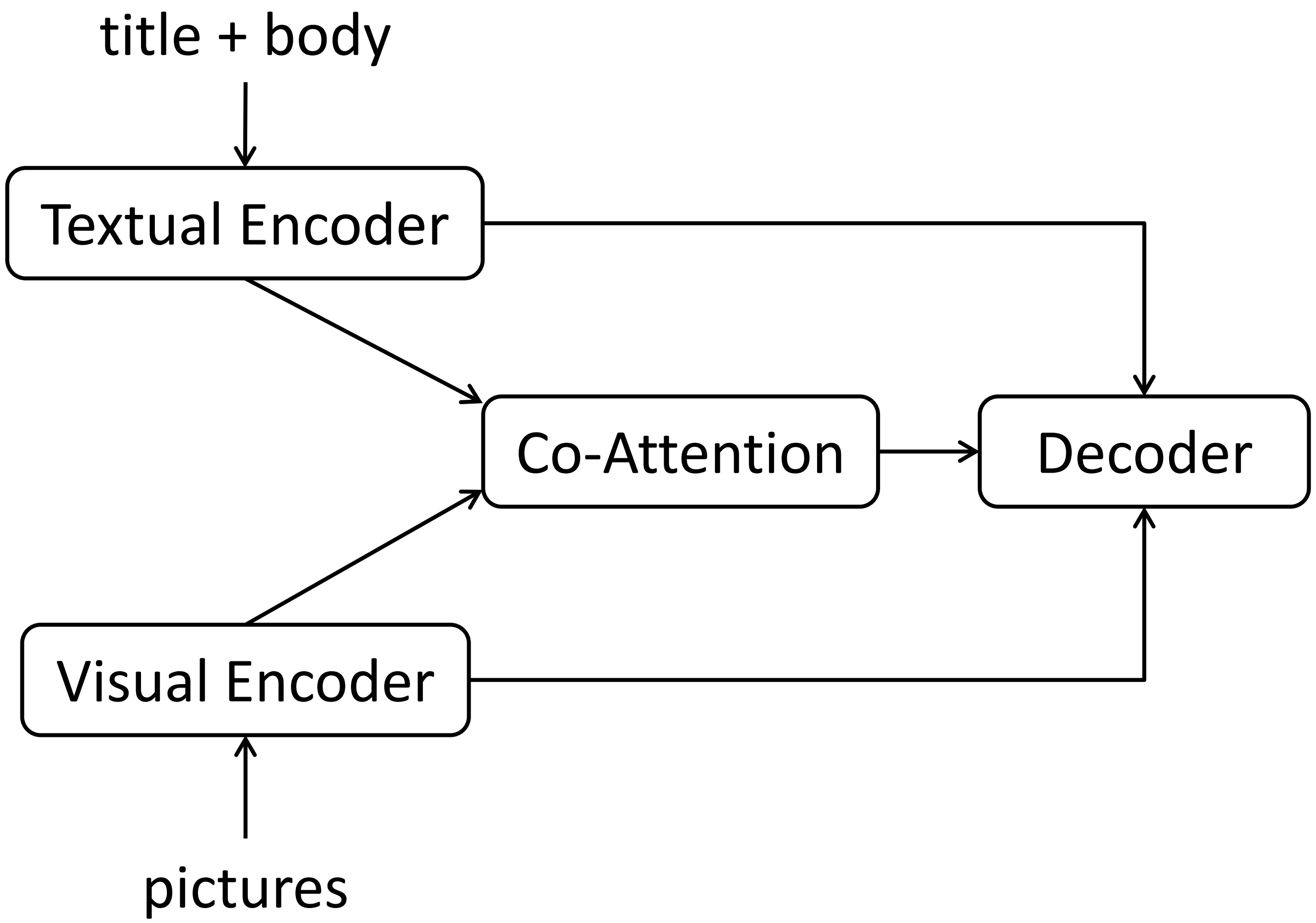
Quality evaluation results of the testing set.
(**Flue.**, **Rele.** and **Info.** denotes fluency, relevance, and informativeness.)



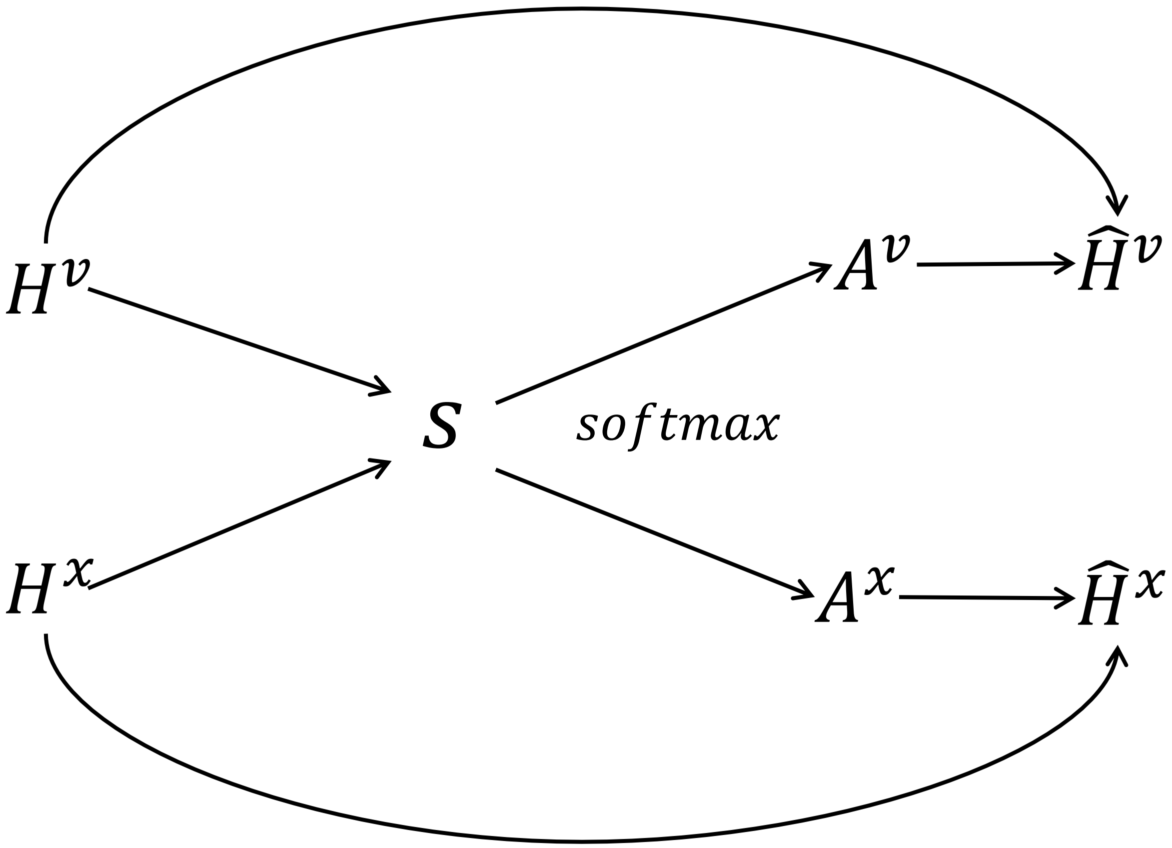

The distribution of lengths for comments in terms of both word-level and character-level.

# Method: Co-Attention Model




1. Encoders: extract information from text and pictures
2. Co-Attention: model intrinsic dependencies between two modalities
3. Decoder: generate output comments

# Method: Co-Attention Model




$$S = H^v W (H^X)^T$$

$$A^x = softmax(S)$$

$$A^v = softmax(S^T)$$

$$\widehat{H}^x = A^x H^x$$

$$\widehat{H}^v = A^v H^v$$

# Experiments

## Dataset:

Cross-Modal Comment Dataset, 930,656 pieces of news comments.

## Baselines:

- Seq2seq based models with separate attention: S2S-V, S2S-T, S2S-VT
- Transformer based models: Trans-V, Trans-T, Trans-VT

# Experiments: results

| Models | BLEU-1 | ROUGE-L | DIST-1 | DIST-2 |
|---|---|---|---|---|
| S2S-V | 6.1 | 7.8 | 1348 | 3293 |
| S2S-T | 6.3 | 8.1 | 1771 | 4285 |
| S2S-VT | 6.6 | 8.5 | 1929 | 4437 |
| **Our (S2S)** | **7.1** | **9.1** | **2279** | **4743** |
| Trans-V | 5.9 | 7.6 | 1336 | 3472 |
| Trans-T | 6.4 | 8.3 | 1772 | 4694 |
| Trans-VT | 6.8 | 8.6 | 1891 | 4739 |
| **Our (Trans)** | **7.7** | **9.4** | **2265** | **4941** |

| Models | Flue. | Rele. | Info. | Overall |
|---|---|---|---|---|
| S2S-V | 3.1 | 2.8 | 2.5 | 3.2 |
| S2S-T | 4.5 | 4.6 | 3.7 | 4.7 |
| S2S-VT | 4.6 | 5.1 | 4.3 | 4.9 |
| **Our (S2S)** | **4.8** | **5.7** | **4.7** | **5.1** |
| Trans-V | 2.9 | 2.3 | 2.8 | 2.9 |
| Trans-T | 4.3 | 4.8 | 4.4 | 4.6 |
| Trans-VT | 4.7 | 4.6 | 4.7 | 5.1 |
| **Our (Trans)** | **4.9** | **5.9** | **5.0** | **5.2** |

- The universality of our co-attention model
- The effectiveness of co-attention mechanism
- The positive impact of images

# Thank you!

The code is available at https://github.com/lancopku/CMAC