Improving Image Captioning by Concept-based Sentence Reranking

Xirong Li & Qin Jin

Multimedia Computing Lab, Renmin University of China

2016-09-15 / PCM 2016

Multimedia Meets Language

- Image captioning is to automatically generate a **natural-language** sentence that describes **main** entities and events in a given image.

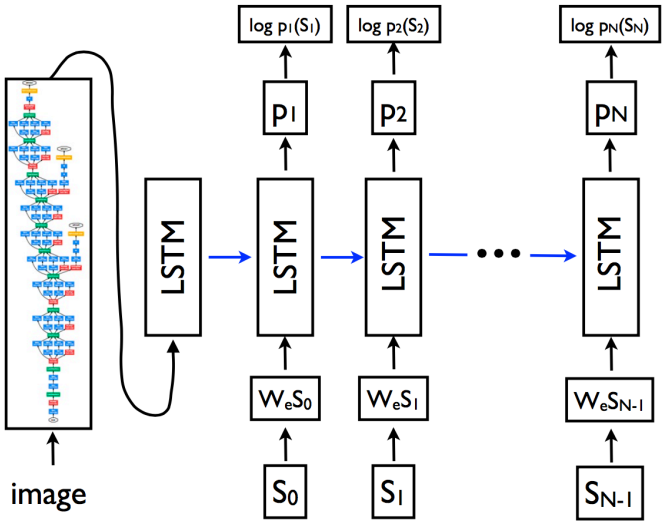
Two dogs playing on beach

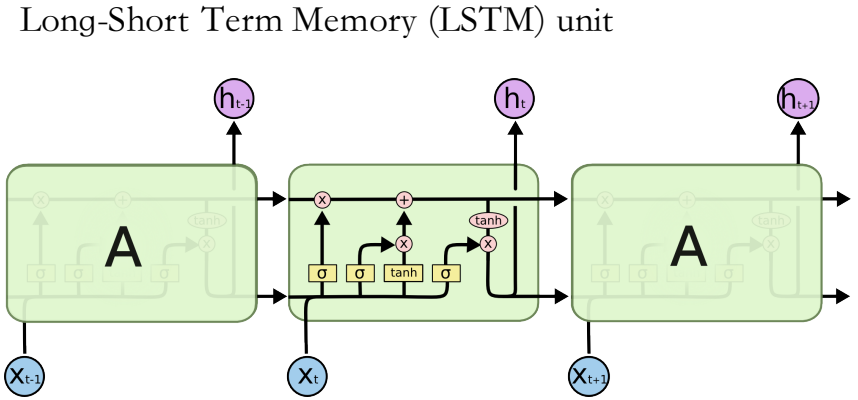
草地上的可爱兔子

麦田里的人物

State-of-the-Art on Image Captioning

- Google's Show and Tell model[Vinyals CVPR2015]
  - Image encoding by CNN and sentence decoding by LSTM

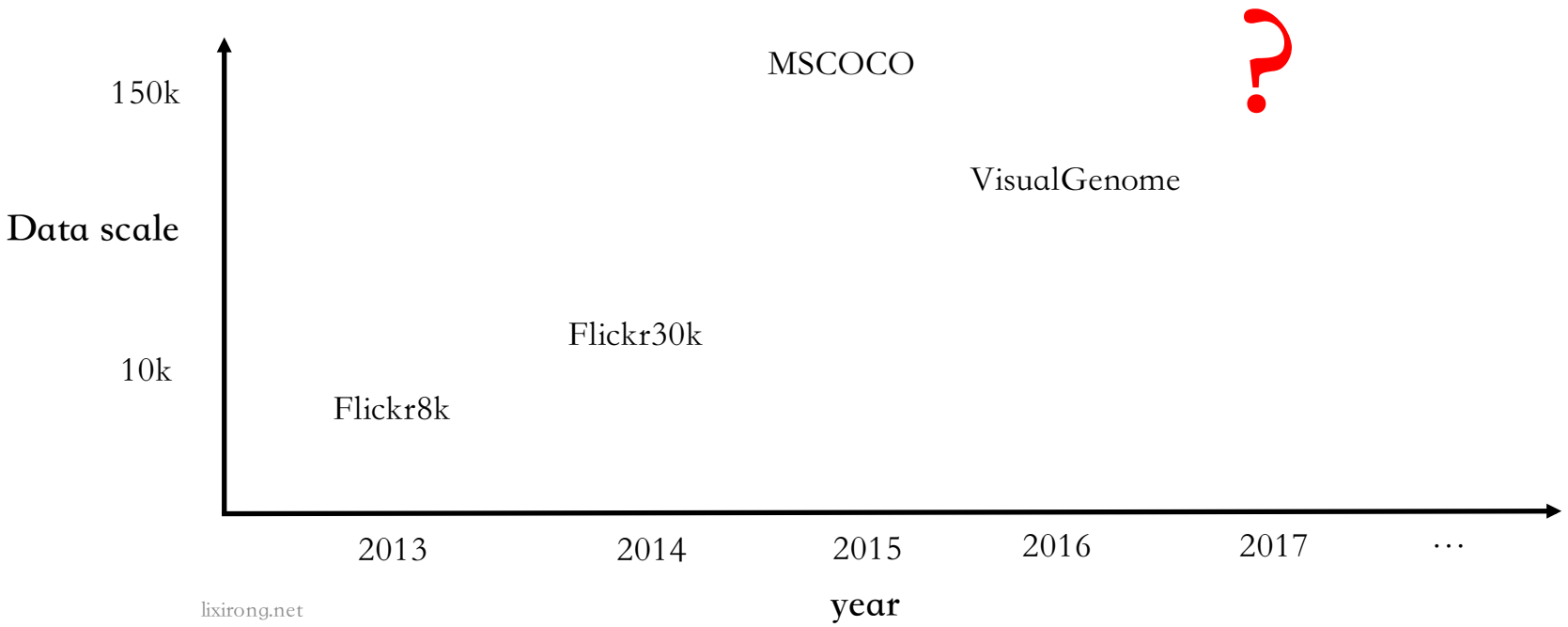




http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Data is Becoming a Bottleneck, Again

- Development of image sentence corpora



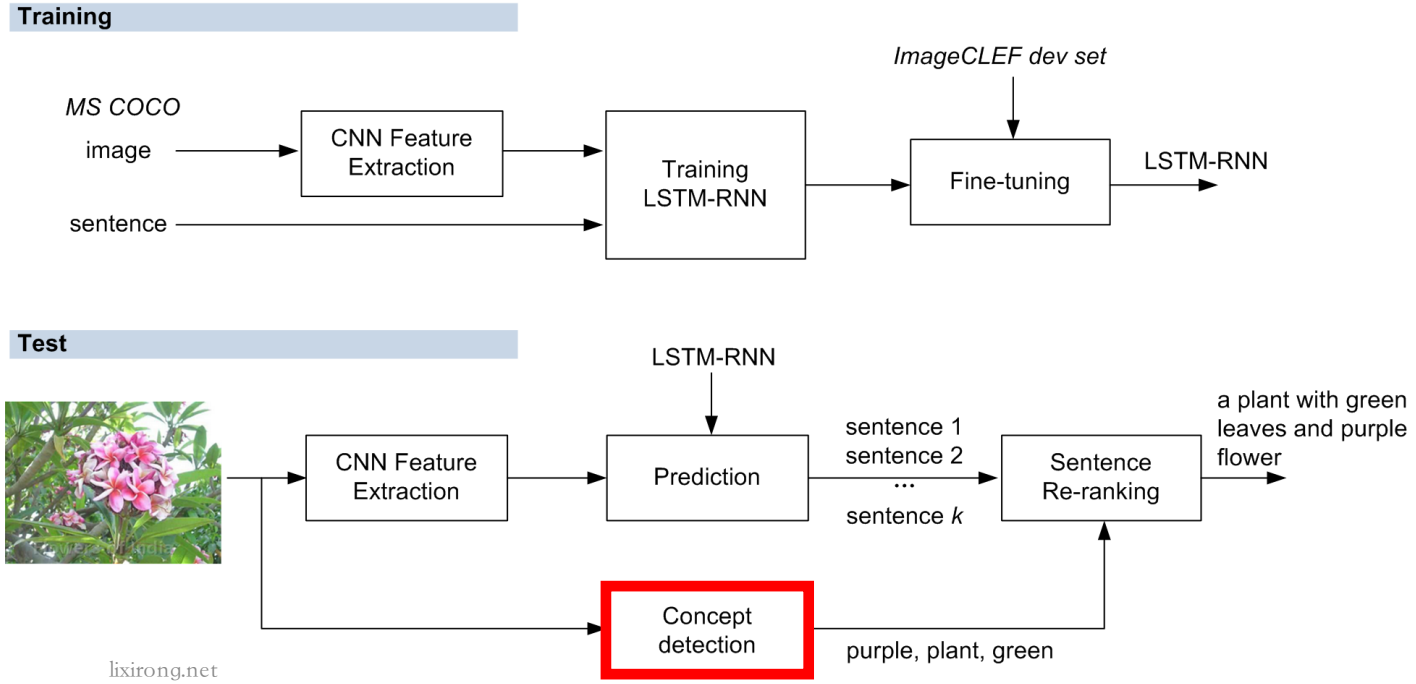
Our Motives

- Many more concept-level annotations than sentence-level annotations
  - 7 million+ hits when searching for 'dog' on the Flickr website
  - Many visual concepts can be effectively predicted

- Can we utilize (noise) concept-level annotations for image captioning?

Our Proposal

- Improving SOTA by concept-based sentence reranking



Sentence Re-Ranking

- Re-compute sentence score in terms of concept detection results

weight * sentScore(hypoSent) + (1-weight) * conceptScore(hypoSent)

- hypoSent: a candidate sentence predicted by the LSTM model
- sentScore(hypoSent): initial score
- conceptScore(hypoSent): averaged score of concepts found in the hypoSent

Sentence Re-Ranking: Real Example

(1) a plane taking off from a runway
(2) a plane taking off from the runway
(3) a plane flying in the sky
(4) an airplane flying in the sky
(5) a plane flying in the air
(6) an airplane is flying in the sky
(7) a plane taking off into the sky

(3) a plane flying in the sky
[flying | plane | sky]

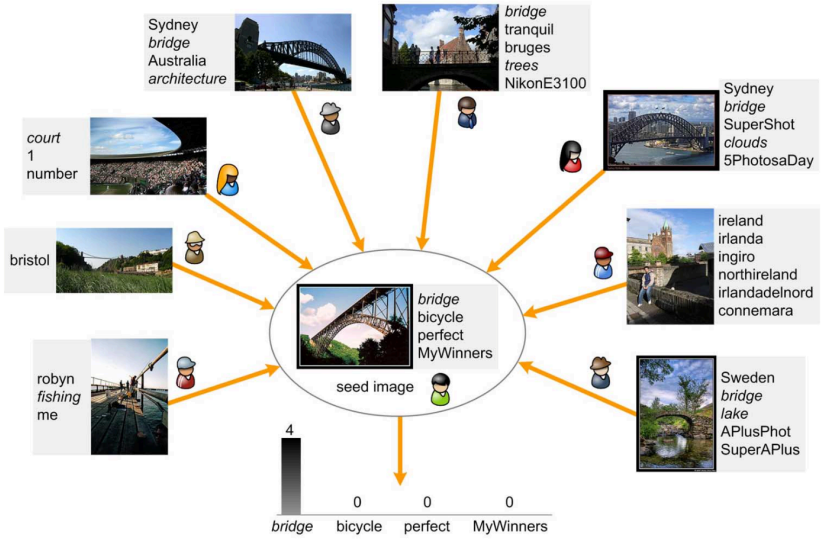
Sentence Re-Ranking: Real Example

(1) a plant with green leaves
(2) a cluster of yellow flowers
(3) a plant is growing in a garden
(4) a plant with green leaves and green leaves
(5) a plant with green leaves and purple flowers
(6) a cluster of yellow flowers in a garden
(7) a cluster of yellow flowers in a field

(5) a plant with green leaves and purple flowers
[purple | plant | green]

Models for Concept Detection I

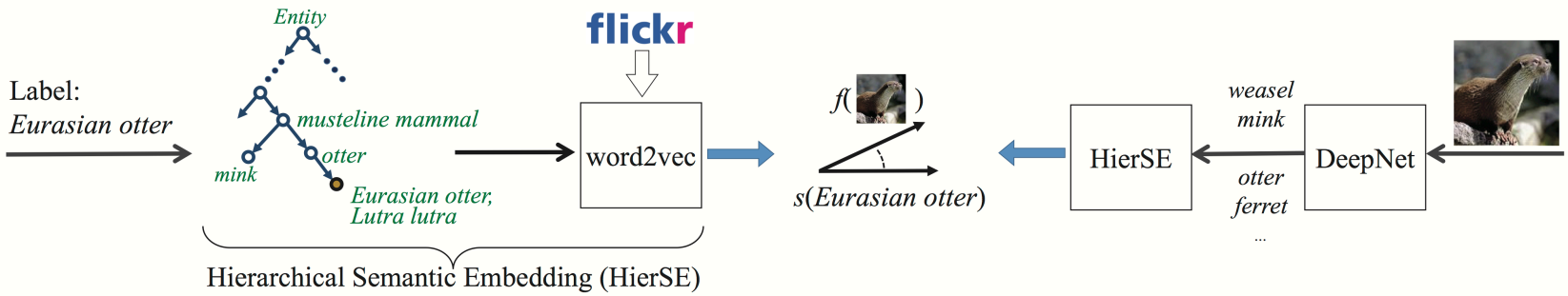
- Neighbor voting (NeiVote)[Li TMM2009]
  - Count image neighbors labeled with a specific concept



Li et al., Socializing the semantic gap: A comparative survey on image tag assignment, refinement and retrieval, *ACM Computing Surveys* 2016

Models for Concept Detection II

- Hierarchical Semantic Embedding (HierSE) [Li SIGIR2015]
  - Compute image-concept relevance in a Word2Vec space [Mikolov NIPS2013]
  - Faster than NeiVote



Evaluation

- ImageCLEF 2015 Image Sentence Generation Task[Gilbert CLEF2015]
  - A dev set of 2k images with manually written captions
    - 5 to 51 captions per image
  - Participants were asked to predict captions for 500k web images
    - The organizers select a subset of 3,070 images (which we do not know) for test
    - Impossible to manually tune for "good" performance
  - Evaluation criterion
    - METEOR

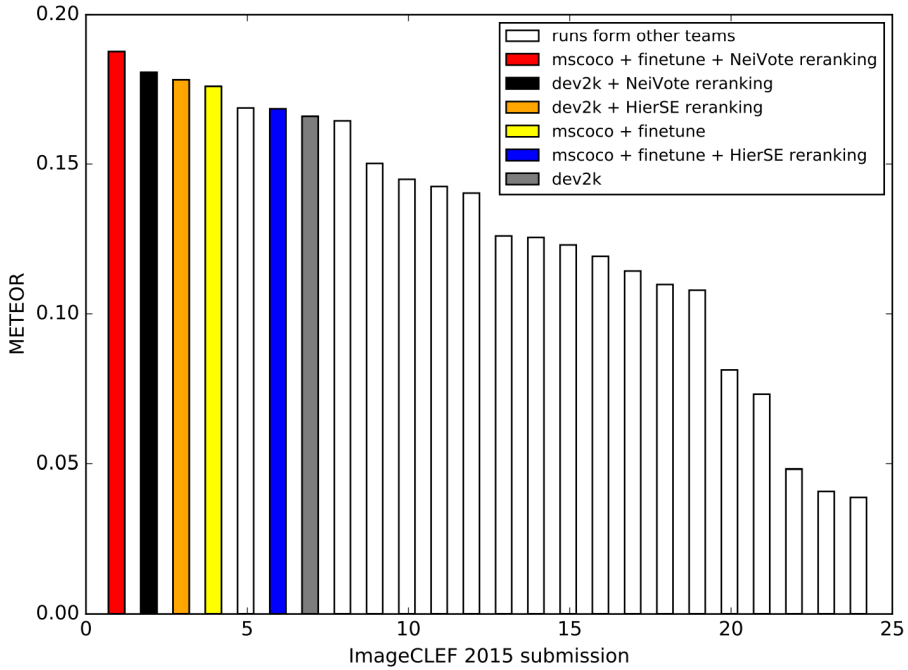
Results

- Performance of our system under varied settings

	Training data	Data for fine tuning	Concept detection	METEOR score
baseline →	ImageCLEF dev	–	–	0.1659
	MSCOCO	ImageCLEF dev	–	0.1759
	ImageCLEF dev	–	HierSE	0.1781
	ImageCLEF dev	–	NeiVote	0.1806
	MSCOCO	ImageCLEF dev	HierSE	0.1684
	MSCOCO	ImageCLEF dev	NeiVote	**0.1875**

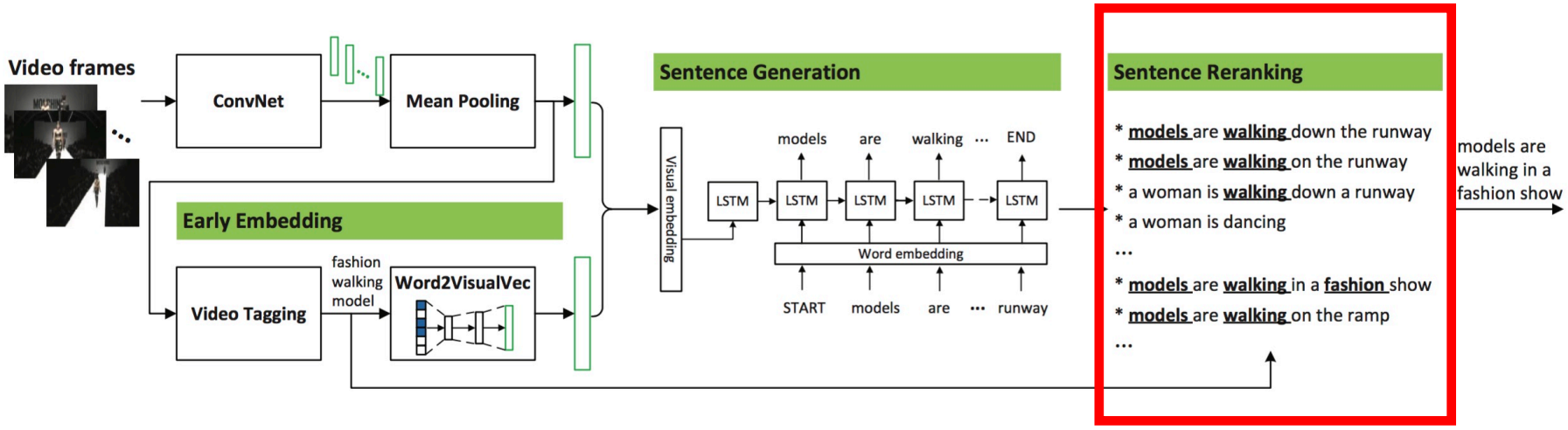
Results

- Comparing with other submissions in ImageCLEF 2015
  - Our runs are shown in color bars



Extension to the video domain

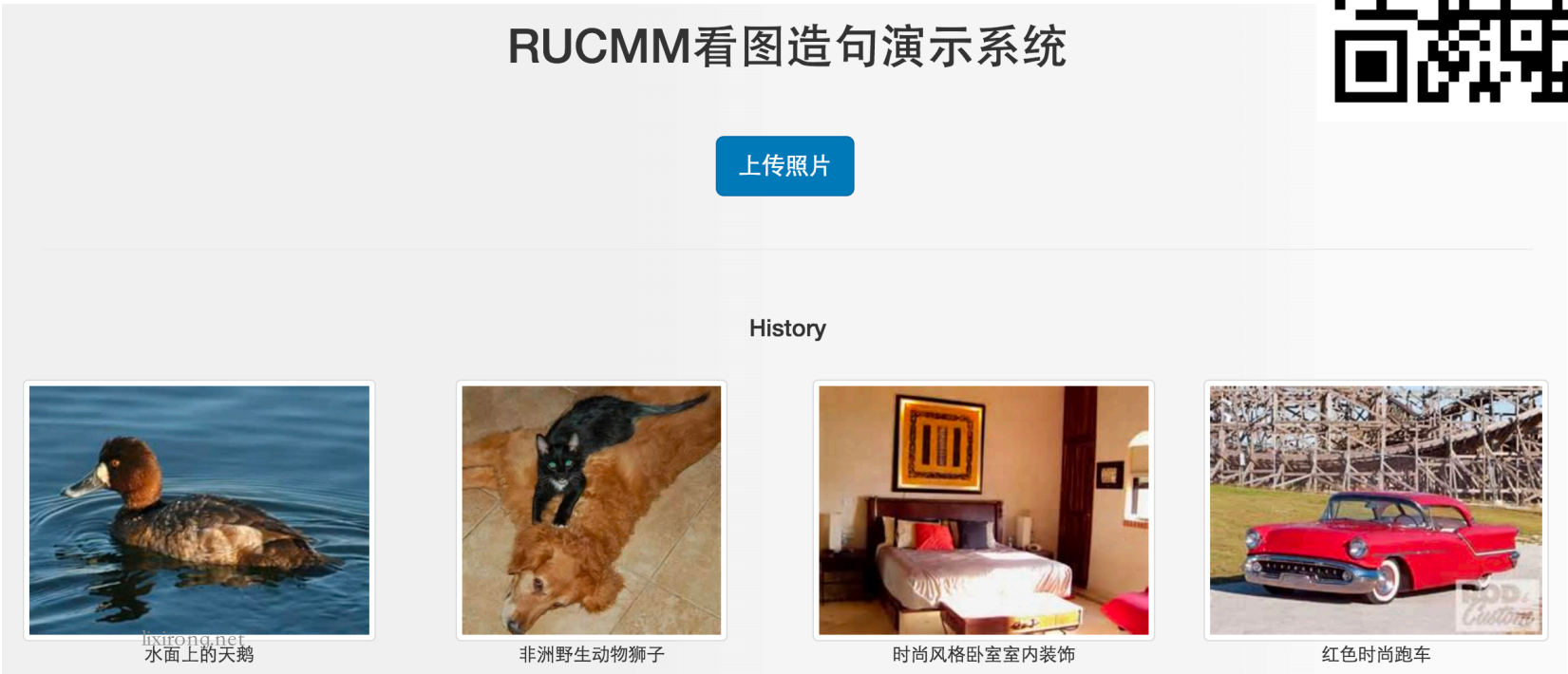
- Concept-based sentence reranking also works for video captioning



Dong et al., Early embedding and late reranking for video captioning, *ACM Multimedia* 2016

Conclusions

- Concept-based sentence reranking is helpful
  - Given proper concept detection (vocabulary, model, ⋯)
- Improve image captioning with no extra cost of manual annotation
  - Learning from Flickr images
- Winning entry in ImageCLEF 2015

Demo: Image captioning in Chinese

http://tinyurl.com/imcap101