Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)

http://www.stat.ucla.edu/~junhua.mao/m-RNN.html

Junhua Mao[1,2], Wei Xu[1], Yi Yang[1], Jiang Wang[1], Zhiheng Huang[1], Alan Yuille[2]

[1]Baidu Research [2]University of California, Los Angeles

a close up of a bowl of food on a table

a train is traveling down the tracks in a city

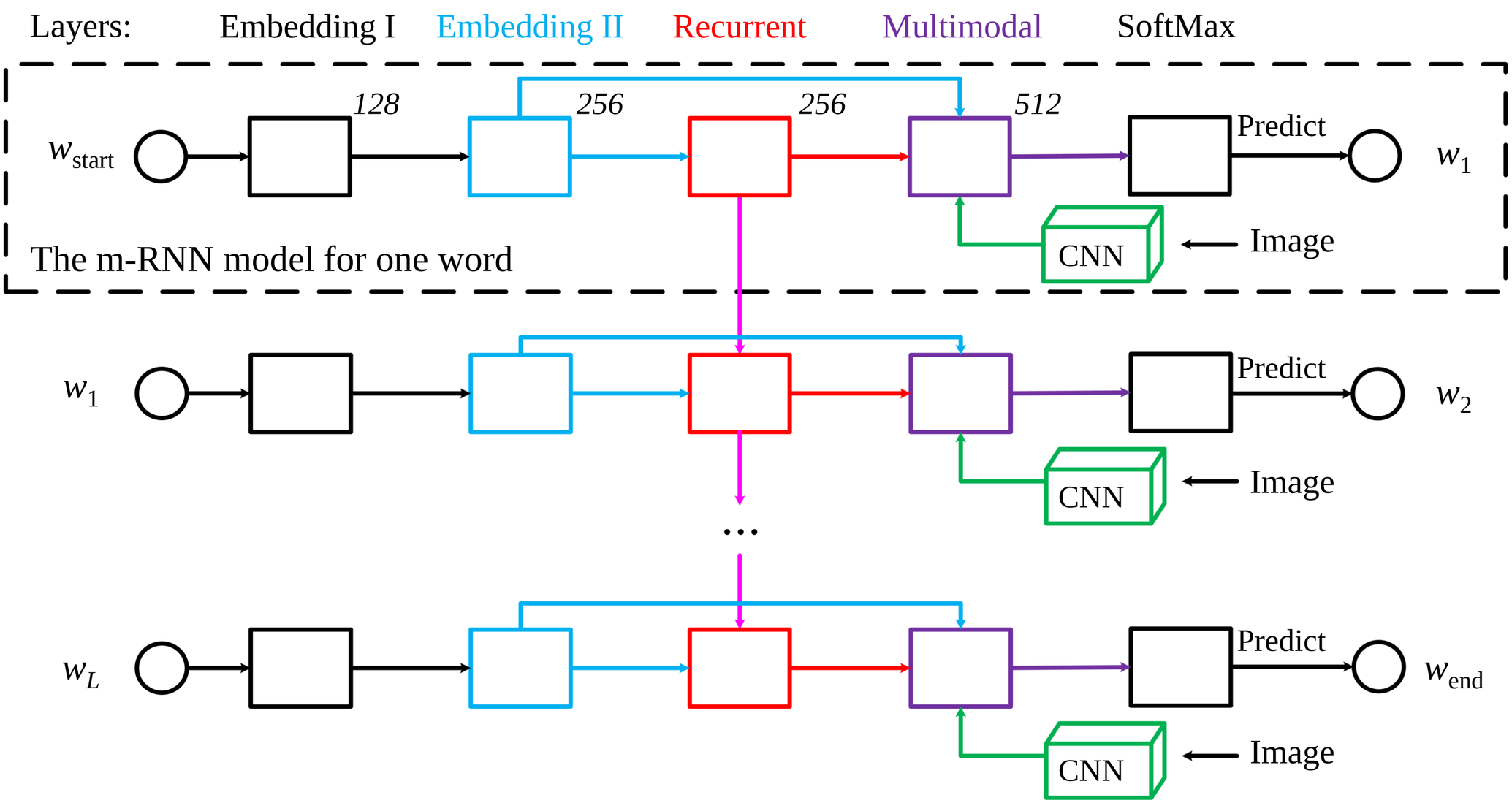
a pizza sitting on top of a table next to a box of pizza

Baidu Research UCLA

Abstract

- Three Tasks:
  - Image caption generation
  - Image retrieval (given query sentence)
  - Sentence retrieval (given query image)
- One model (m-RNN):
  - A deep Recurrent NN (RNN) for the sentences
  - A deep Convolutional NN (CNN) for the images
  - A multimodal layer connects the first two components
- State-of-the-art Performance:
  - For three tasks
  - On four datasets: IAPR TC-12 [Grubinger et al. 06'], Flickr 8K [Rashtchian et al. 10'], Flickr 30K [Young et al. 14'] and MS COCO [Lin et al. 14']

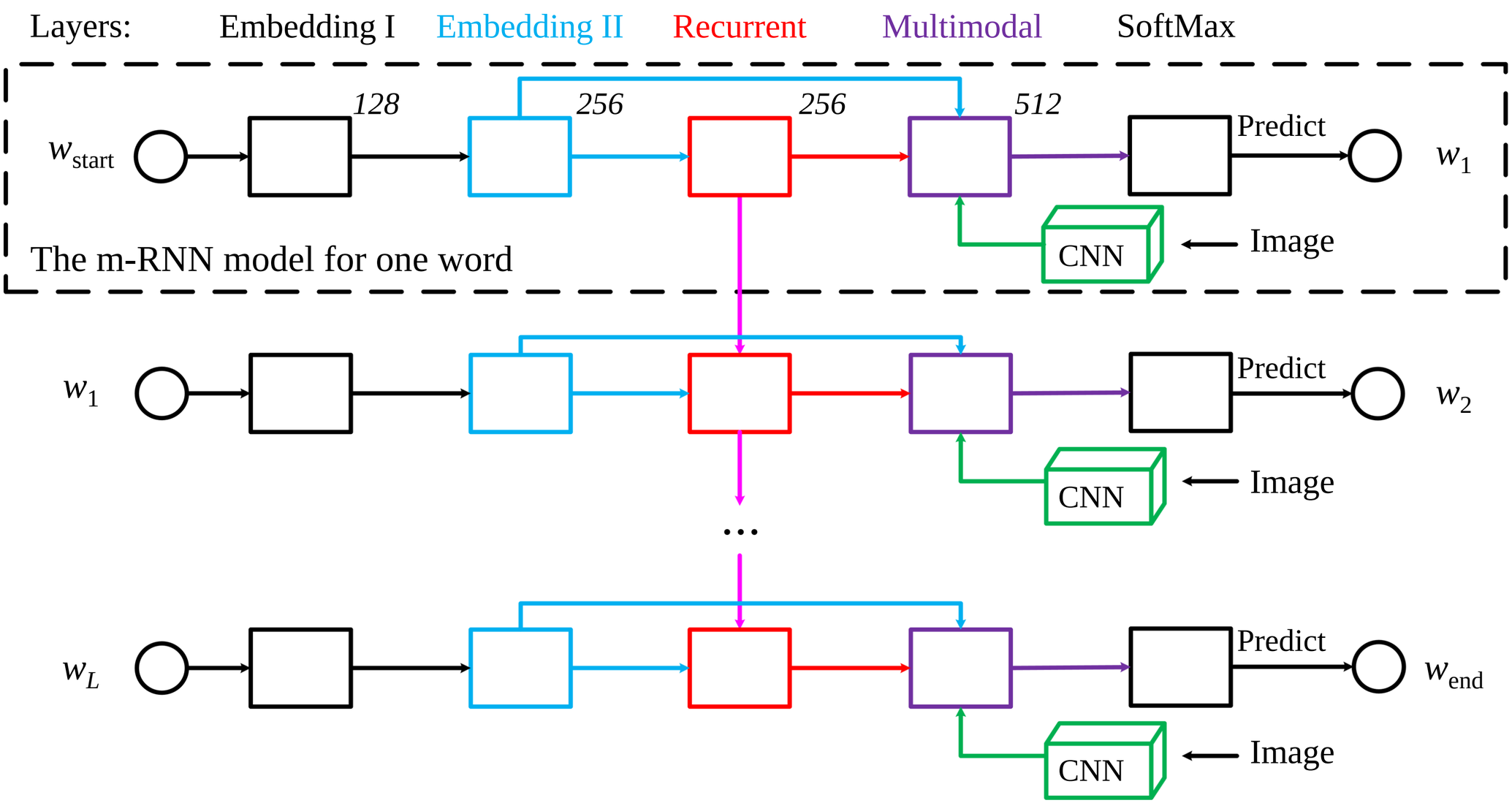
The m-RNN Model



$w_1, w_2, \ldots, w_L$ is the sentence description of the image

w_{start}, w_{end} is the start and end sign of the sentence

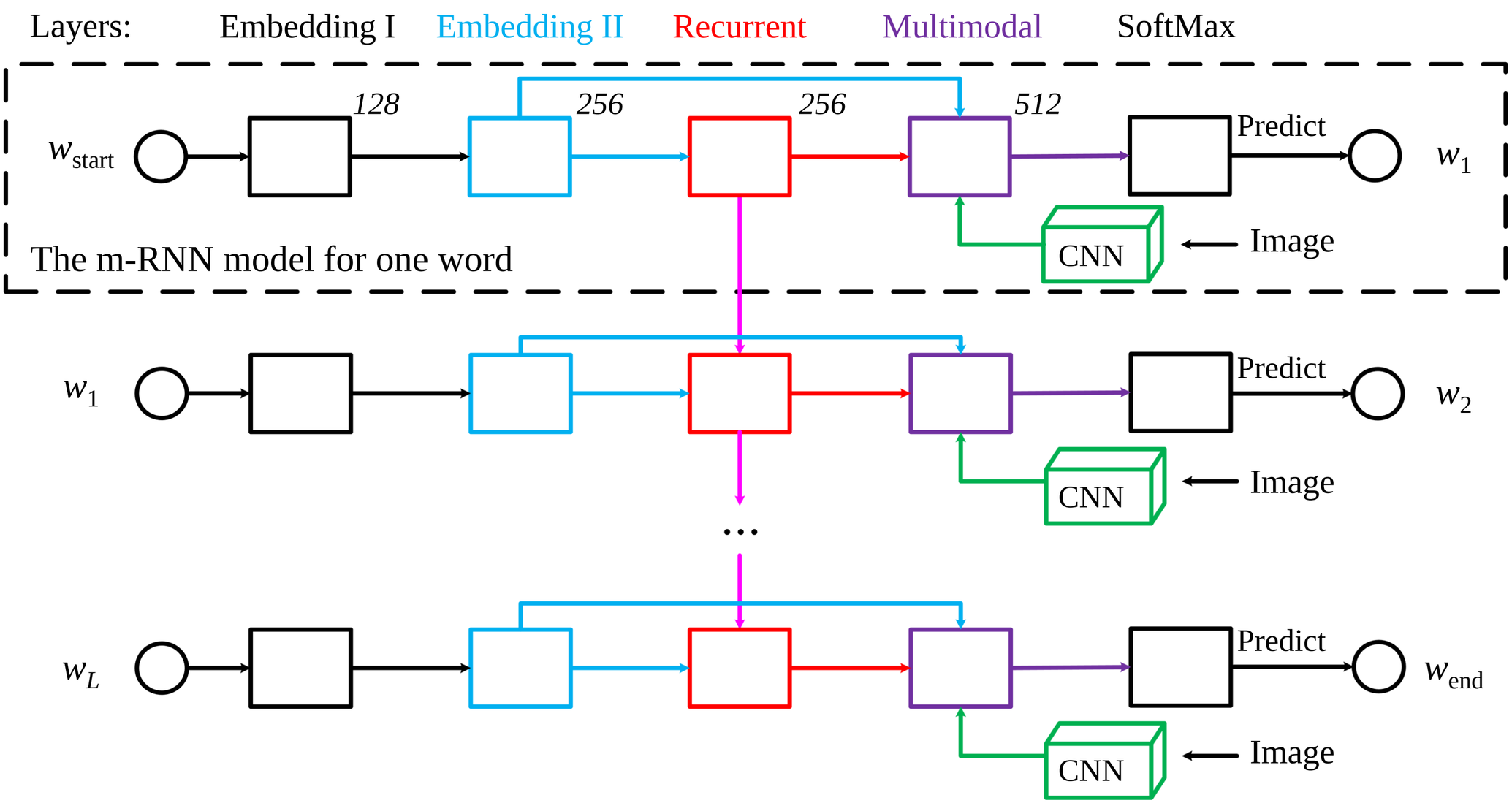
The m-RNN Model



Detailed calculation for recurrent $\mathbf{r}(t)$ and multimodal layer $\mathbf{m}(t)$

- $\mathbf{r}(t) = f(\mathbf{U}_r \cdot \mathbf{r}(t-1) + \mathbf{w}(t))$, $\mathbf{w}(t)$ is the activation of embedding layer II for the word w_t
- $\mathbf{m}(t) = g(\mathbf{V}_w \cdot \mathbf{w}(t) + \mathbf{V}_r \cdot \mathbf{r}(t) + \mathbf{V}_I \cdot \mathbf{I})$, $\mathbf{I}$ is the image representation
- "+" here means element-wise plus

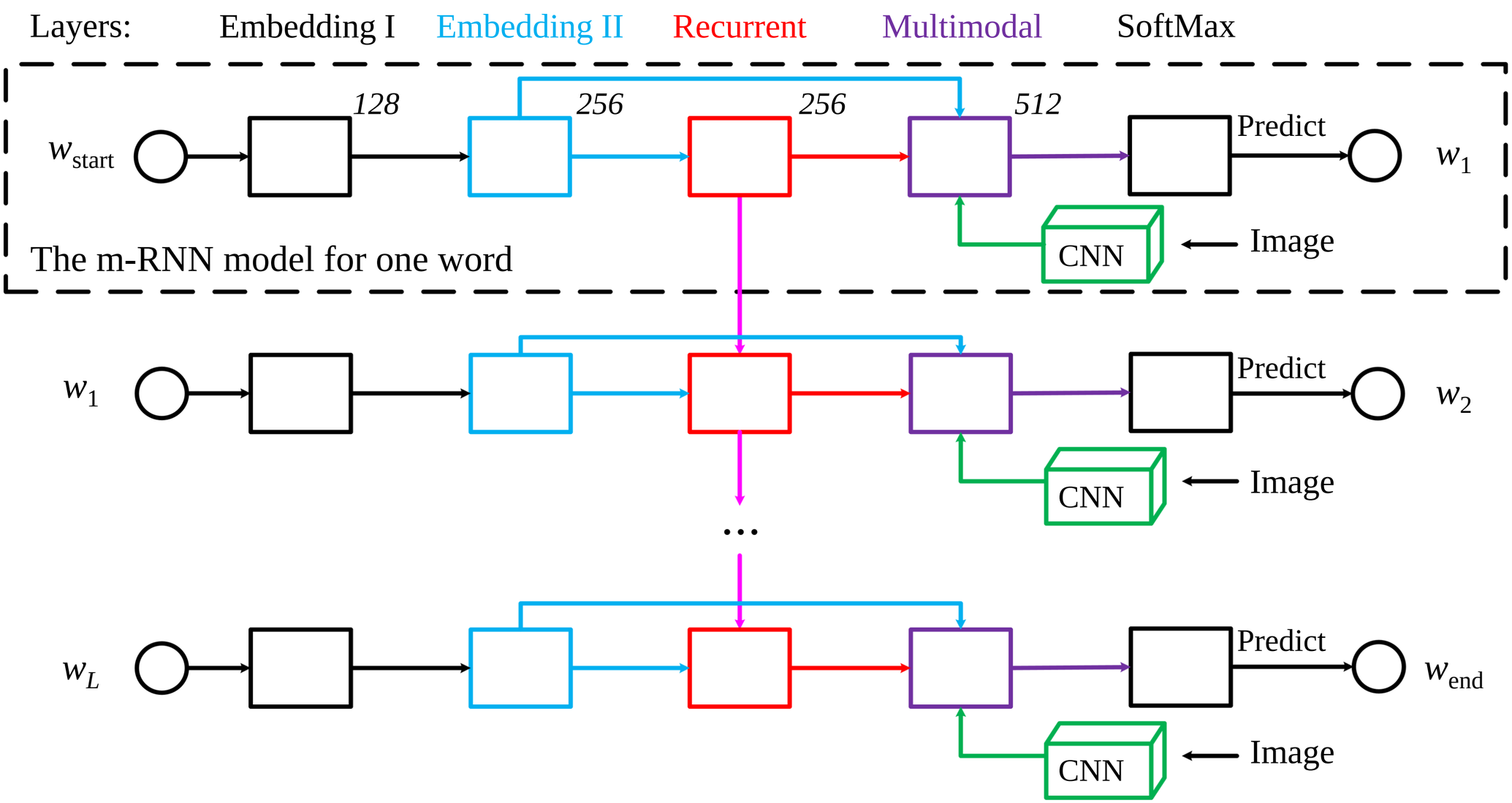
The m-RNN Model



Non-linear activation functions:

- For the recurrent layer: ReLU [Nair and Hinton 10'] $f(x) = \max(0, x)$
- For the word embedding layers and the multimodal layer: S-tanh [LeCun et.al 12']: $g(x) = 1.72 \cdot \tanh(\frac{2}{3}x)$

The m-RNN Model



The output of the trained model:

$$P(w_n|w_{1:n-1}, \mathbf{I})$$

Application

- Image caption generation:
  - Begin with the start sign w_{start}
  - Sample next word from $P(w_n|w_{1:n-1}, \mathbf{I})$
  - Repeat until the model generates the end sign w_{end}

Application

- Image caption generation:
  - Begin with the start sign w_{start}
  - Sample next word from $P(w_n|w_{1:n-1}, \mathbf{I})$
  - Repeat until the model generates the end sign w_{end}
- Image retrieval given query sentence:
  - Ranking score: $P\left(w_{1:L}^{Q}\middle|\mathbf{I}^{D}\right) = \prod_{n=2}^{L} P\left(w_{n}^{Q}\middle|w_{1:n-1}^{Q}, \mathbf{I}^{D}\right)$
  - Output the top ranked images

Application

- Image caption generation:
  - Begin with the start sign w_{start}
  - Sample next word from $P(w_n|w_{1:n-1}, \mathbf{I})$
  - Repeat until the model generates the end sign w_{end}
- Image retrieval given query sentence:
  - Ranking score: $P\left(w_{1:L}^{Q}\middle|\mathbf{I}^{D}\right) = \prod_{n=2}^{L} P\left(w_{n}^{Q}\middle|w_{1:n-1}^{Q}, \mathbf{I}^{D}\right)$
  - Output the top ranked images
- Sentence retrieval given query image:
  - *Problem*: Some sentences have high probability for any image query
  - *Solution*: **Normalize** the probability. $\mathbf{I}'$ are images sampled from the training set:

$$\frac{P\left(w_{1:L}^{D}\middle|\mathbf{I}^{Q}\right)}{P\left(w_{1:L}^{D}\right)} \qquad P\left(w_{1:L}^{D}\right) = \sum_{\mathbf{I}'} P\left(w_{1:L}^{D}\middle|\mathbf{I}'\right) \cdot P(\mathbf{I}')$$

Application

- Image caption generation:
  - Begin with the start sign w_{start}
  - Sample next word from $P(w_n|w_{1:n-1}, \mathbf{I})$
  - Repeat until the model generates the end sign w_{end}
- Image retrieval given query sentence:
  - Ranking score: $P\left(w_{1:L}^{Q}\middle|\mathbf{I}^{D}\right) = \prod_{n=2}^{L} P\left(w_{n}^{Q}\middle|w_{1:n-1}^{Q}, \mathbf{I}^{D}\right)$
  - Output the top ranked images
- Sentence retrieval given query image:
  - *Problem*: Some sentences have high probability for any image query
  - *Solution*: **Normalize** the probability. $\mathbf{I}'$ are images sampled from the training set:

$$\frac{P\left(w_{1:L}^{D}\middle|\mathbf{I}^{Q}\right)}{P\left(w_{1:L}^{D}\right)} \qquad P\left(w_{1:L}^{D}\right) = \sum_{\mathbf{I}'} P\left(w_{1:L}^{D}\middle|\mathbf{I}'\right) \cdot P(\mathbf{I}')$$

  - Equivalent to using a ranking score: $P\left(\mathbf{I}^{Q}\middle|w_{1:L}^{D}\right) = \frac{P\left(w_{1:L}^{D}\middle|\mathbf{I}^{Q}\right) \cdot P(\mathbf{I}^{Q})}{P\left(w_{1:L}^{D}\right)}$

Experiment: Retrieval

Table 1. Retrieval results on Flickr 30K and MS COCO

	Sentence Retrival (Image to Text)				Image Retrival (Text to Image)			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Flickr30K								
Random	0.1	0.6	1.1	631	0.1	0.5	1.0	500
DeepFE-RCNN (Karpathy et al. 14′)	16.4	40.2	54.7	8	10.3	31.4	44.5	13
RVR (Chen & Zitnick 14′)	12.1	27.8	47.8	11	12.7	33.1	44.9	12.5
MNLM-AlexNet (Kiros et al. 14′)	14.8	39.2	50.9	10	11.8	34.0	46.3	13
MNLM-VggNet (Kiros et al. 14′)	23.0	50.7	62.9	5	16.8	42.0	56.5	8
NIC (Vinyals et al. 14′)	17.0	56.0	/	7	17.0	**57.0**	/	7
LRCN (Donahue et al. 14′)	14.0	34.9	47.0	11	/	/	/	/
DeepVS-RCNN (Karpathy et al. 14′)	22.2	48.2	61.4	4.8	15.2	37.7	50.5	9.2
Ours-**m-RNN**-AlexNet	18.4	40.2	50.9	10	12.6	31.2	41.5	16
Ours-**m-RNN**-VggNet	**35.4**	**63.8**	**73.7**	**3**	**22.8**	50.7	**63.1**	**5**
MS COCO								
Random	0.1	0.6	1.1	631	0.1	0.5	1.0	500
DeepVS-RCNN (Karpathy et al. 14′)	29.4	62.0	75.9	2.5	20.9	**52.8**	69.2	4
Ours-**m-RNN**-VggNet	**41.0**	**73.0**	**83.5**	**2**	**29.0**	42.2	**77.0**	**3**

R@K: The recall rate of the groundtruth among the top K retrieved candidates

Med r: Median rank of the top-ranked retrieved groundtruth

(*) Results reported on 04/10/2015. The deadline for our camera ready submission.

Experiment: Captioning

Table 2. Caption generation results on Flickr 30K and MS COCO

	Flickr30K					MS COCO				
	PERP	B-1	B-2	B-3	B-4	PERP	B-1	B-2	B-3	B-4
RVR (Chen & Zitnick 14′)	-	-	-	-	0.13	-	-	-	-	0.19
DeepVS-AlexNet (Karpathy et al. 14′)	-	0.47	0.21	0.09	-	-	0.53	0.28	0.15	-
DeepVS-VggNet (Karpathy et al. 14′)	21.20	0.50	0.30	0.15	-	19.64	0.57	0.37	0.19	-
NIC (Vinyals et al. 14′)	-	**0.66**	-	-	-	-	**0.67**	-	-	-
LRCN (Donahue et al. 14′)	-	0.59	0.39	0.25	0.16	-	0.63	0.44	0.31	0.21
DMSM (Fang et al. 14′)	-	-	-	-	-	-	-	-	-	0.21
Ours-m-RNN-AlexNet	35.11	0.54	0.36	0.23	0.15	-	-	-	-	-
Ours-m-RNN-VggNet	**20.72**	0.60	**0.41**	**0.28**	**0.19**	**13.60**	**0.67**	**0.49**	**0.34**	**0.24**

B-K: BLEU-K score

PERP: Perplexity $\mathcal{PPL}$ $\log_2 \mathcal{PPL}(w_{1:L}|\mathbf{I}) = -\frac{1}{L}\sum_{n=1}^{L} \log_2 P(w_n|w_{1:n-1}, \mathbf{I})$

(*) Results reported on 04/10/2015. The deadline for our camera ready submission.

Experiment: Captioning

Table 4. Results on the MS COCO test set

	B1	B2	B3	B4	CIDEr	$ROUGE_L$	METEOR
Human-c5 (**)	0.663	0.469	0.321	0.217	0.854	0.484	0.252
m-RNN-c5	0.668	0.488	0.342	0.239	0.729	0.489	0.221
m-RNN-beam-c5	0.680	0.506	0.369	0.272	0.791	0.499	0.225
Human-c40 (**)	0.880	0.744	0.603	0.471	0.910	0.626	0.335
m-RNN-c40	0.845	0.730	0.598	0.473	0.740	0.616	0.291
m-RNN-beam-c40	0.865	0.760	0.641	0.529	0.789	0.640	0.304

c5 and c40: evaluated using 5 and 40 reference sentences respectively.

"-beam" means that we generate a set of candidate sentences, and then selects the best one. (beam search)

(**) Provided in https://www.codalab.org/competitions/3221#results

(***) We evaluate it on the MS COCO evaluation server: https://www.codalab.org/competitions/3221

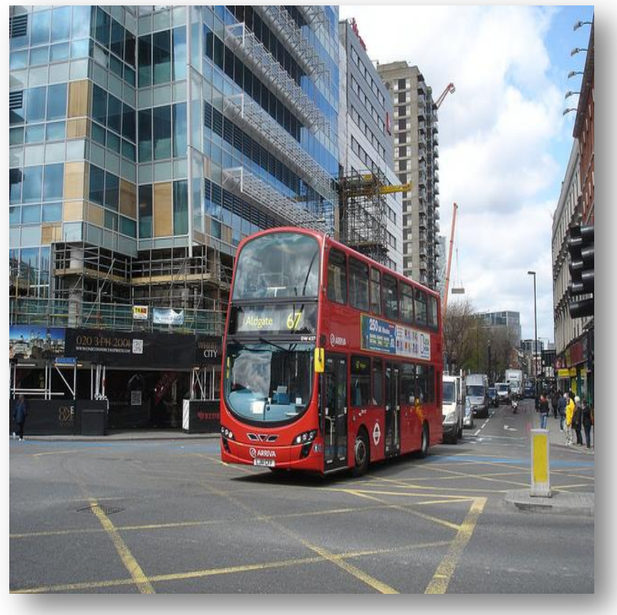
Discussion

Discussion

Other language: Chinese

一个年轻的男孩坐在长椅上。

一列火车在轨道上行驶。



一辆双层巴士停在一个城市街道上。

We acknowledge Haoyuan Gao and Zhiheng Huang from Baidu Research for designing the Chinese image captioning system

Discussion

Other language: Chinese

一个年轻的男孩坐在长椅上。

A young boy sitting on a bench.

一列火车在轨道上行驶。

A train running on the track.

一辆双层巴士停在一个城市街道上。

A double decker bus stop on a city street.

We acknowledge Haoyuan Gao and Zhiheng Huang from Baidu Research for designing the Chinese image captioning system

Discussion

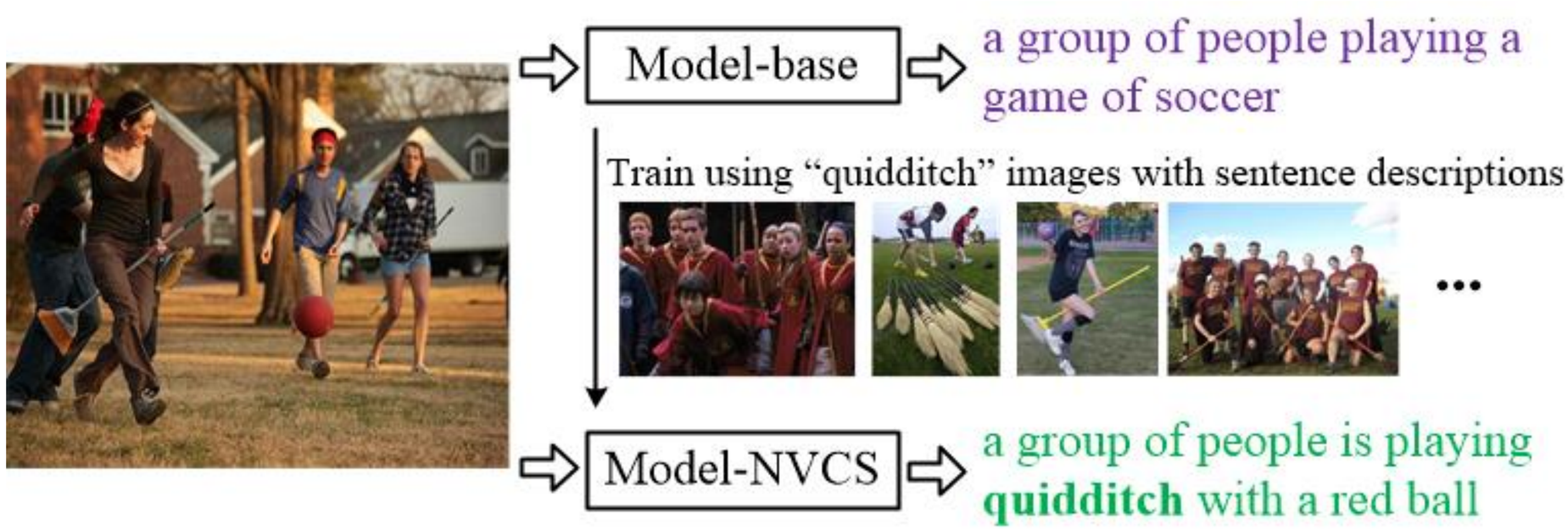
Can we design a system that learns to describe new visual concepts from a few examples?

Discussion

Can we design a system that learns to describe new visual concepts from a few examples?

Learning like a Child: Fast Novel Visual Concept Learning from Sentence Descriptions of Images, arXiv 1504.06692

- Efficiently enlarge the vocabulary
- Needs only *a few images* with only *a few minutes*
- Datasets for evaluation



Thank you

For more details, please visit the project page:
http://www.stat.ucla.edu/~junhua.mao/m-RNN.html

The updated version of our paper: http://arxiv.org/abs/1412.6632
J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, A. Yuille, "Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)", *arXiv:1412.6632*.

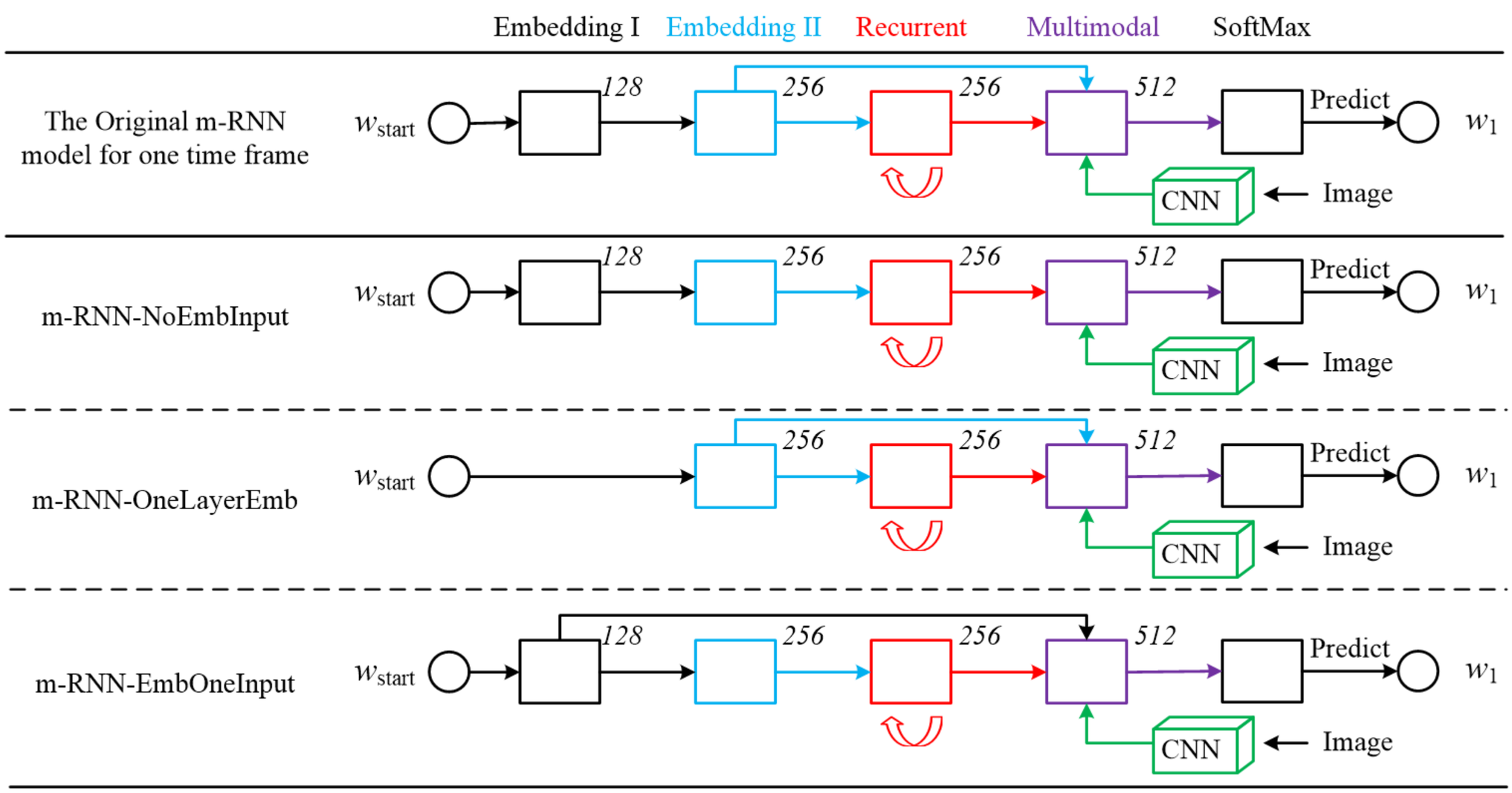
The novel visual concept learning paper: http://arxiv.org/abs/1504.06692
J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, A. Yuille, "Learning like a Child: Fast Novel Visual Concept Learning from Sentence Descriptions of Images", *arXiv:1412.6632*.

a young girl brushing his teeth with a toothbrush

a group of people flying kites in a field

a man is doing a trick on a skateboard

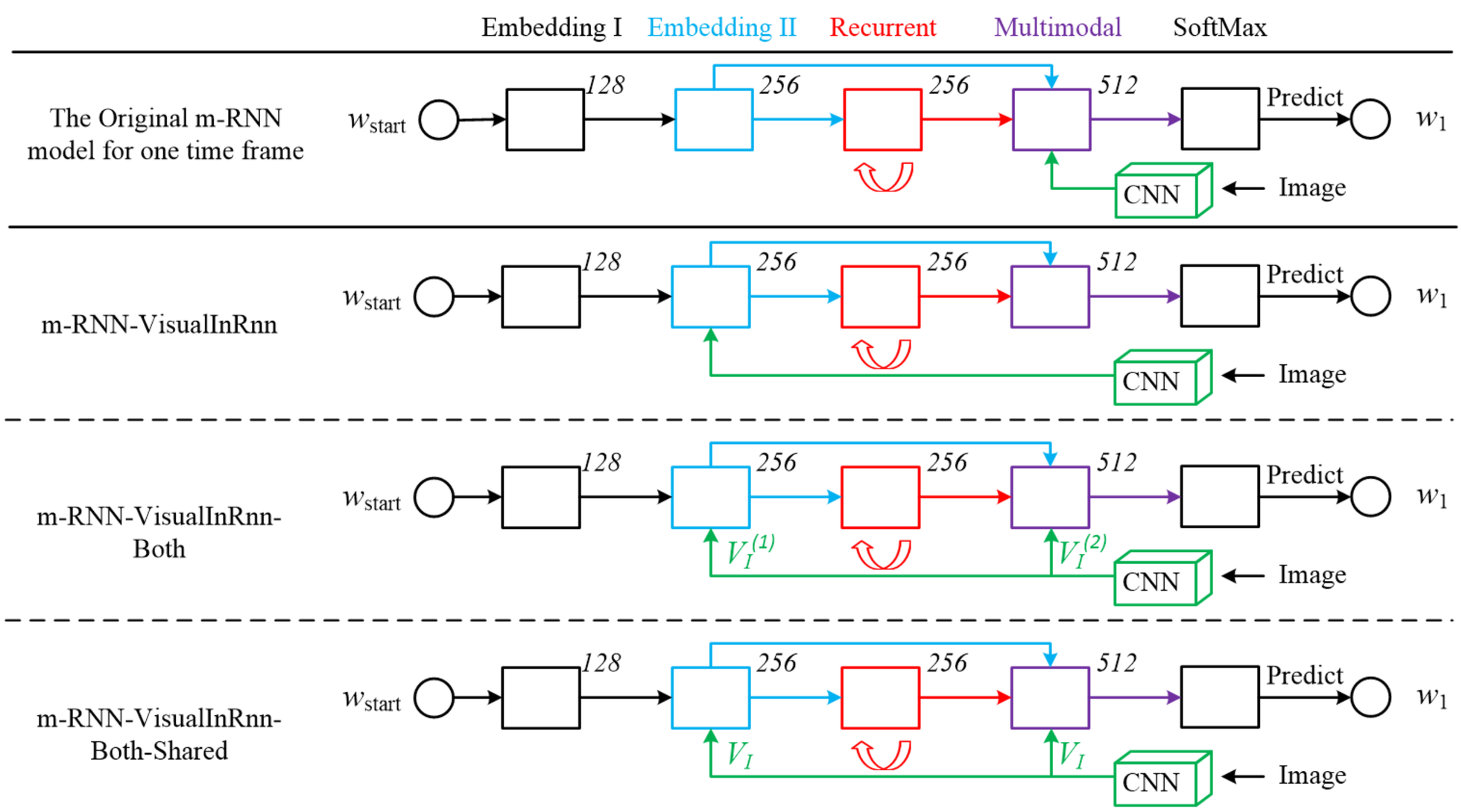
Appendix



	B-1	B-2	B-3	B-4
m-RNN	0.600	0.412	0.278	0.187
m-RNN-NoEmbInput	0.592	0.408	0.277	0.188
m-RNN-OneLayerEmb	0.594	0.406	0.274	0.184
m-RNN-EmbOneInput	0.590	0.406	0.274	0.185

Performance comparison with different word-embedding configuration

Appendix



	B-1	B-2	B-3	B-4
m-RNN	0.600	0.412	0.278	0.187
m-RNN-visInRnn	0.466	0.267	0.157	0.101
m-RNN-visInRnn-both	0.546	0.333	0.191	0.120
m-RNN-visInRnn-both-shared	0.478	0.279	0.171	0.110

Performance comparison with different image representation input methods

Table 3. Recurrent layer size and whether LSTM is used

	MNLM	NIC	LRCN	RVR	DeepVS	Our m-RNN
Size	300	512	1000 (x4)	100	300-600	256
LSTM	Yes	Yes	Yes	No	No	No