课程详情扫码咨询

www.mooc.ai

From
Natural Language Processing
to
Computer Vision + Natural Language Processing

玖强

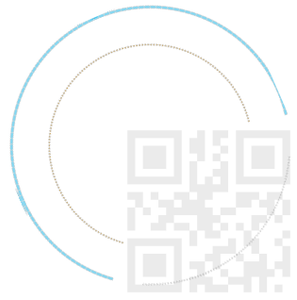
http://jxgu.cc
http://ertou.net

Outline

- **Recent Trends in *Deep Learning Based* NLP [10 Minutes]**
- **CNN-based vs. RNN-based Language Model [15 Minutes]**
- **Reinforcement Learning for NLP [15 Minutes]**
- **Cross-Modality Retrieval [10 Minutes]**
- **Q / A [10 Minutes]**

Recent Trends in Deep Learning Based NLP

- Gu, Jiuxiang, et al. "Recent advances in convolutional neural networks." *Pattern Recognition* (2017).
- 新智元导读：【珍藏】了解CNN这一篇就够了：卷积神经网络技术及发展：http://www.voidcn.com/article/p-zxvegkbr-bnn.html

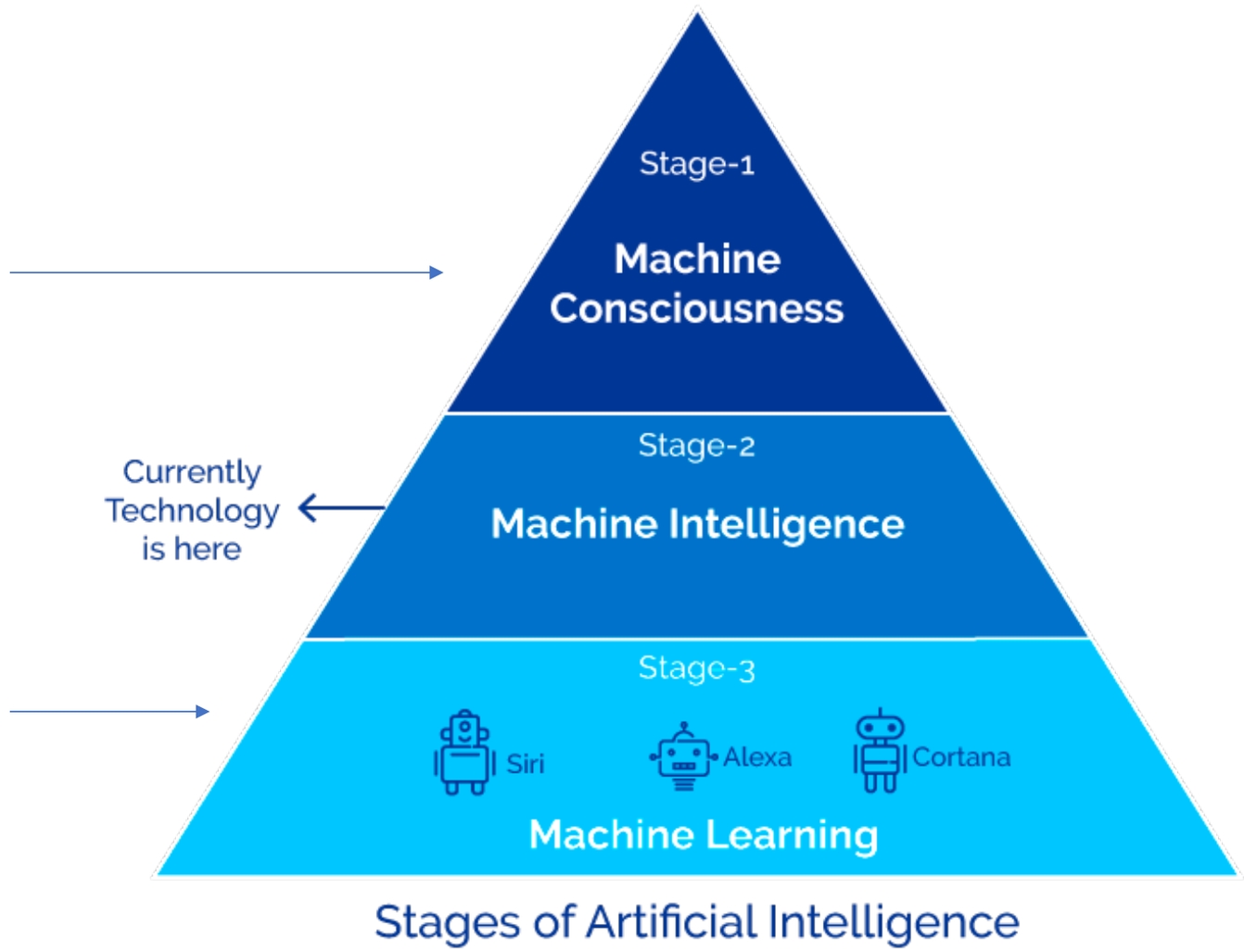


WHAT MAKES SYSTEM AI ENABLED

机器意识

机器智能

机器学习



Stages of Artificial Intelligence

≠ True AI

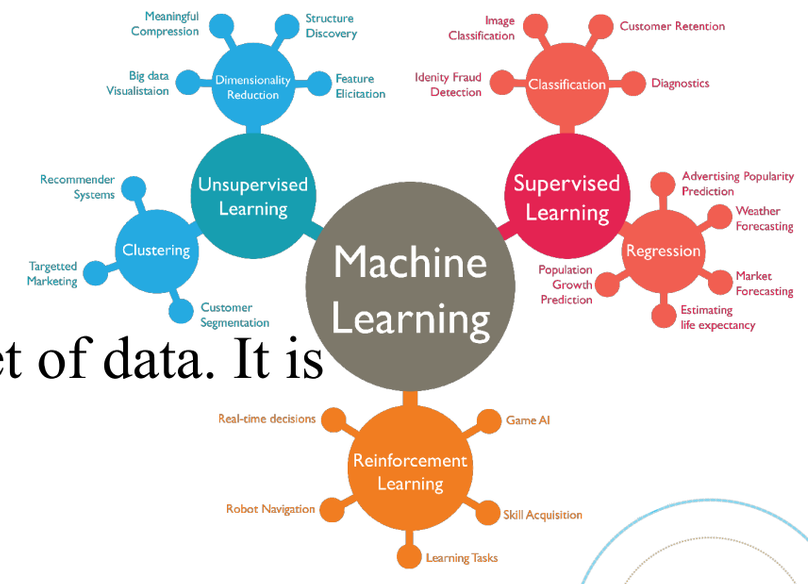
NLP, AI, ML, DL & NN的区别

- **AI or Artificial Intelligence**
  - Building systems that can do intelligent things.

- **NLP or Natural Language Processing**
  - Building systems that can understand language. It is a subset of Artificial Intelligence.

- **ML or Machine Learning**
  - Building systems that can learn from experience. It is also a subset of Artificial Intelligence.

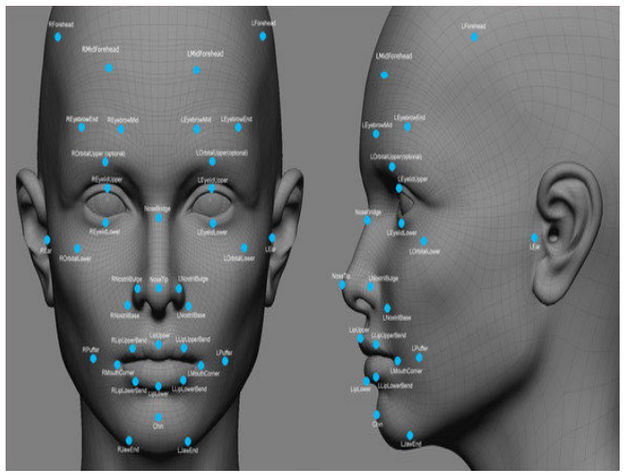
- **NN or Neural Network**
  - Biologically inspired network of Artificial Neurons.



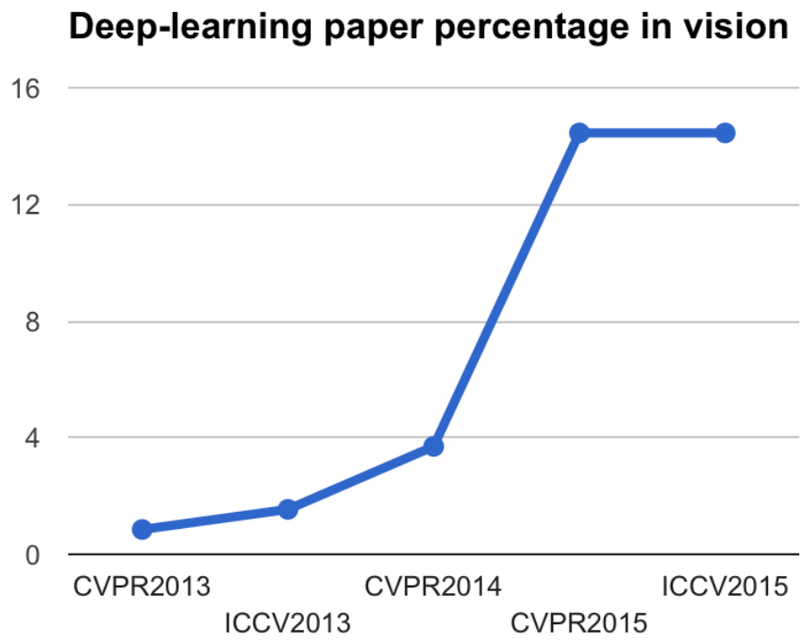
- **DL or Deep Learning**
  - Building systems that use Deep Neural Network on a large set of data. It is a subset of Machine Learning.

Recent Trends in Deep Learning Based NLP

1. Deep Learning in Computer Vision



2. Following this trend, recent NLP research is now increasingly focusing on the use of new deep learning methods.



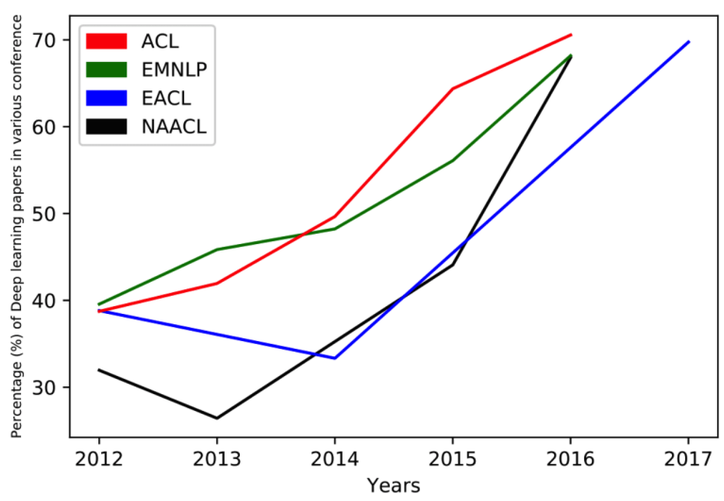


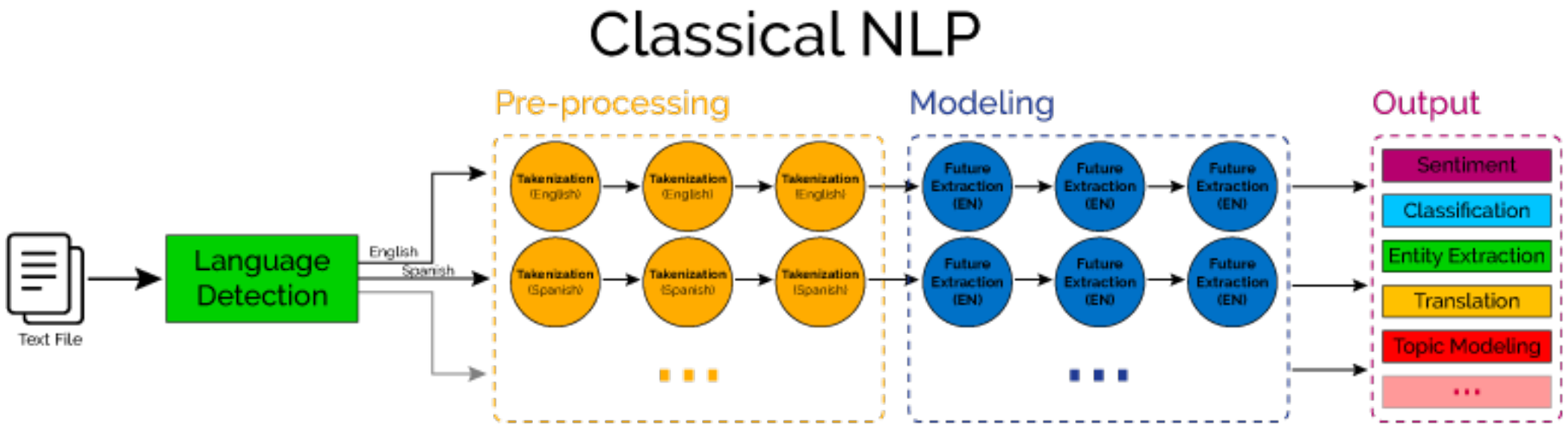
Fig. 1: Percentage of deep learning papers in ACL, EMNLP, EACL, NAACL over the last 6 years (long papers).

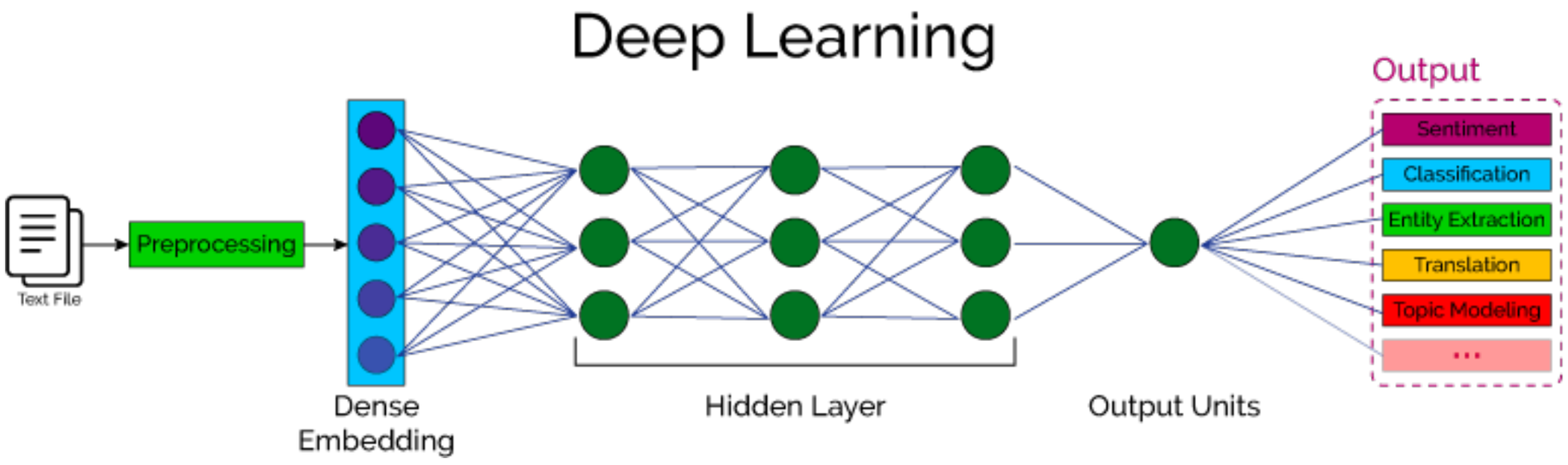
Young, Tom, et al. "Recent trends in deep learning based natural language processing." arXiv preprint arXiv:1708.02709(2017).

COMMON TASKS OF DEEP LEARNING IN NLP

Deep Learning Algorithms	NLP Usage
Neural Network – NN (feed)	•Part-of-speech Tagging •Tokenization •Named Entity Recognition •Intent Extraction
Recurrent Neural Networks -(RNN)	**•Machine Translation** **•Question Answering System** **•Image Captioning** **•Visual Question Answering**
Recursive Neural Networks	**•Parsing sentences** **•Sentiment Analysis** •Paraphrase detection •Relation Classification **•Object detection**
Convolutional Neural Network -(CNN)	**•Sentence/ Text classification** •Relation extraction and classification •Spam detection •Categorization of search queries **•Semantic relation extraction**

Difference Between Classical NLP & Deep Learning NLP





RNN-based Statistics Language Model

RNN-based Statistics Language Model

❑ The Goal of RNN-based LM

➢ ***Predicting the next word in textual data given context***

$$p(w_1, \ldots, w_N) = \prod_{i=1}^{N} p(w_i | w_1, \ldots, w_{i-1})$$

➢ Recurrent Neural Networks based LMs employ the chain rule to model joint probabilities over word sequences, where the context of all previous words is encoded with an LSTM, and the probability over words uses a Softmax.

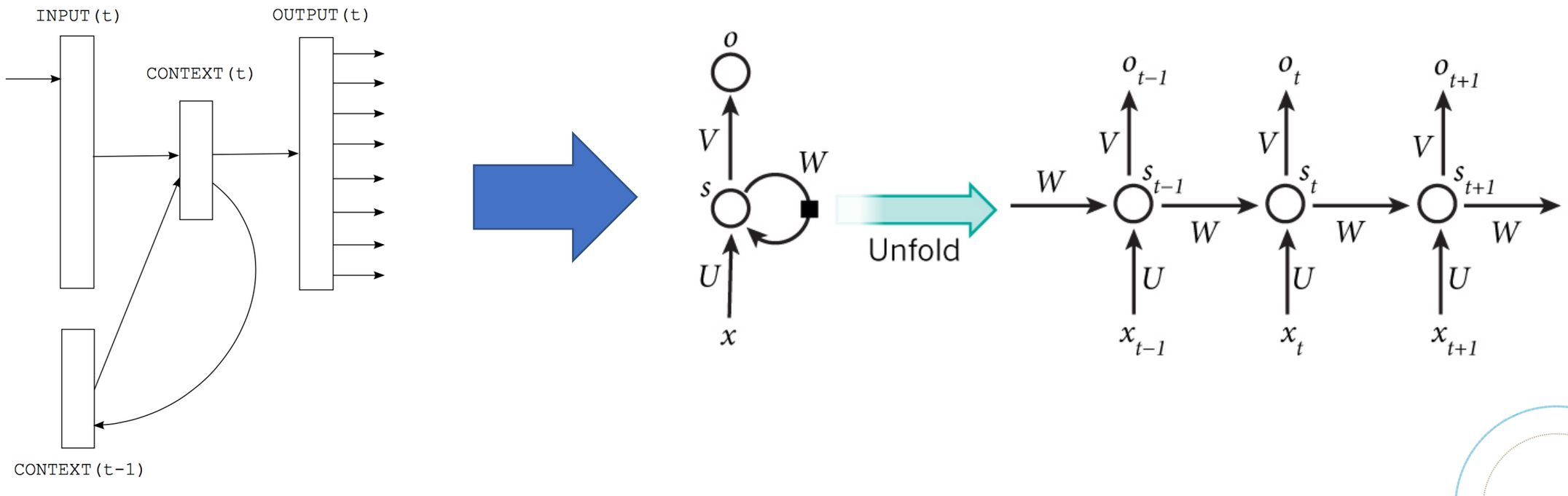


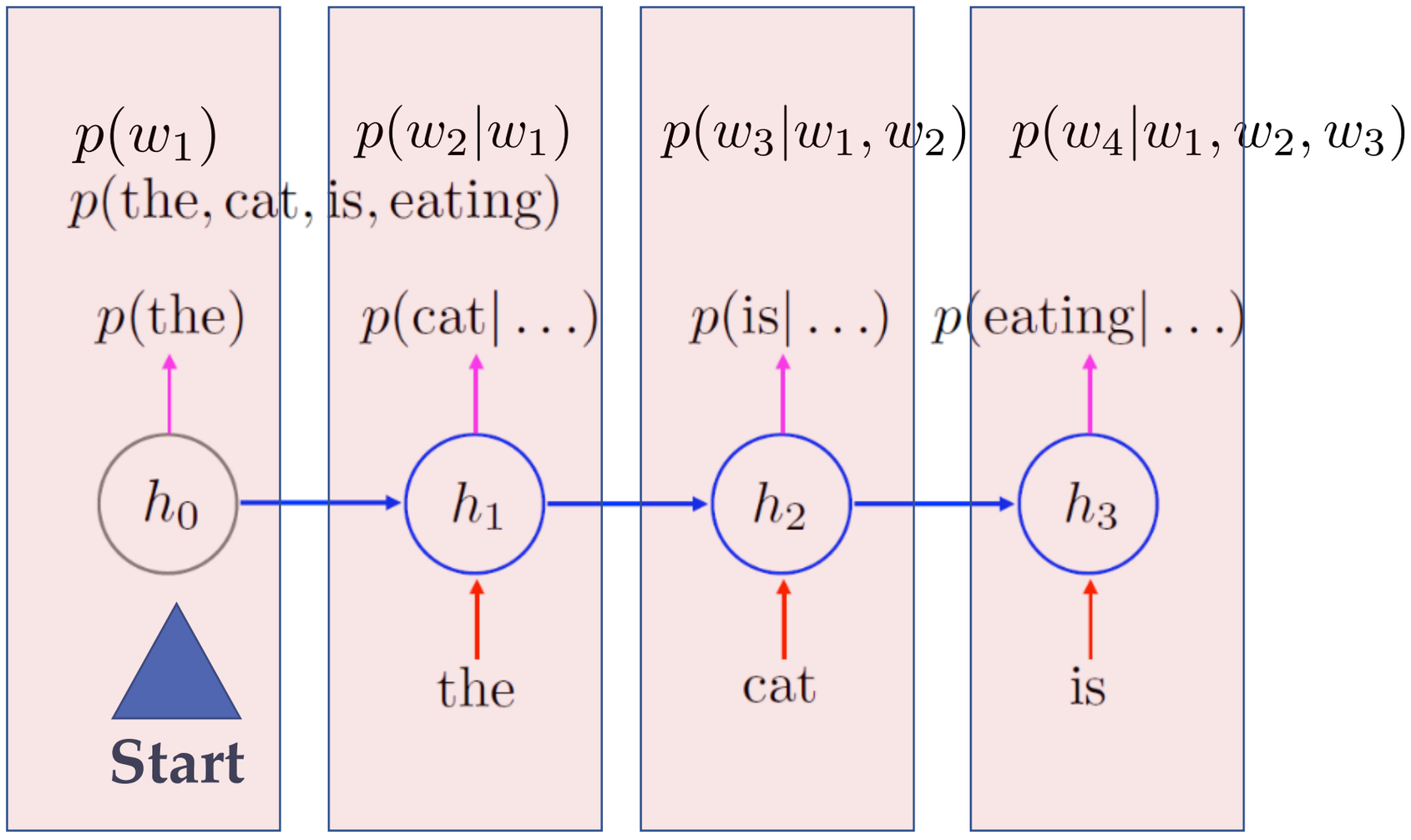
Figure 1: *Simple recurrent neural network.*

Mikolov, Tomáš, et al. "Recurrent neural network based language model." *Eleventh Annual Conference of the International Speech Communication Association*. 2010.

RNN-based Statistics Language Model

- **To compute $P(w_1, w_2, \ldots, w_N)$ by RNN**

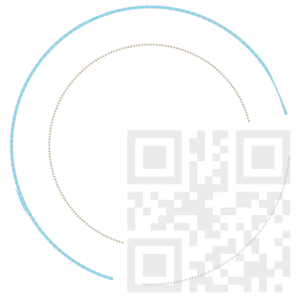
$$p(w_1, \cdots, w_n) = \Pi_{i=1}^{N} p(w_i | w_1, \cdots, w_i - 1)$$
$$= p(w_1) p(w_1 | w_2) \cdots p(w_n | w_1, \cdots, w_{n-1})$$



RNN在Language Model中的角色

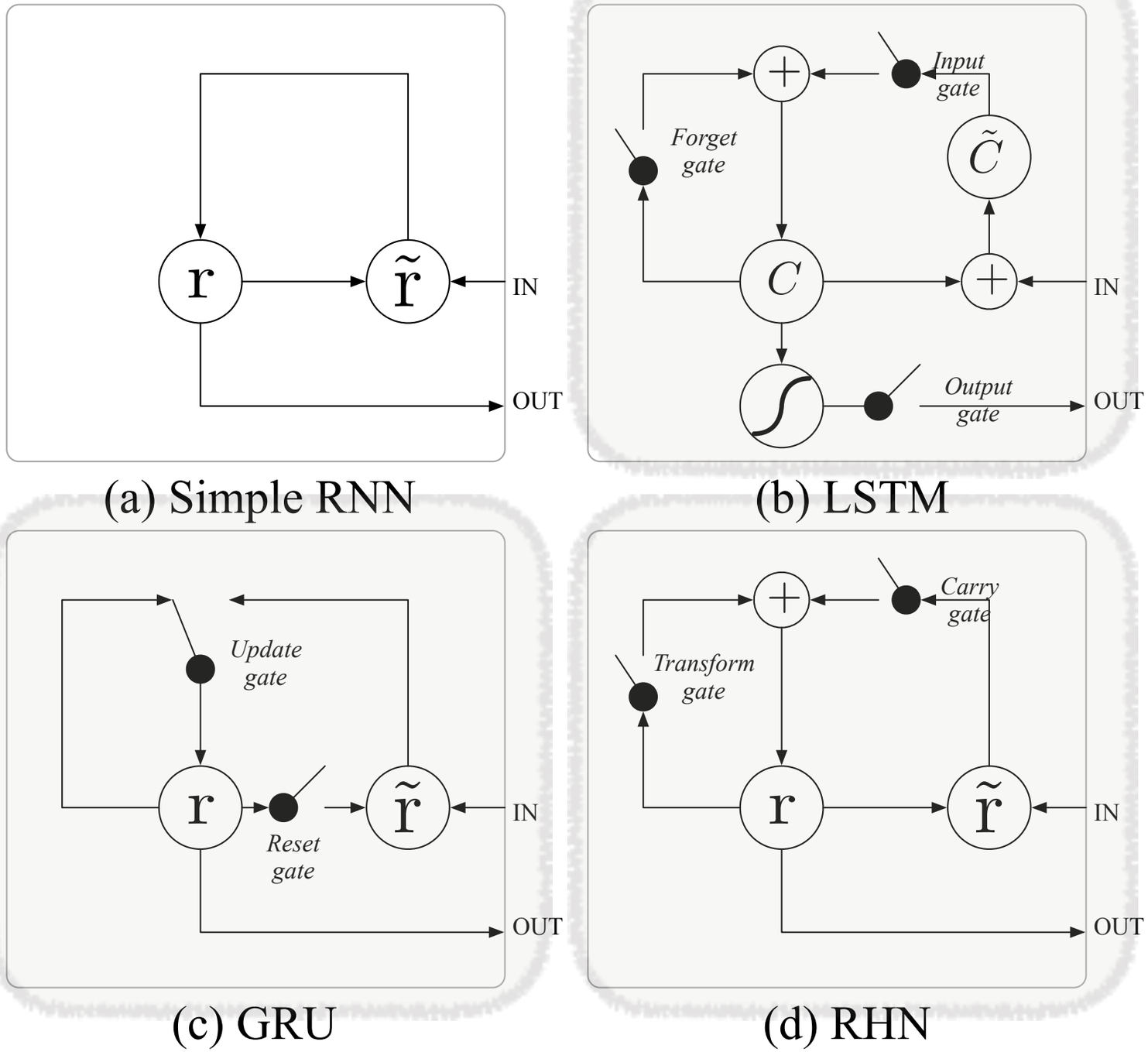
❑ Recurrent Neural Networks (RNNs) 在Sentence/Sequence Predictor扮演了什么角色呢?

- In neural machine translation systems, a recurrent neural network (RNN) is typically viewed as the primary '**generation/生成器**' component.
- The goal of statistical language modeling is **to predict/预测 the next word in textual data given context.**
- Thus we are dealing with **sequential data prediction problem** when constructing language models.
- It is well known that humans can exploit **longer context** with great success. However, it is also often claimed that learning long-term dependencies by stochastic gradient descent can be quite difficult.
  - LSTM据说可以model long-term dependency，但是没理论证明，只有实验分析



有多少种RNN？很多。。。

- Many variants of RNN have been proposed, e.g., GRU, LSTM, GRH, etc.



(a) Simple RNN

(b) LSTM

(c) GRU

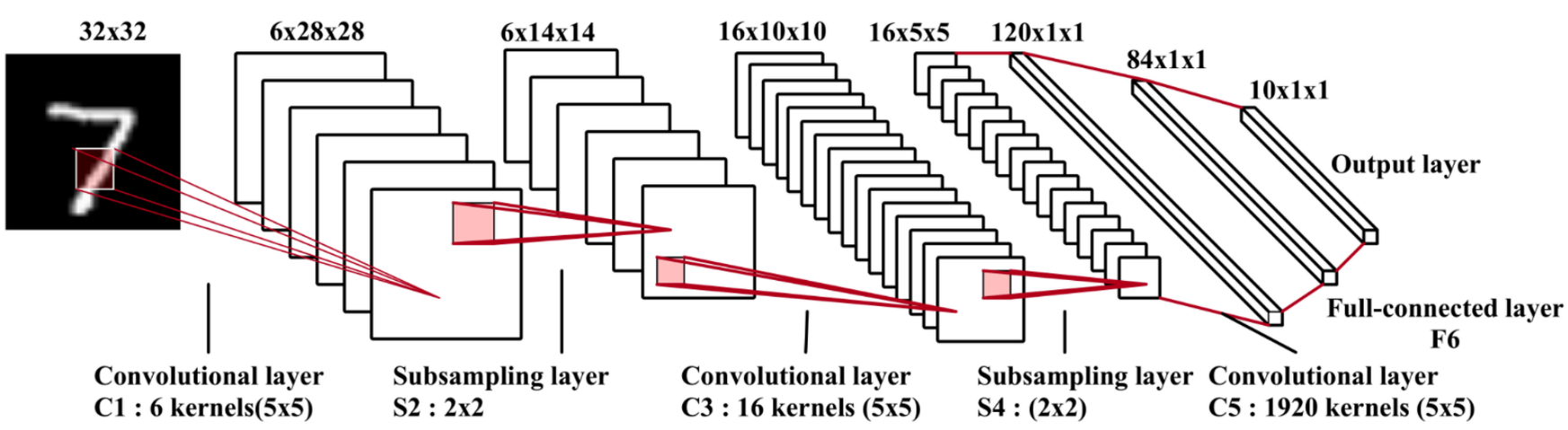
(d) RHN

CNN-based vs. RNN-based Language Model

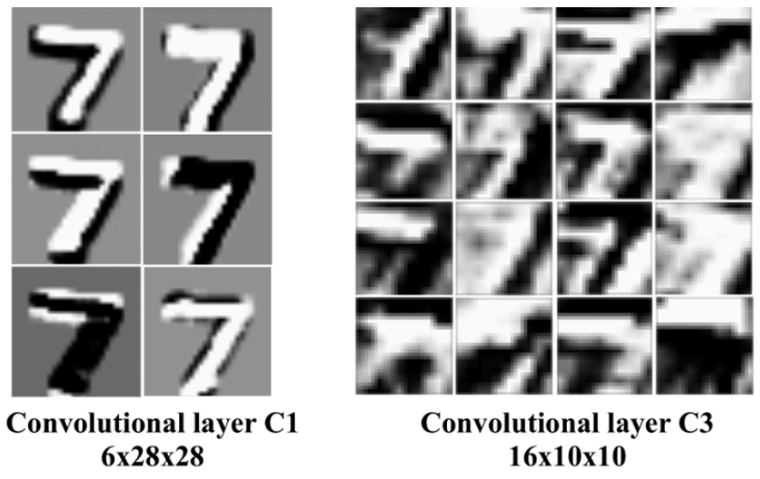
CNN vs. RNN

❑ What is CNN?

- A simple ConvNet is a sequence of layers, and every layer of a ConvNet transforms one volume of activations to another through a differentiable function. We use three main types of layers to build ConvNet architectures:
  - Convolutional Layer
  - Pooling Layer
  - Fully-Connected Layer



(a) LeNet-5 network

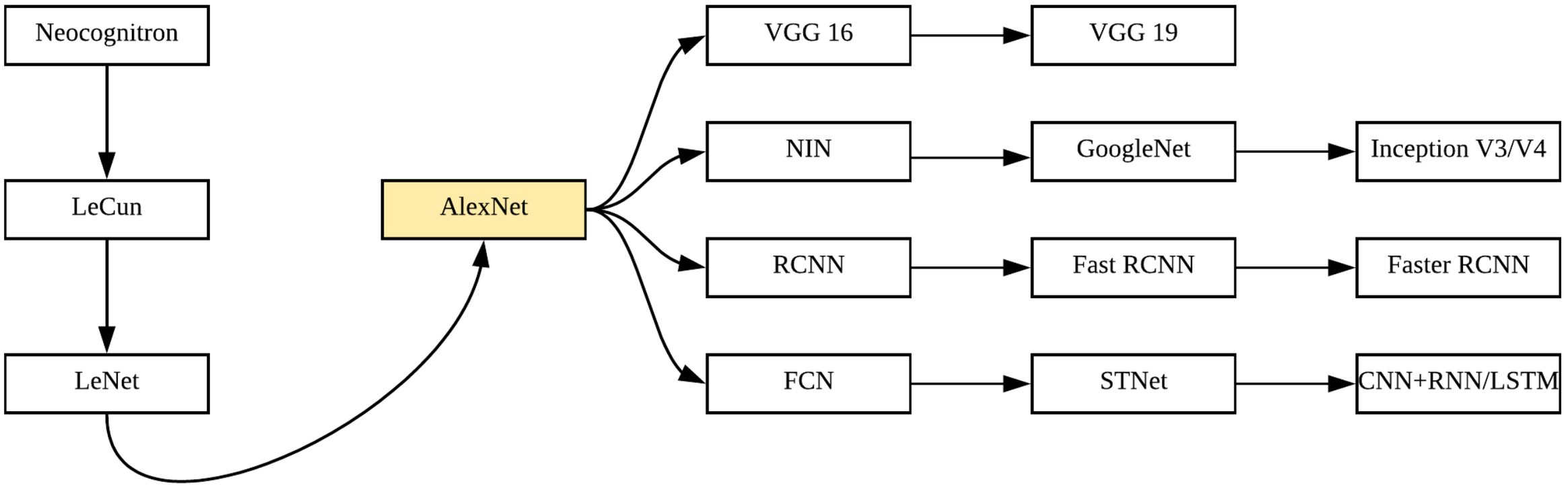


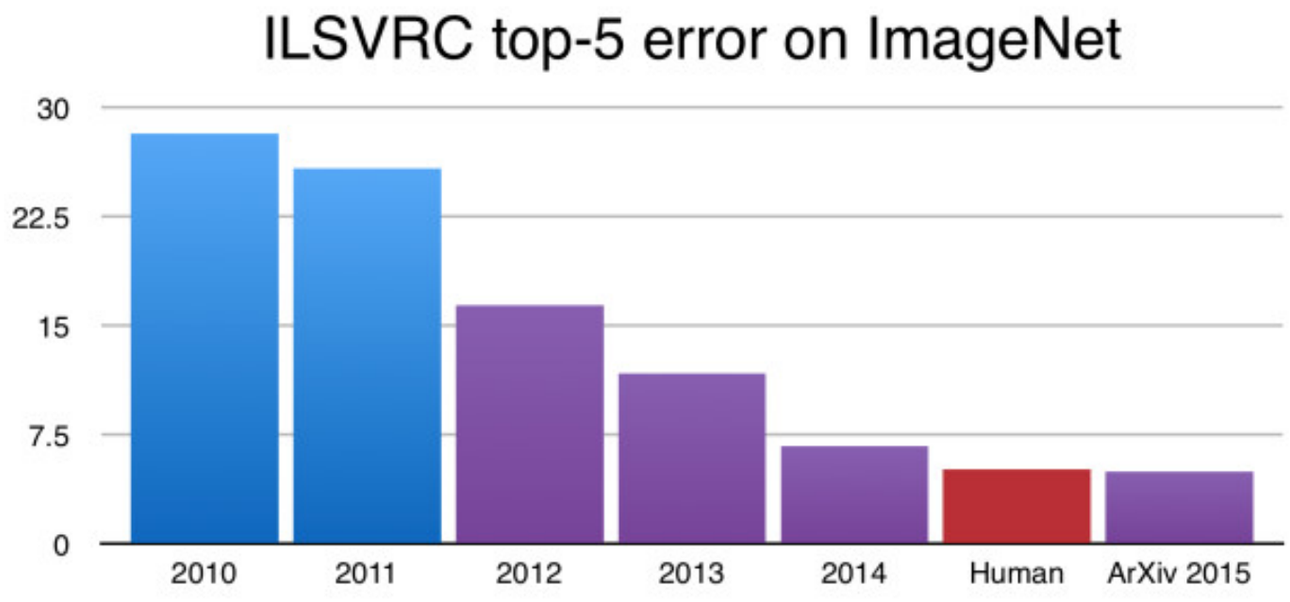
(b) Learned features

了解CNN这一篇就够了：卷积神经网络技术及发展:
http://news.hexun.com/2016-08-08/185382572.html

CNN vs. RNN

❑ Evolution





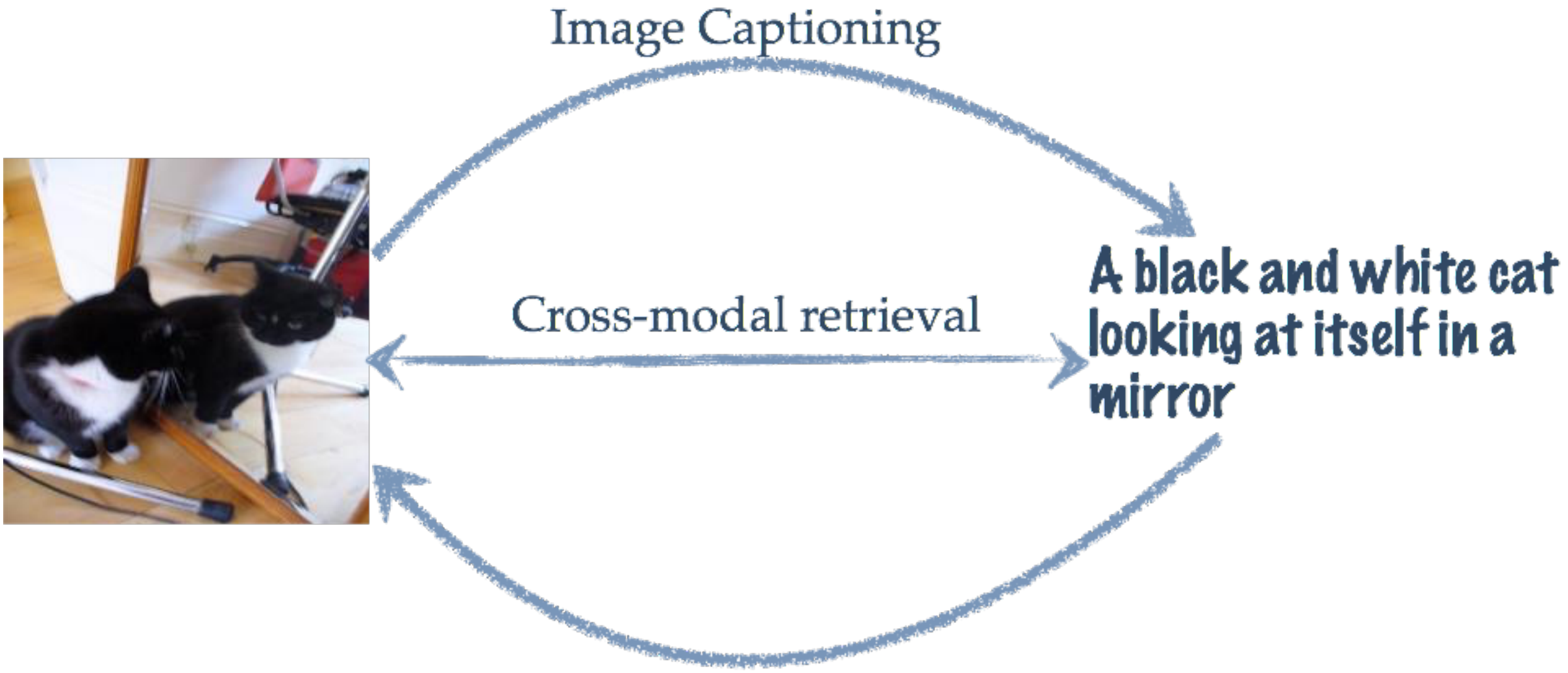
CNN vs. RNN 哪个最好？

❑ Which one is the best?

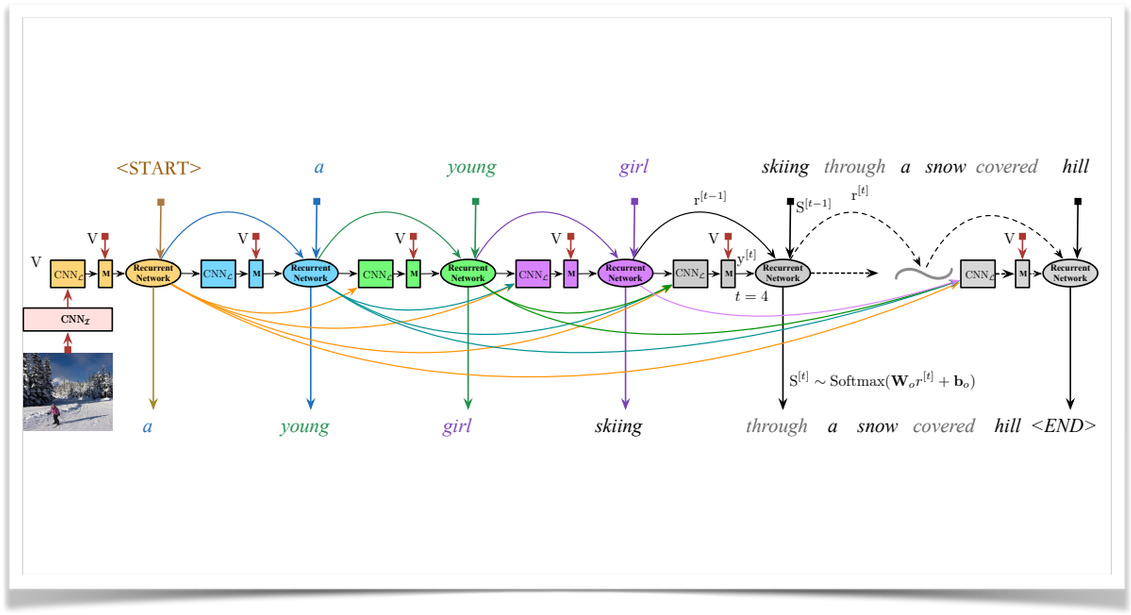
- The best RNNs are the ones that work well for you ！ ！ ！
- We should select the appropriate recurrent network for our
  - LSTM并不是最合适的！！

CV + NLP

From NLP to CV+NLP



An Empirical Study of Language CNN for Image Captioning, ICCV, 2017

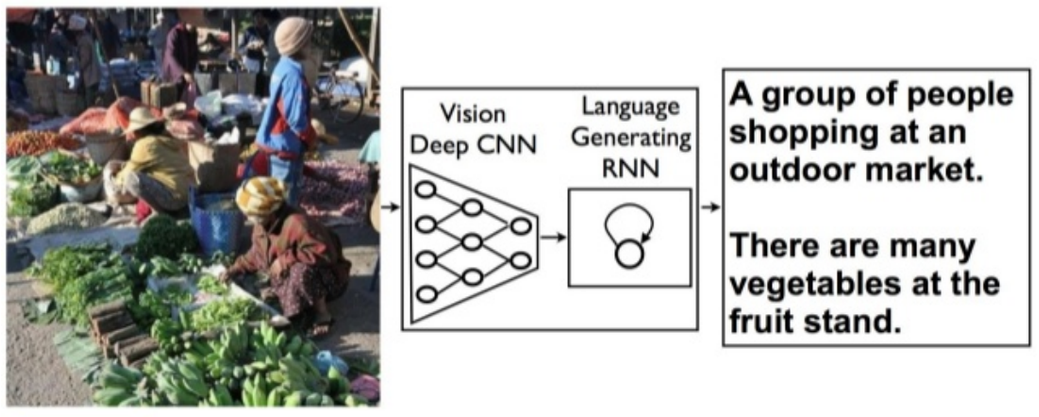


阿里 AI LAB ICCV 2017 录用论文详解：语言卷积神经网络应用于图像标题生成的经验学习: https://www.jiqizhixin.com/articles/2017-10-31-6

Motivation

Describes without errors	Describes with minor errors	Somewhat related to the image
		
A person riding a motorcycle on a dirt road.	Two dogs play in the grass.	A skateboarder does a trick on a ramp.
		
A group of young people playing a game of frisbee.	Two hockey players are fighting over the puck.	A little girl in a pink hat is blowing bubbles.

Captioning: Show & Tell



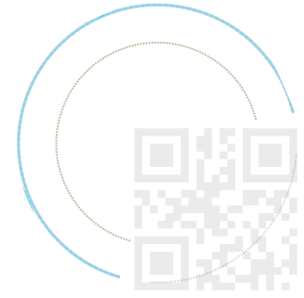
Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and tell: A neural image caption generator." CVPR 2015.

回顾LSTM

❑ Let's take a close look at LSTM.

The full name of LSTM is Long Short-Term Memory (LSTM).

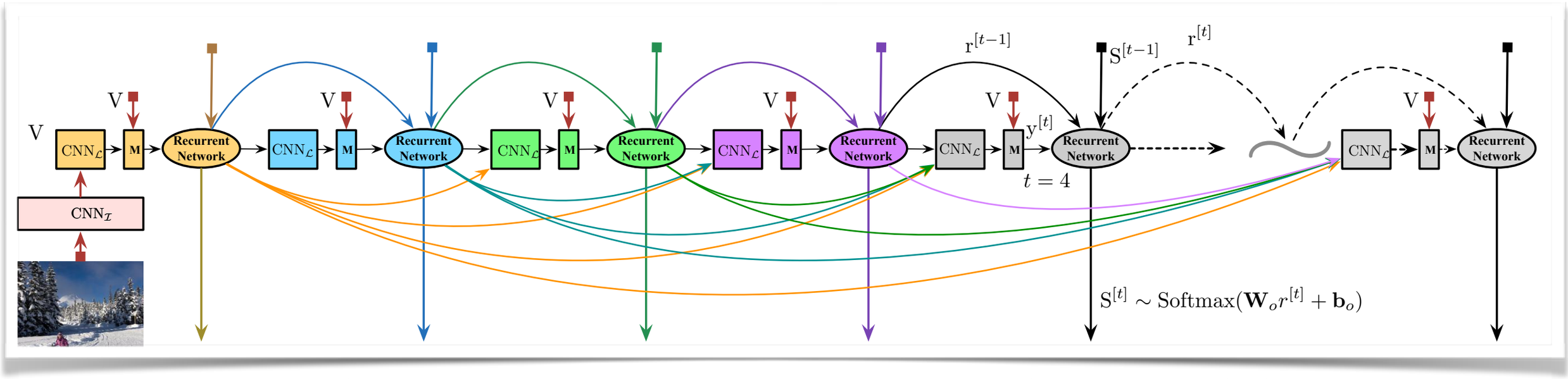
- **Long-Term**
  - The expression long short-term refers to the fact that LSTM is a model has a memory which can last for a long period of time.
  - Long-term dependencies are hard to learn, especially when we have limited number of data.
  - However, we can not trust LSTM completely, as the information will be dropped each time step (input gate, forget gate, and output gate)
- **Short-Term**
  - Each word prediction is highly depended on their adjacent words (n-gram model).
  - Modeling the dynamic temporal behavior.

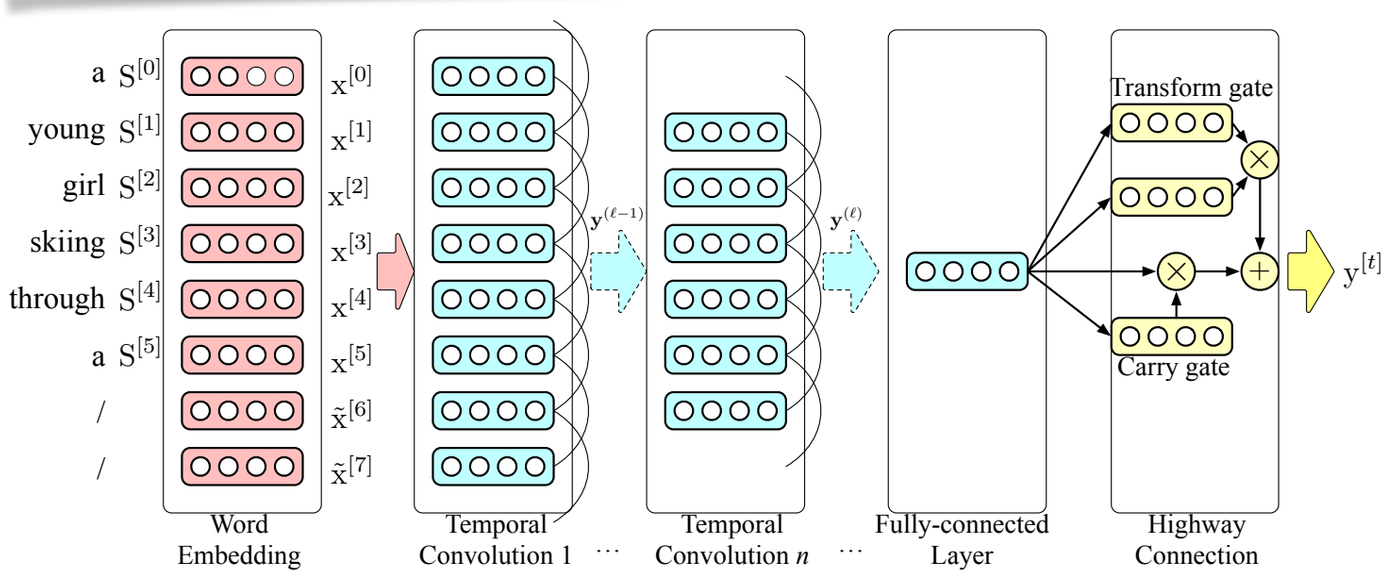


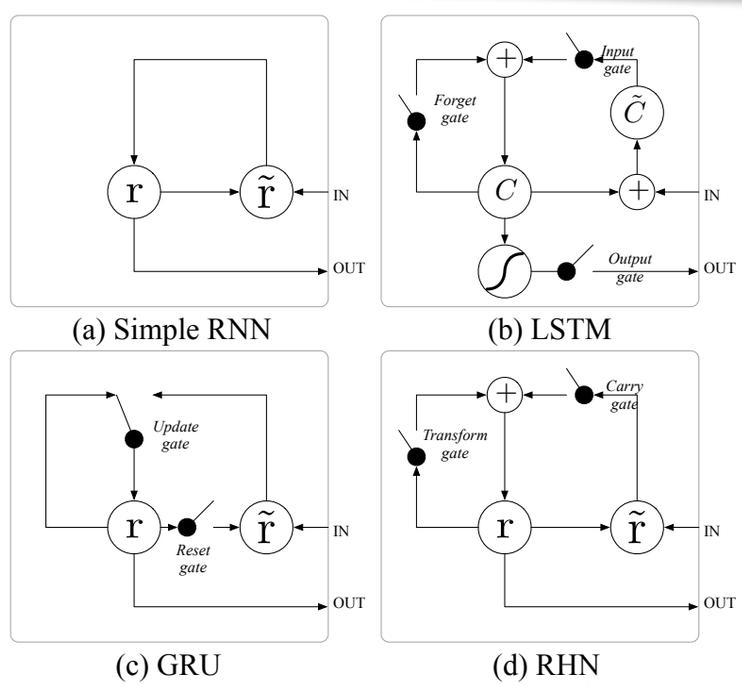
回顾LSTM

- LSTM sometime works, but not always, there is no strict theoretical proof ！
- We need a network which can learn both long-term and short-term behavior.
- Let’s explore a new network !!!

Motivation







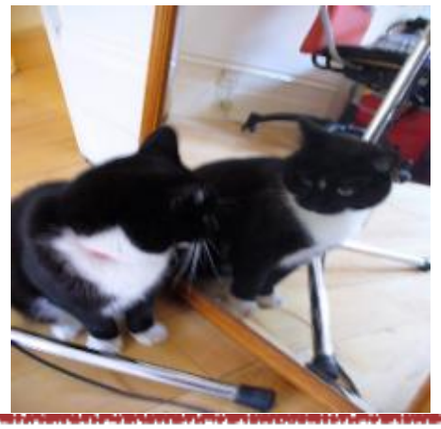
(a) Simple RNN (b) LSTM

(c) GRU (d) RHN

Our model estimates the probability distribution of the next word given previous words and image. It consists of four parts: a CNNI for image feature extraction, a deep CNNL for language modelling, a multimodal layer (M) that connects the CNNI and CNNL, and a Recurrent Network for word prediction. The weights are shared among all time frames.

* Our model is trained with cross-entropy loss

Results



CNNL+RHN **:** a black and white cat looking at itself in a mirror
CNNL+RNN **:** a black and white cat sitting in front of a mirror
GRU **:** a black and white cat standing next to a mirror
LSTM **:** a black and white cat sitting in a bathroom sink
RNN **:** a cat sitting on the floor in a bathroom

- there is a black tuxedo cat looking in the mirror
- two cats sitting on top of a wooden floor
- a cat looking at itself in the mirror next to a tripod
- a cat and a tripod sitting in front of a mirror
- a close up of a cat in a mirror

CNNL+RHN **:** a man standing next to a child on a snow covered slope
CNNL+RNN **:** a man and a woman standing on a snow covered slope
GRU **:** a man and a child standing on a snow covered slope
LSTM **:** a man and a child are standing in the snow
RNN **:** a man and a woman are skiing on the snow

- a woman and child in ski gear next to a lodge
- a man and a child are smiling while standing on skiis
- a young man poses with a little kid in the snow
- an adult and a small child dressed for skiing
- a man and a little girl in skis stand in front of a mountain lodge

CNNL+RHN **:** a man talking on a cell phone while walking down a street
CNNL+RNN **:** a man is talking on a cell phone
GRU **:** a man is talking on a cell phone in the street
LSTM **:** a man is talking on his cell phone
RNN **:** a man standing next to a woman talking on a cell phone

- a man talking on the phone in front of a blue car
- a man on a telephone holds his hand up to his other ear as he walks
- a man standing next to a car with a cellphone
- a man is talking on a cell phone next to a city street
- a man standing on the side of the street with a cell phone up to his

CNNL+RHN **:** a cat looking at a dog in a door
CNNL+RNN **:** a cat is looking at a dog in front of a window
GRU **:** a cat standing next to a door looking out a window
LSTM **:** a dog and a cat are standing in front of a window
RNN **:** a cat sitting on the side of the road

- a dog looking at a cat through a glass window
- a cat is outside looking through in at a dog
- the dog wants to go outside with the cat
- a cat sitting outside of a door next to a dog
- a cat sitting at a sliding glass door

Results

Approach	*Flickr30k*					*MS COCO*					
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
BRNN [19]	57.3	36.9	24.0	15.7	—	62.5	45.0	32.1	23.0	19.5	66.0
Google NIC [46]	—	—	—	—	—	—	—	—	27.7	23.7	85.5
LRCN [6]	58.8	39.1	25.1	16.5	—	66.9	48.9	34.9	24.9	—	—
MSR [7]	—	—	—	—	—	—	—	—	25.7	23.6	—
m-RNN [35]	60.0	41.0	28.0	19.0	—	67.0	49.0	35.0	25.0	—	—
Hard-Attention [51]	66.9	43.9	29.6	19.9	18.5	70.7	49.2	34.4	24.3	23.9	—
Soft-Attention [51]	66.7	43.4	28.8	19.1	18.5	71.8	50.4	35.7	25.0	23.0	—
ATT-FCN [53]	64.7	46.0	32.4	23.0	18.9	70.9	53.7	40.2	30.4	24.3	—
ERD+GoogLeNet [52]	—	—	—	—	—	—	—	—	29.8	24.0	88.6
emb-gLSTM [15]	64.6	44.6	30.5	20.6	17.9	67.0	49.1	35.8	26.4	22.7	81.3
VAE [40]	72.0	53.0	38.0	25.0	—	72.0	52.0	37.0	28.0	24.0	90.0
	State-of-the-art results using model assembling or extra information										
Google NICv2 [47]	—	—	—	—	—	—	—	—	32.1	25.7	99.8
Attributes-CNN+RNN [50]	73.0	55.0	40.0	28.0	—	74.0	56.0	42.0	31.0	26.0	94.0
	Our results										
$CNN_{\mathcal{L}}$+RNN	71.3	53.8	39.6	28.7	22.6	72.2	55.0	40.7	29.5	24.5	95.2
$CNN_{\mathcal{L}}$+RHN	73.8	56.3	41.9	30.7	21.6	72.3	55.3	41.3	30.6	25.2	98.9
$CNN_{\mathcal{L}}$+LSTM	64.5	45.8	32.2	22.4	19.0	72.1	54.6	40.9	30.4	25.1	99.1
$CNN_{\mathcal{L}}$+GRU	71.4	54.0	39.5	28.2	21.1	72.6	55.4	41.1	30.3	24.6	96.1

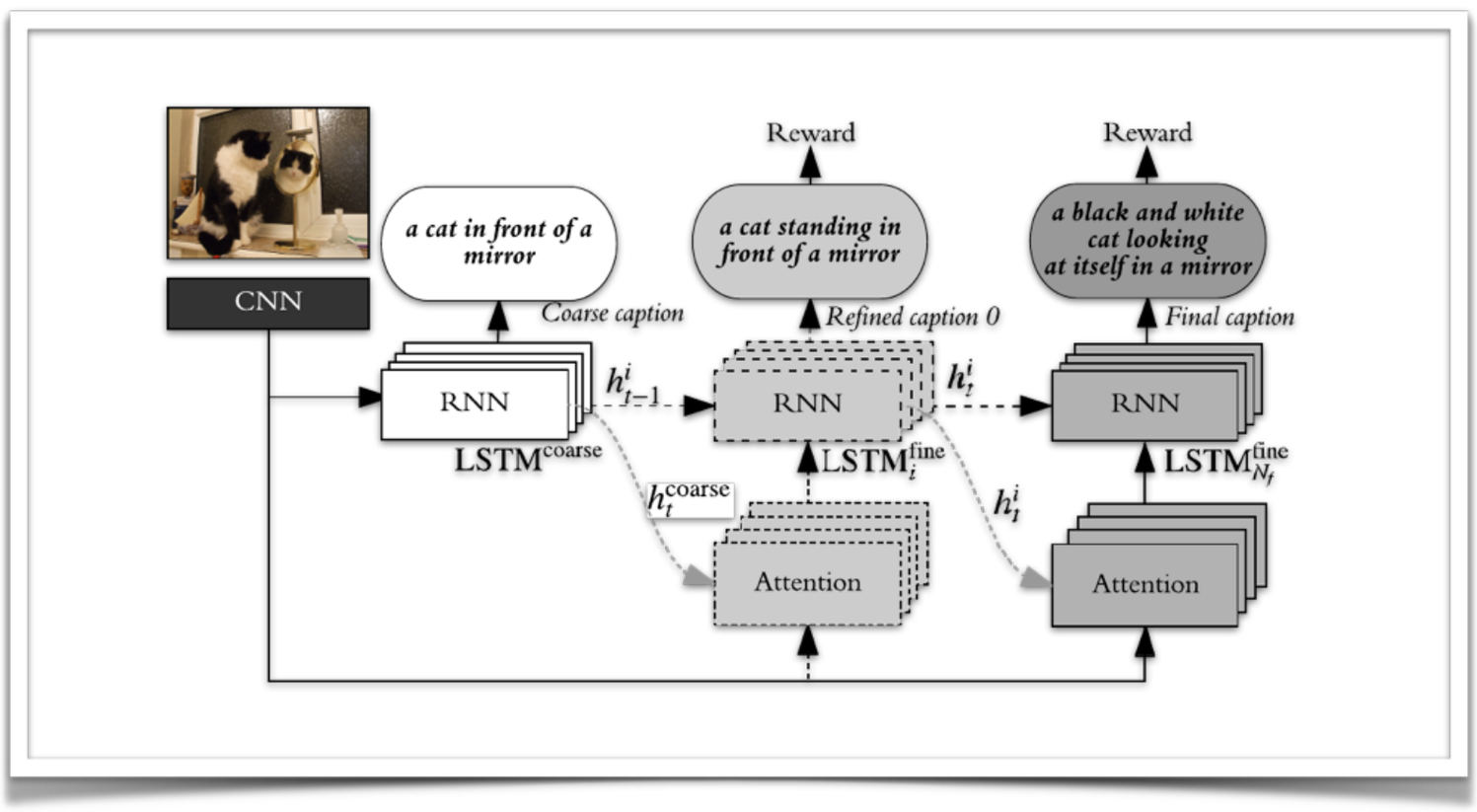
Table 5. Performance in terms of BLEU-n, METEOR, and CIDEr compared with other state-of-the-art methods on the MS COCO and Flickr30k datasets. For those competing methods, we extract their performance from their latest version of papers.

CV + NLP + Reinforcement Learning

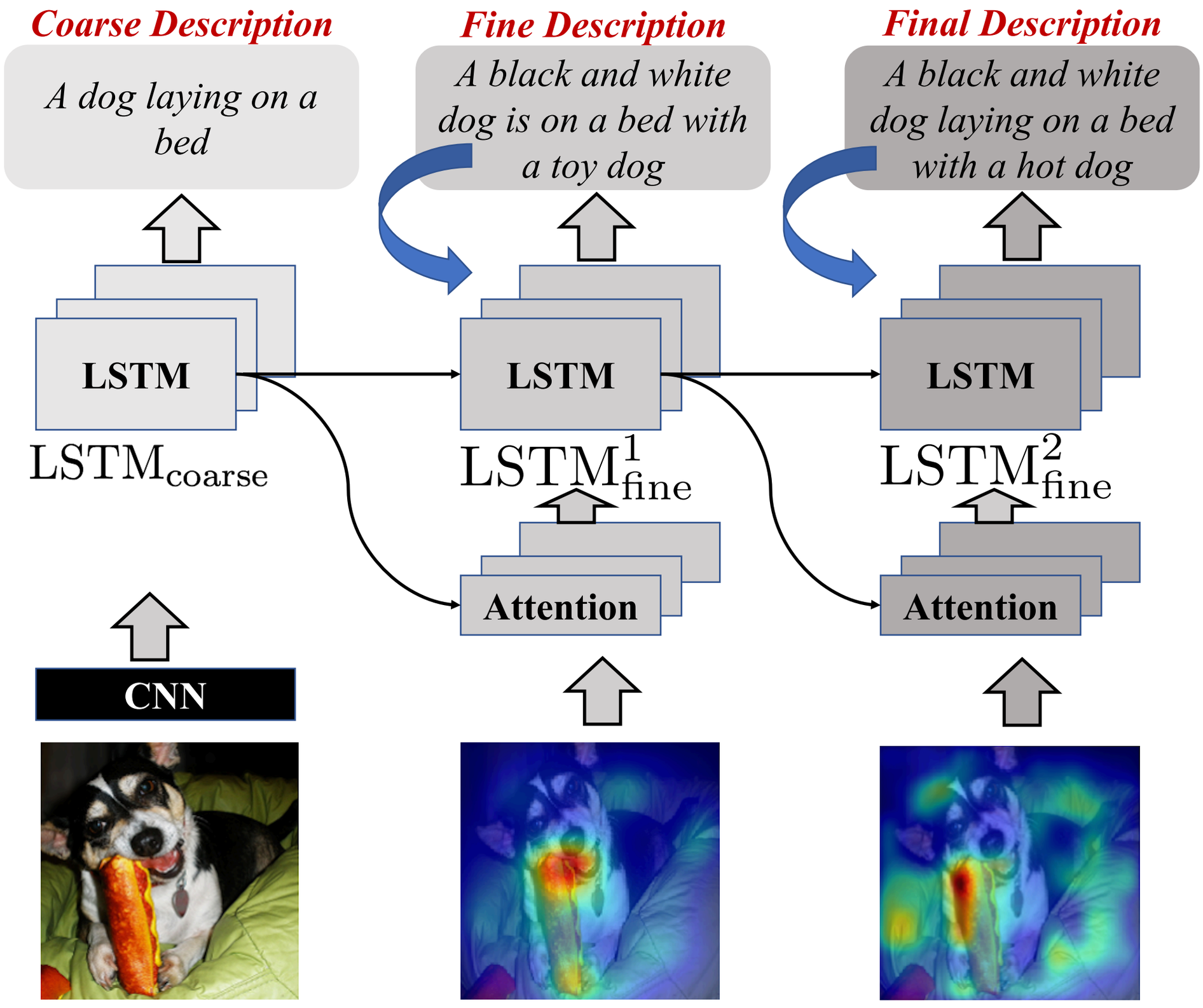
Stack-Captioning:
Coarse-to-Fine Learning for Image Captioning

AAAI oral presentation

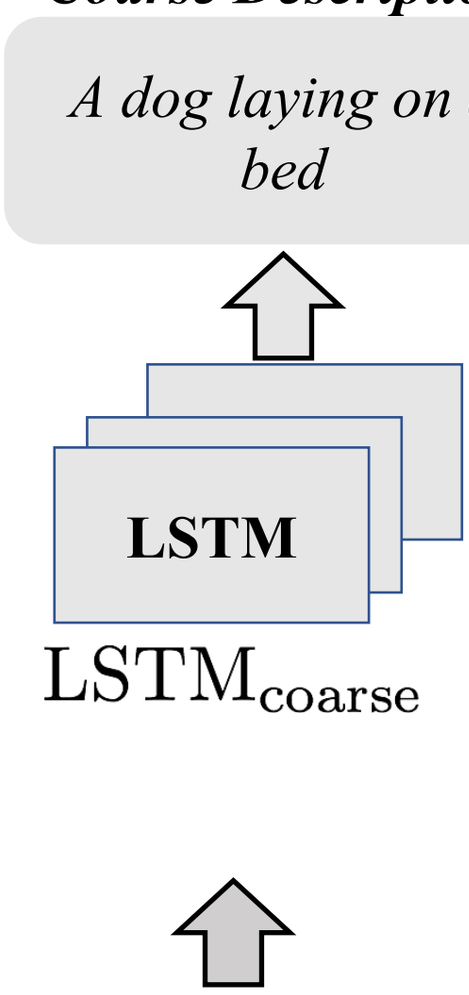
Feb 2-7 2018



Methodology



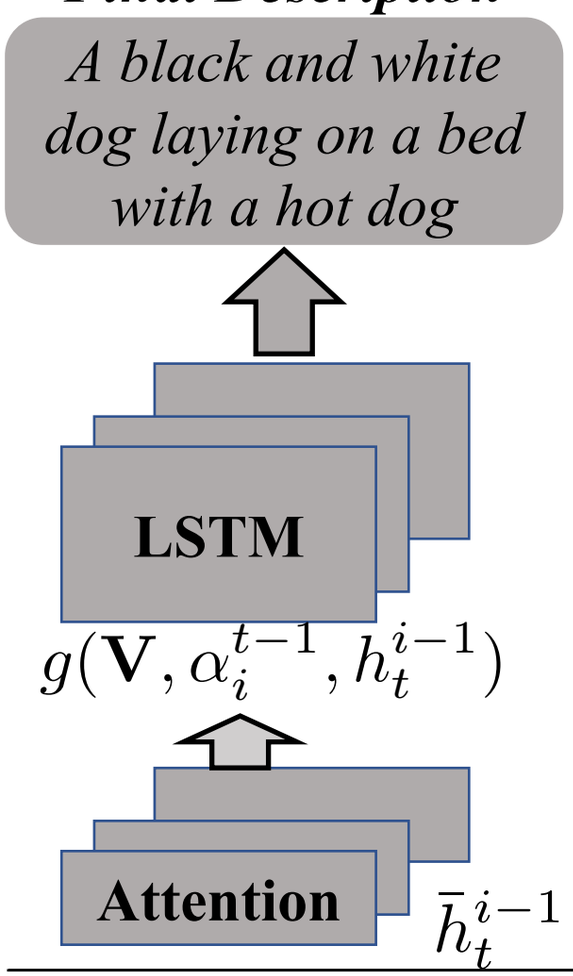
Methodology (Coarse-Decoder)



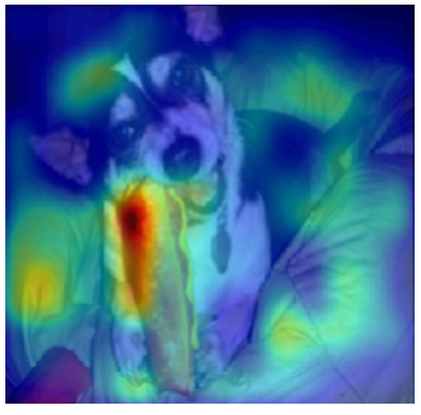
- **Coarse Decoder**
  - We start by decoding in a coarse search space in the first stage (i = 0) ;
  - The operation of the $\mathrm{LSTM}_{\mathrm{coarse}}$ can be described as:
    $$o_t^0, h_t^0 = \mathrm{LSTM}_{\mathrm{coarse}}(h_{t-1}^0, i_t^0, y_{t-1})$$
    $$i_t^0 = [f(\mathbf{V}); h_{t-1}^{N_f}]$$
  - where h_{t-1}^0 and $h_{t-1}^{N_f}$ are the hidden states;
  - $y_{t-1} = \boldsymbol{W}_e\, Y_{t-1}$ is the embedding of previous word Y_{t-1};
  - $\hat{Y}_t^0 \sim \mathrm{Softmax}(\boldsymbol{W}_o^0 o_t^0 + b_o^0)$ is decoded word drawn from the dictionary according to the o_t^0.

- **Encoder**
  - For coarse-stage, we take a mean-pooling over the spatial image features. The global image feature is: $f(\mathbf{V}) = \frac{1}{k \times k} \sum_{i=0}^{k \times k - 1} V_i$, where $\mathbf{V} = \mathrm{CNN}(\mathbf{I})$.

Methodology (Fine-Decoder)



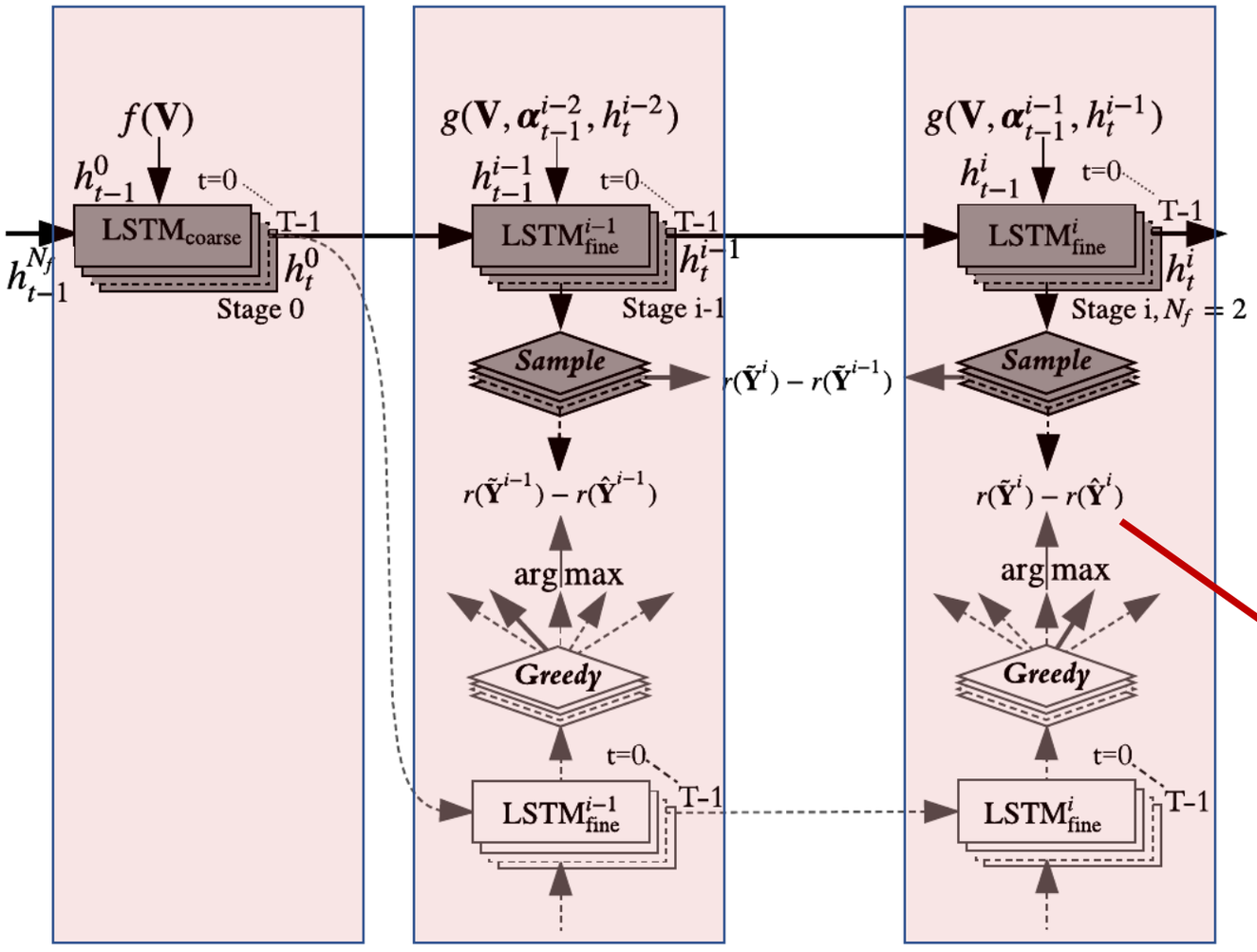
- **Fine Decoder**
  - Each fine decoder consists of an $\text{LSTM}_{\text{fine}}$ network and an attention model;
  - The updating procedure of $\text{LSTM}_{\text{fine}}$ can be written as:
    $$o_t^i, h_t^i = \text{LSTM}_{\text{fine}}^i(h_{t-1}^i, i_t^i, y_{t-1})$$
    $$i_t^i = [\text{g}(\mathbf{V}, \boldsymbol{\alpha}_i^{t-1}, h_t^{i-1}); h_{t-1}^{i-1}]$$
  - where h_{t-1}^{i-1} is are the hidden state of fine decoder;
  - In each fine stage i, our attention model operates on both image features V and attention weights $\boldsymbol{\alpha}_i^{t-1}$ from the preceding stage, the new attended feature is $\text{g}(\mathbf{V}, \boldsymbol{\alpha}_i^{t-1}, h_t^{i-1})$.



- **Stack-Attention**
  - $\text{g}(\mathbf{V}, \boldsymbol{\alpha}_i^{t-1}, h_t^{i-1}) = \sum_{n=0}^{k\times k-1} \alpha_t^{i,n}(\mathbf{W}_{v\alpha}^i V_n + \mathbf{b}_{v\alpha}^i)$;
  - Attention probability: $\boldsymbol{\alpha}_i^i = \text{Softmax}(\mathbf{W}_\alpha^i \boldsymbol{A}_t^i + \mathbf{b}_\alpha^i)$
  - $A_t^{i,n} = \tanh(\mathbf{W}_{va}^i V_n + \mathbf{W}_{ha}^i \bar{h}_t^{i-1})$
  - $\bar{h}_t^{i-1} = h_t^{i-1} + \sum_{n=0}^{k\times k-1} \alpha_t^{i-1,n}(\mathbf{W}_{v\alpha}^{i-1} V_n + \mathbf{b}_{v\alpha}^{i-1})$

Training (1. cross-entropy loss)

- We first incorporate supervised training objectives to the intermediate layers. Each stage of the coarse-to-fine sentence decoder is trained to predict the words repeatedly.



- We first training the network by defining a loss function for each stage i that minimizes the **cross-entropy (XE) loss**:

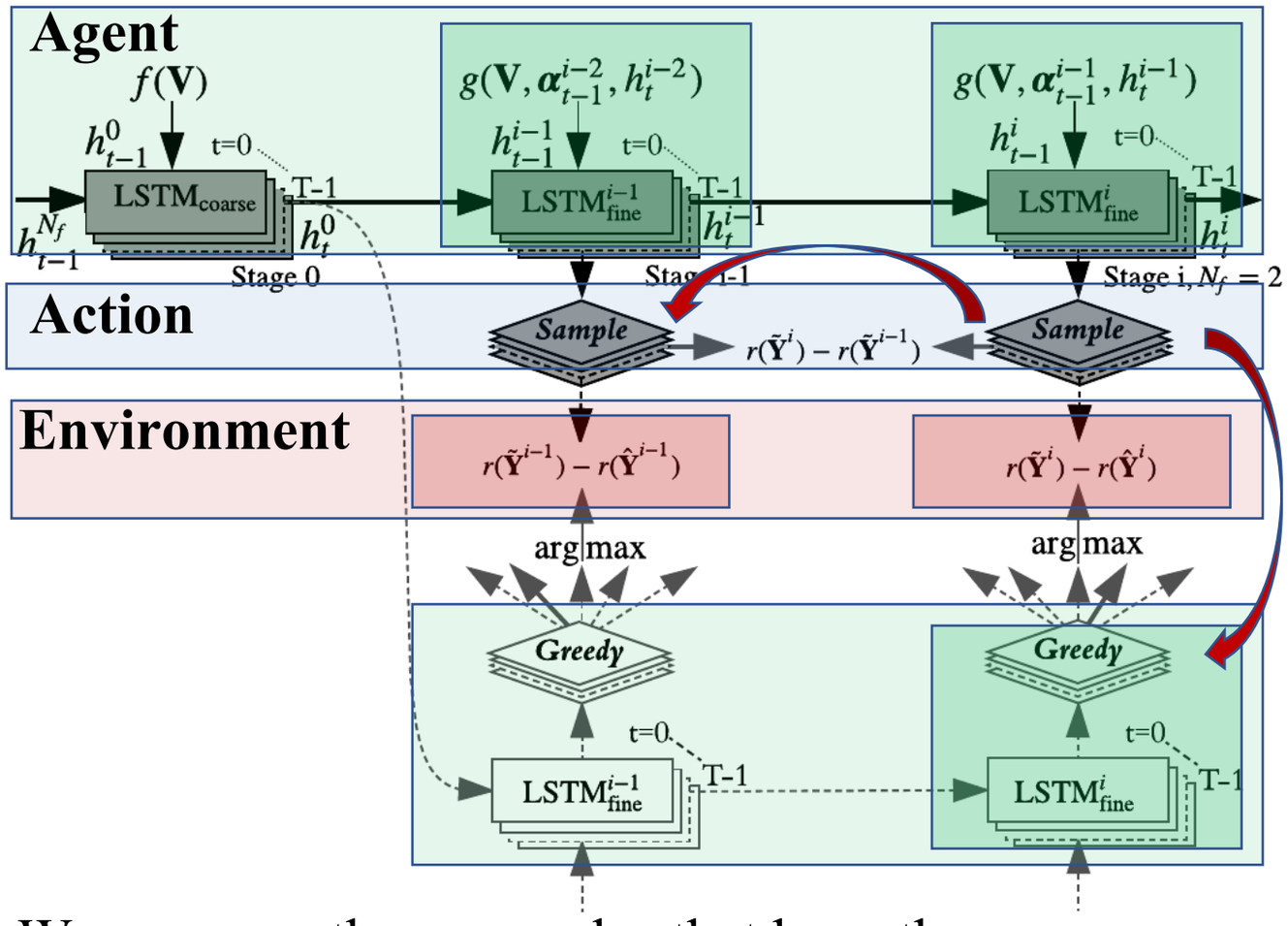
$$\mathcal{L}_{\text{XE}}(\theta) = \sum_{i=0}^{N_f} \mathcal{L}_{\text{XE}}^i(\theta_{0:i})$$

$$= -\sum_{i=0}^{N_f} \sum_{t=0}^{T-1} \log(p_{\theta_{0:i}}(Y_t \mid Y_{0:t-1}, \mathbf{I}))$$

$$\mathcal{L}_{\text{XE}}^i(\theta_{0:i}) = -\sum_{t=0}^{T-1} \log(p_{\theta_{0:i}}(Y_t \mid Y_{0:t-1}, \mathbf{I})),$$

Training (2. REINFORCE-based approach)

- We first incorporate supervised training objectives to the intermediate layers. Each stage of the coarse-to-fine sentence decoder is trained to predict the words repeatedly.



- We suppress those samples that have the worse scores than the greedy decoding results.
- The second term increases the probability of the samples from stage i that outperform the samples from stage i-1, and suppresses the inferior samples.

- Then, the goal of RL-based training is to minimize the **negative expected rewards** (**punishments**) of multi-stages, :

$$\mathcal{L}_{\text{RL}}(\theta) = -\sum_{i=1}^{N_f} \mathbb{E}_{\tilde{\mathbf{Y}}^i \sim p_{\theta_{0:i}}}[r(\tilde{\mathbf{Y}}^i)] \approx -\sum_{i=1}^{N_f} r(\tilde{\mathbf{Y}}^i)$$

$$\nabla_\theta \mathcal{L}_{\text{RL}}(\theta) = \sum_{i=1}^{N_f} \nabla_{\theta_{0:i}} \mathcal{L}_{\text{RL}}(\theta_{0:i})$$

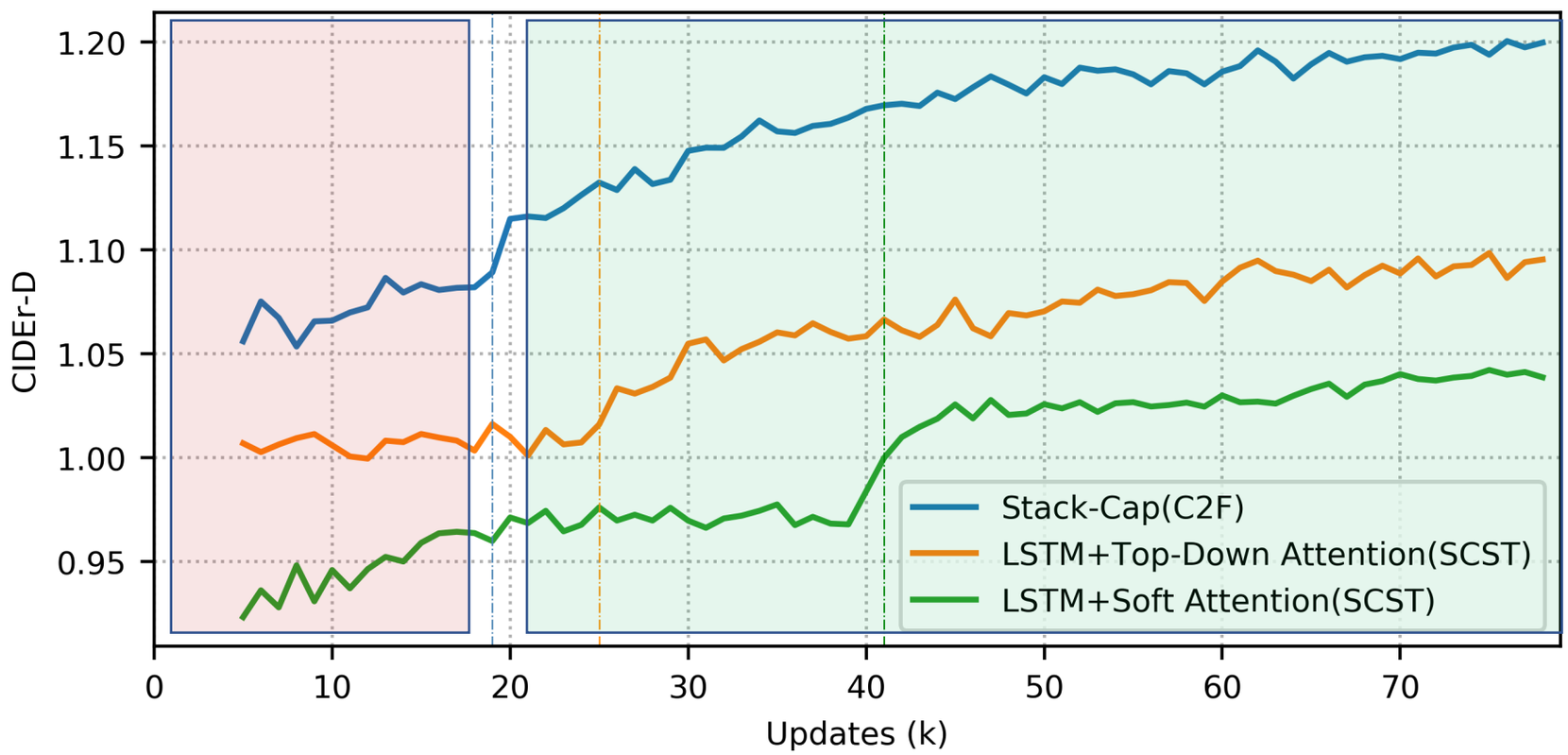
$$\approx -\sum_{i=1}^{N_f} r(\tilde{\mathbf{Y}}^i) \cdot \nabla_{\theta_{0:i}} \log p_{\theta_{0:i}}(\tilde{\mathbf{Y}}^i)$$

$$\nabla_\theta \mathcal{L}_{\text{RL}}(\theta) \approx -\sum_{i=1}^{N_f} \Delta r(\tilde{\mathbf{Y}}^i) \cdot \nabla_{\theta_{0:i}} \log p_{\theta_{0:i}}(\tilde{\mathbf{Y}}^i)$$

$$\Delta r(\tilde{\mathbf{Y}}^i) = \left[r(\tilde{\mathbf{Y}}^i) - r(\hat{\mathbf{Y}}^i) \right] + \left[r(\tilde{\mathbf{Y}}^i) - r(\tilde{\mathbf{Y}}^{i-1}) \right]$$

Experiments

- Dataset
  - We evaluate the proposed approach on MSCOCO dataset. The dataset contrains 123,000 images, where each image has five reference captions.
  - We further test on the MSCOCO test set consisting of 40,775 images, and then conduct the online comparison against the state-of-the-art via the online MSCOCO evaluation server.
- Implementation Details
  - We first train our model under the cross-entropy cost using Adam optimizer with an initial learning rate of 4 × 10−4 and a momentum parameter of 0.9.
  - After that, we run the proposed RL-based approach on the just trained model to be optimized for the CIDEr metric.

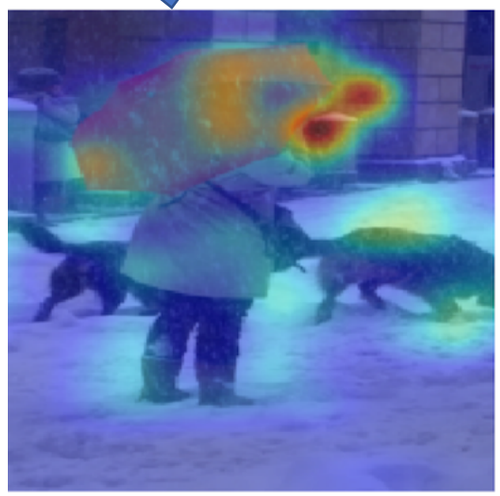


Experiments (Qualitative Analysis)

- *a woman with a red umbrella is walking two dogs in the snow*
- *a person with a white umbrella with two dogs*
- *a woman is walking her dogs on the city sidewalks through the newly fallen snow*
- *a person with an umbrella and two dogs walking in the snow*
- *a woman is walking two dogs in the snow*

a woman walking in the snow

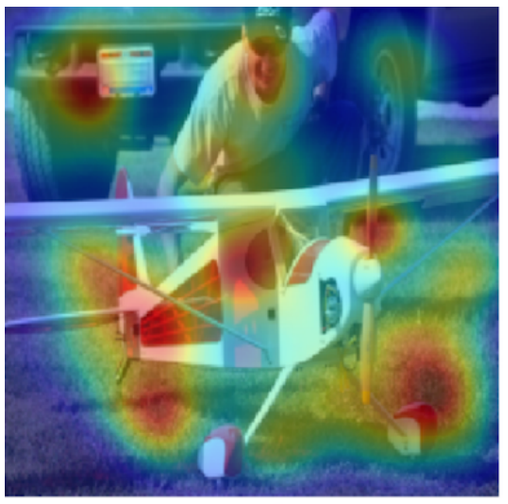
a woman walking a dog in the snow



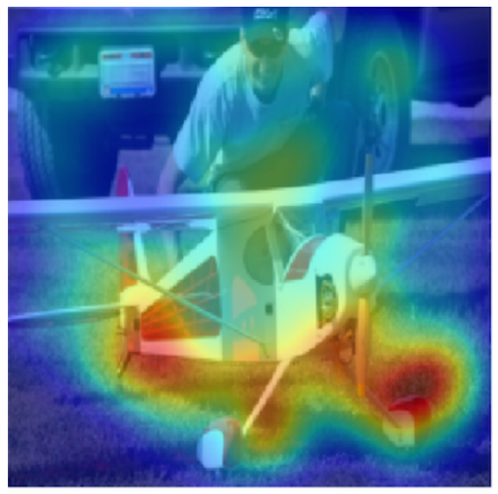
a woman walking a dog in the snow with an umbrella

- *man bending down to check out a model plain that is parked in the grass*
- *man behind a radio operated model airplane on the ground*
- *the man is crouched down with a small airplane model*
- *a small red and white toy plane in the grass*
- *a man smiles while kneeling beside a miniature airplane*

a man is on front grass with a toy



a man is on front grass with a toy truck



a man sitting in the grass with a toy plane

Quantitative Results (offline+online)

Approach	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Google NIC (Vinyals et al., 2015)	—	—	—	27.7	—	23.7	85.5	—
Hard-Attention Xu et al. (2015)	70.7	49.2	34.4	24.3	23.9	—	—	—
Soft-Attention (Xu et al., 2015)	71.8	50.4	35.7	25.0	23.0	—	—	—
VAE (Pu et al., 2016)	72.0	52.0	37.0	28.0	24.0	—	90.0	—
Google NICv2 (Vinyals et al., 2016)	—	—	—	32.1	25.7	—	99.8	—
Attributes-CNN+RNN (Wu et al., 2016)	74.0	56.0	42.0	31.0	26.0	—	94.0	—
$CNN_{\mathcal{L}}$+RHN (Gu et al., 2017b)	72.3	55.3	41.3	30.6	25.2	—	98.9	18.3
PG-SPIDEr-TAG (Liu et al., 2016)	75.4	59.1	44.5	33.2	25.7	55.0	101.3	—
Adaptive (Lu et al., 2017)	74.2	58.0	43.9	33.2	26.6	—	108.5	—
SCST:Att2in (Rennie et al., 2017)	—	—	—	33.3	26.3	55.3	111.4	—
SCST:Att2in (Ens. 4) (Rennie et al., 2017)	—	—	—	34.8	26.9	56.3	115.2	—
Stack-Cap (C2F)	**78.6**	**62.5**	**47.9**	**36.1**	**27.4**	**56.9**	**120.4**	**20.9**

Table 3: Comparisons of the image captioning performance of the existing methods on MSCOCO Karpathy test split. Our Stack-Cap (C2F) model with the coarse-to-fine learning achieves significant gains across all metrics.

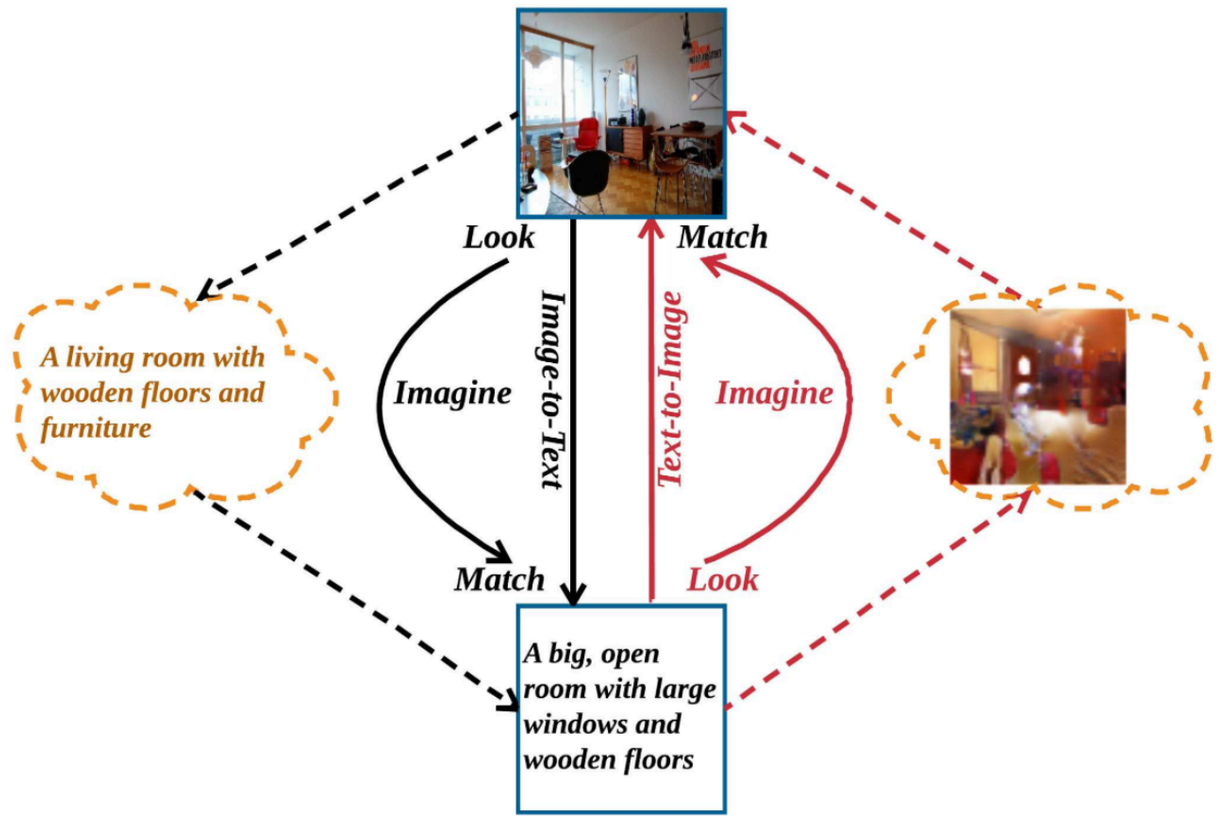
	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr	
Approach	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Google NIC	71.3	89.5	54.2	80.2	40.7	69.4	30.9	58.7	25.4	34.6	53.0	68.2	94.3	94.6
Hard-Attention	70.5	88.1	52.8	77.9	38.3	65.8	27.7	53.7	24.1	32.2	51.6	65.4	86.5	89.3
PG-SPIDEr-TAG	75.1	91.6	59.1	84.2	44.5	73.8	33.1	62.4	25.5	33.9	55.1	69.4	104.2	107.1
Adaptive	74.8	92.0	58.4	84.5	44.4	74.4	33.6	63.7	26.4	35.9	55.0	70.5	104.2	105.9
SCST:Att2in (Ens. 4)	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
Ours: Stack-Cap (C2F)	77.8	93.2	61.6	**86.1**	46.8	**76.0**	34.9	**64.6**	**27.0**	**35.6**	56.2	70.6	**114.8**	**118.3**

Table 4: Leaderboard of the published image captioning models (as of 10/09/2017) on the online MSCOCO test server. Our single Stack-Cap model trained with the coarse-to-fine learning yields comparable performance with the state-of-the-art approaches on all reported metrics.

CV + NLP + Cross-Modal Retrieval

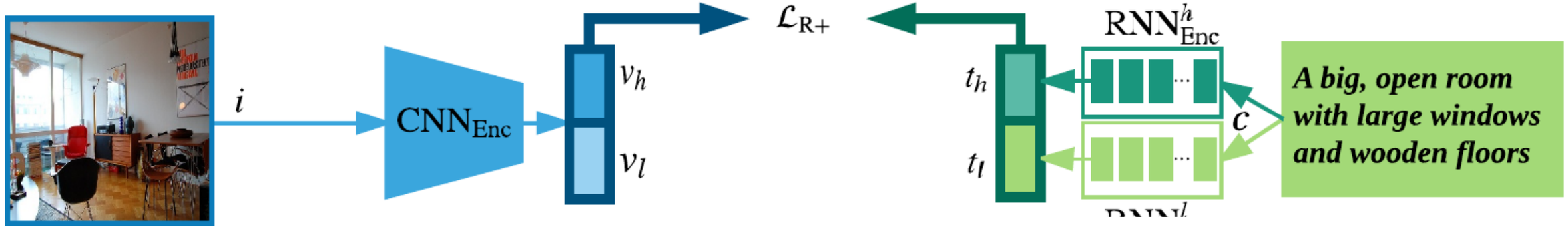
Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval with Generative Models

CVPR 2018



-【论文】所见所想所真，对抗学习GAN提升跨模态检索效果！阿里巴巴AI Labs等团队最新工作：http://www.zhuanzhi.ai/knowledge/9fc760c56299ce049557cf1abaeaf72d

Cross-modal Feature Embedding



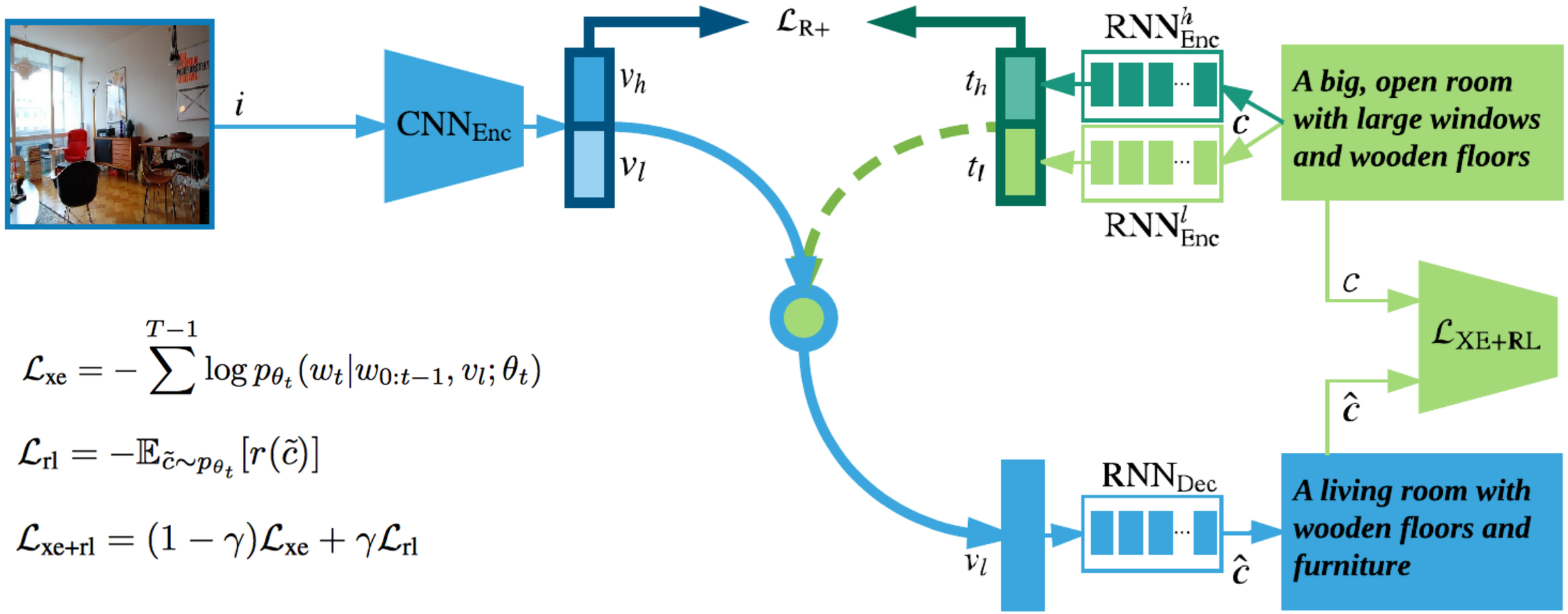
$$
\begin{aligned}
v_k &= P_v^k(\text{CNN}_{\text{Enc}}(i;\theta_i)) \\
t_k &= P_t^k(\text{RNN}^k_{\text{Enc}}(c;\theta_c^k))
\end{aligned}, \quad k \in \{h, l\}
$$

$$
\mathcal{L}_{\text{R}} = \sum_{t'} [\alpha - s(t,v) + s(t',v)]_+ + \sum_{v'} [\alpha - s(t,v) + s(t,v')]_+
$$

$$
\mathcal{L}_{\text{R+}} = \sum_{t'} [\alpha - s^*(t_{h,l}, v_{h,l}) + s^*(t'_{h,l}, v_{h,l})]_+ + \sum_{v'} [\alpha - s^*(t_{h,l}, v_{h,l}) + s^*(t_{h,l}, v'_{h,l})]_+
$$

$$
s^*(t_{h,l}, v_{h,j}) = \lambda s(t_h, v_h) + (1-\lambda) s(t_l, v_l)
$$

Image-to-Text Retrieval

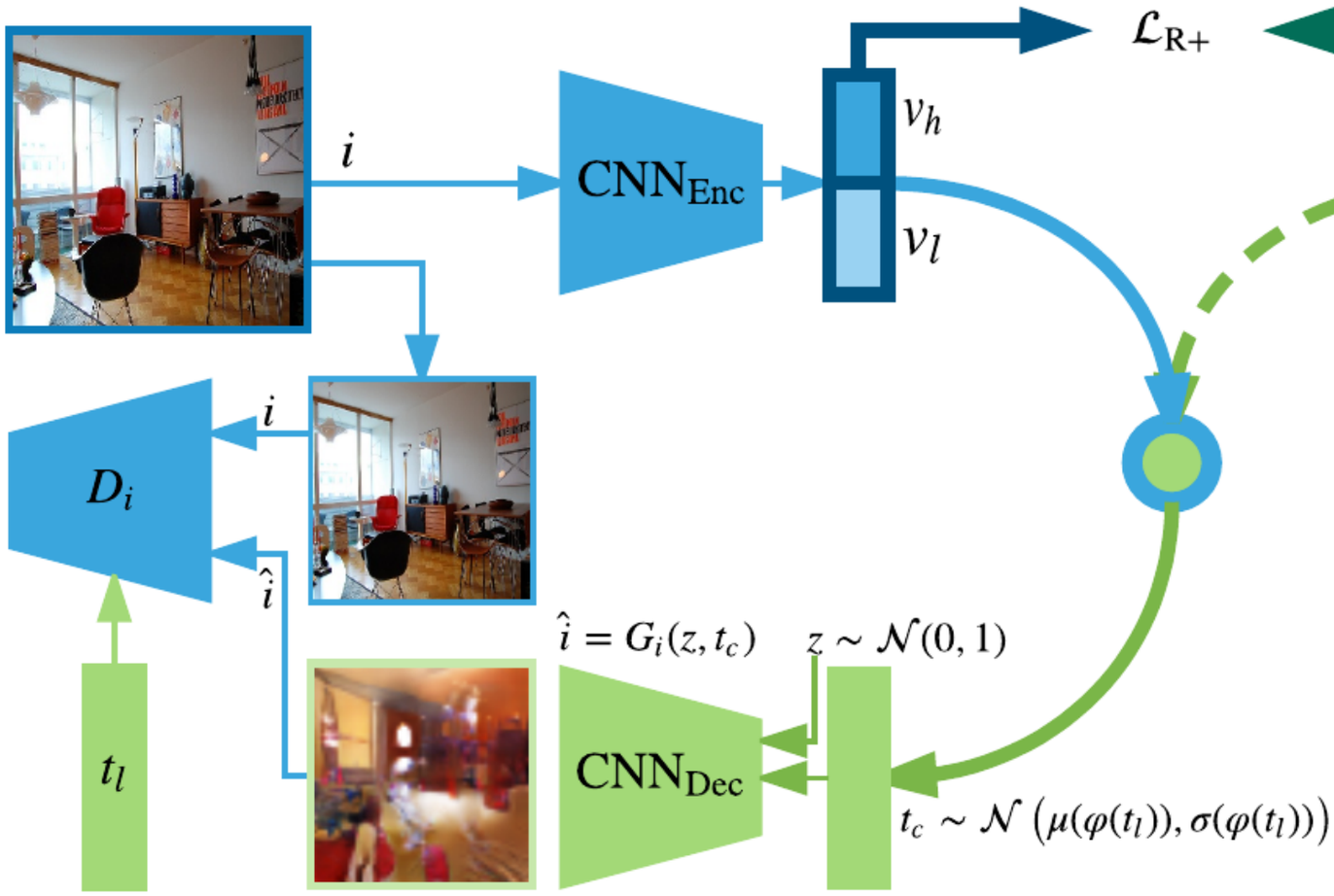


$$\mathcal{L}_{\text{xe}} = -\sum^{T-1} \log p_{\theta_t}(w_t | w_{0:t-1}, v_l; \theta_t)$$

$$\mathcal{L}_{\text{rl}} = -\mathbb{E}_{\tilde{c} \sim p_{\theta_t}}[r(\tilde{c})]$$

$$\mathcal{L}_{\text{xe+rl}} = (1-\gamma)\mathcal{L}_{\text{xe}} + \gamma \mathcal{L}_{\text{rl}}$$

Text-to-Image Retrieval



$$\min_{G_i} \max_{D_i} V(D_i, G_i) = \mathcal{L}_{D_i} + \mathcal{L}_{G_i}.$$

$$\begin{aligned}\mathcal{L}_{D_i} =& \mathbb{E}_{i\sim p_{\text{data}}}[\log D_i(i, t_l)] + \beta_f \mathbb{E}_{\hat{i}\sim p_G}[\log(1 - D_i(\hat{i}, t_l))] + \\ & \beta_w \mathbb{E}_{i\sim p_{\text{data}}}[\log(1 - D_i(i, t'_l))] \qquad (9)\\ \mathcal{L}_{G_i} =& \mathbb{E}_{\hat{i}\sim p_G}[\log(1 - D_i(\hat{i}, t_l))] \qquad (10)\end{aligned}$$

$$\begin{aligned}\mathcal{L}_{D_i} =& \mathbb{E}_{i\sim p_{\text{data}}}[\log D_i(i, t_l)] + \beta_f \mathbb{E}_{\hat{i}\sim p_G}[\log(1 - D_i(\hat{i}, t_l))] + \\ & \beta_w \mathbb{E}_{i\sim p_{\text{data}}}[\log(1 - D_i(i, t'_l))] \qquad (11)\\ \mathcal{L}_{G_i} =& \mathbb{E}_{\hat{i}\sim p_G}[\log(1 - D_i(\hat{i}, t_l))] + \\ & \beta_s \mathcal{D}_{\text{KL}}(\mathcal{N}(\mu(\varphi(t_l)), \sigma(\varphi(t_l))) \parallel \mathcal{N}(0, 1)) \qquad (12)\end{aligned}$$

Qualitative Results (Image-to-Text)

Retrieved Captions:
Ours: 1. a young man doing a skateboard trick while others watch
2. man doing a skate trick during a competition event with a audience
3. skateboarders are doing tricks as a crowd watches
4. a group of people doing skateboarding tricks on a car
5. a boy riding on his skateboard at a skate patk whike other guys watch
VSE0 : Young skateboarder displaying skills on sidewalk near field
VSE++: Two young men are outside skateboarding together
Generated caption: A man doing a trick on a skateboard on a skate park

GT
- **Some skateboarders doing tricks and people watching them**
- **Boys skateboarding over a rusty car onto platforms, with an audience watching**
- **A group of kids doing stunts at a skating event**
- **Skateboarders are doing tricks as a crowd watches**
- **A group of teens performing stunts at a skateboard park**

Retrieved Captions:
Ours: 1. a beautiful young lady holding a tennis racquet on top of a court
2. a lady dressed in pink playing a game of tennis
3. a woman holding a racquet on top of a tennis court
4. a woman is standing while holding a tennis racket
5. a woman in pink dress playing a game of tennis
VSE0 : A man playing tennis and holding back his racket to hit the ball
VSE++: A woman is standing while holding a tennis racket
Generated caption: A woman holding a tennis racket on a court

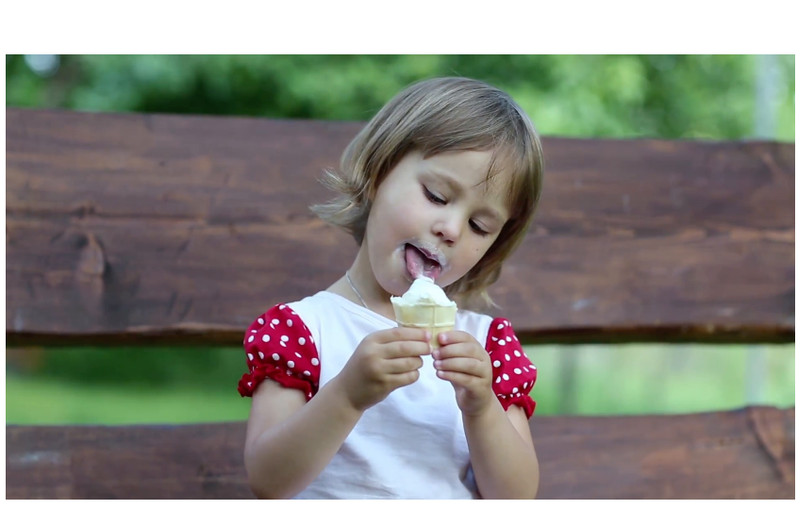
GT
- **A woman in a short pink skirt holding a tennis racquet**
- **A woman is standing while holding a tennis racket**
- **A female tennis player is positioning herself for her next move**
- **A young girl is holding a tennis racket**
- **A woman in a skirt is holding a tennis racket**

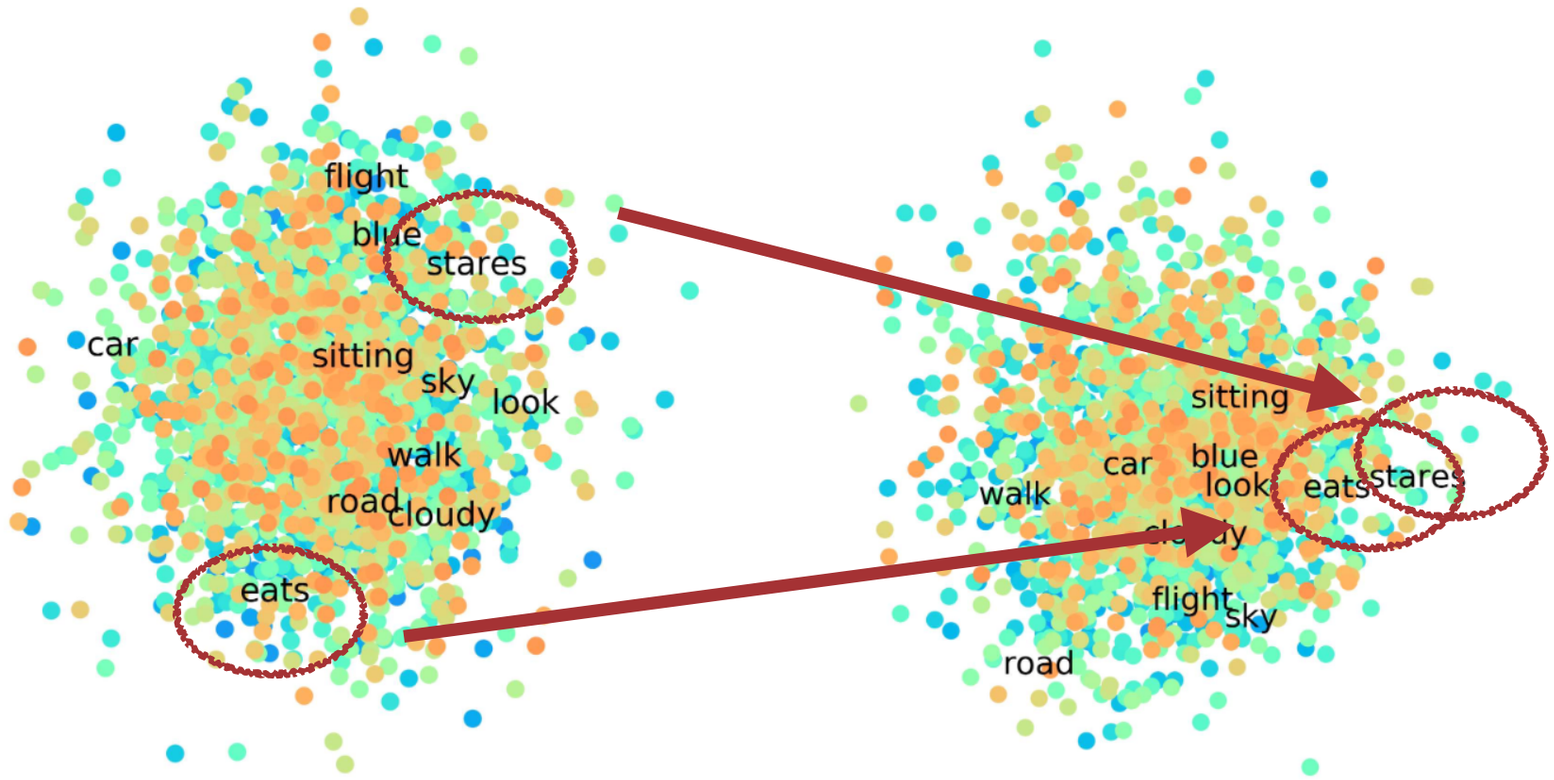
Figure 3: Visual results of image-to-text retrieval, where the top-5 retrieved captions and the generated caption are shown in red color.

Qualitative Results (Text-to-Image)

Figure 4: Visual results of text-to-image retrieval. 2nd row: retrieved images. 3rd row: image samples generated by our conditional GAN.

Qualitative Results





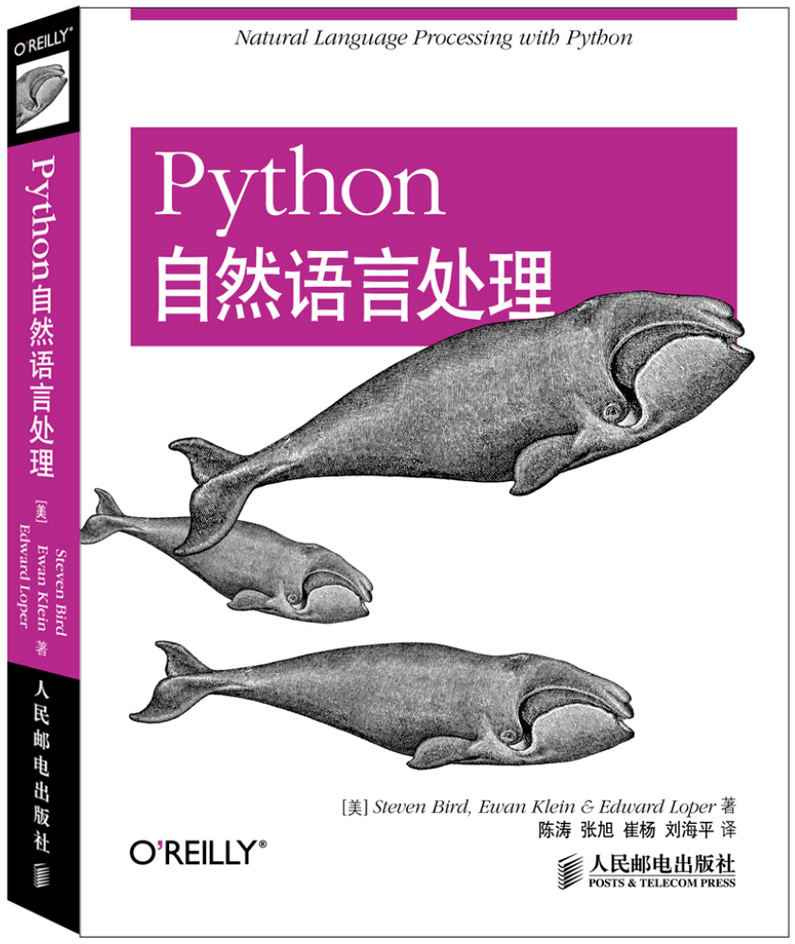
(a) XRN (fine-tune) (b) Gen-XRN (i2t+t2i)

Figure 6: Visualization of word embedding.

课程详情扫码咨询

www.mooc.ai

Q / A



《Python自然语言处理》
【美】Steven Bird , Ewan Klein , Edward Loper