Bài học từ cuộc thi tạo mô tả cho ảnh MSCOCO 2015

Môn: Nhập môn học máy

Thông tin nhóm

Nhóm 11	
Họ và tên	**MSSV**
Trần Thanh Tùng	**18120258**
Trần Hữu Chí Bảo	**18120288**
Vòng Cảnh Chi	**18120293**
Cao Tất Cường	**18120296**
Hà Văn Duy	**18120339**

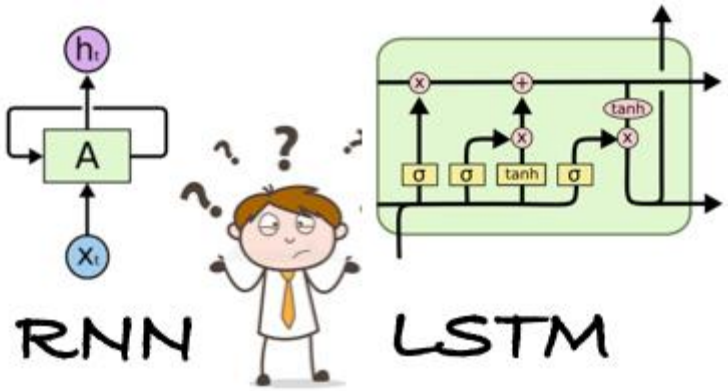
Giới thiệu

"man in black shirt is playing guitar."

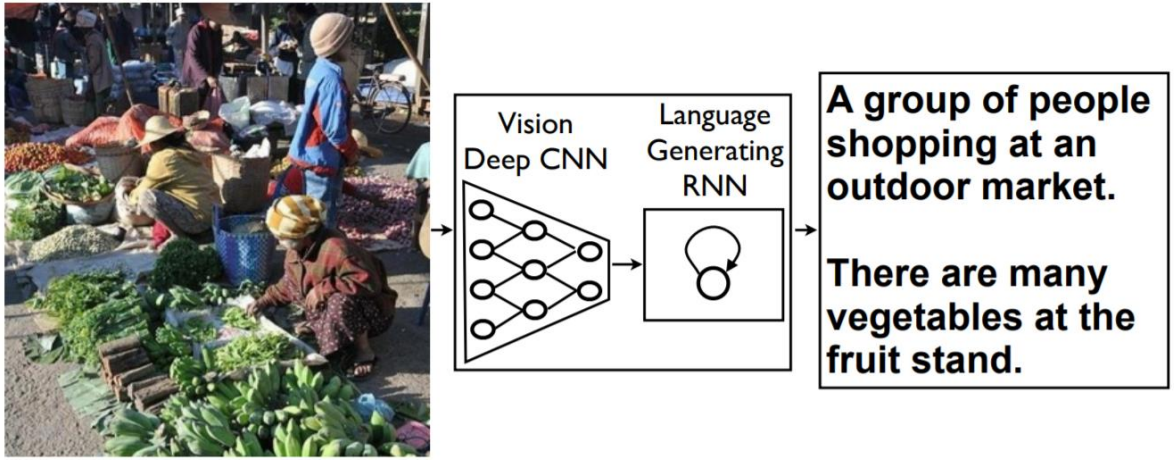
"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

Những nghiên cứu liên quan



Mô hình đề xuất



Ý tưởng mô hình

Ta dùng RNN để lập mô hình p(S_t|I, S_0, . . . , S_{t-1}). Trạng thái ẩn được cập nhật khi có đầu vào mới bằng cách sử dụng hàm phi tuyến tính f:

$$h_{t+1} = f(h_t, x_t) \ . \qquad (3)$$

Đối với hình ảnh, chúng tôi sử dụng CNN sử dụng phương pháp batch normalization

Các từ được biểu diễn bằng một mô hình nhúng

Ý tưởng mô hình

Công thức để tối đa hóa xác suất của mô tả đúng

$$\theta^{\star} = \arg\max_{\theta} \sum_{(I,S)} \log p(S|I;\theta) \qquad (1)$$

Trong đó:

- θ là các tham số mô hình
- I là một hình ảnh
- S là câu mô tả đúng của ảnh

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1}) \qquad (2)$$

LSTM-based Sentence Generator

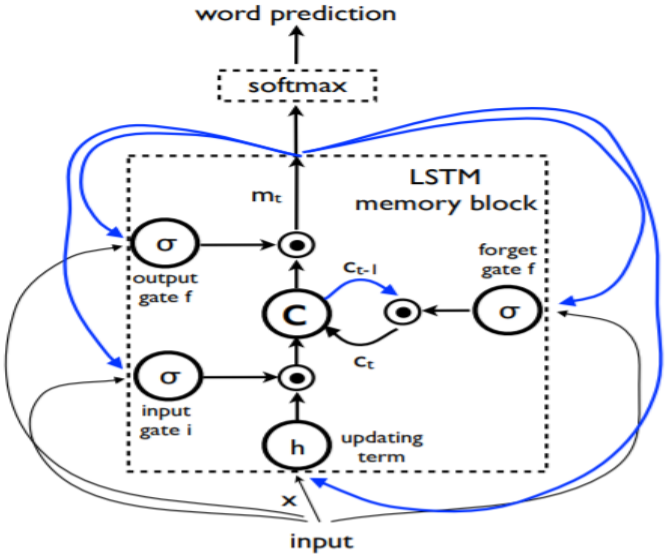


Fig. 2. LSTM: the memory block contains a cell c which is controlled by three gates. In blue we show the recurrent connections – the output m at time $t-1$ is fed back to the memory at time t via the three gates; the cell value is fed back via the forget gate; the predicted word at time $t-1$ is fed back in addition to the memory output m at time t into the Softmax for word prediction.

LSTM-based Sentence Generator

$$
\begin{aligned}
i_t &= \sigma(W_{ix}x_t + W_{im}m_{t-1}) && (4)\\
f_t &= \sigma(W_{fx}x_t + W_{fm}m_{t-1}) && (5)\\
o_t &= \sigma(W_{ox}x_t + W_{om}m_{t-1}) && (6)\\
c_t &= f_t \odot c_{t-1} + i_t \odot h(W_{cx}x_t + W_{cm}m_{t-1}) && (7)\\
m_t &= o_t \odot c_t && (8)\\
p_{t+1} &= \text{Softmax}(m_t) && (9)
\end{aligned}
$$

Training

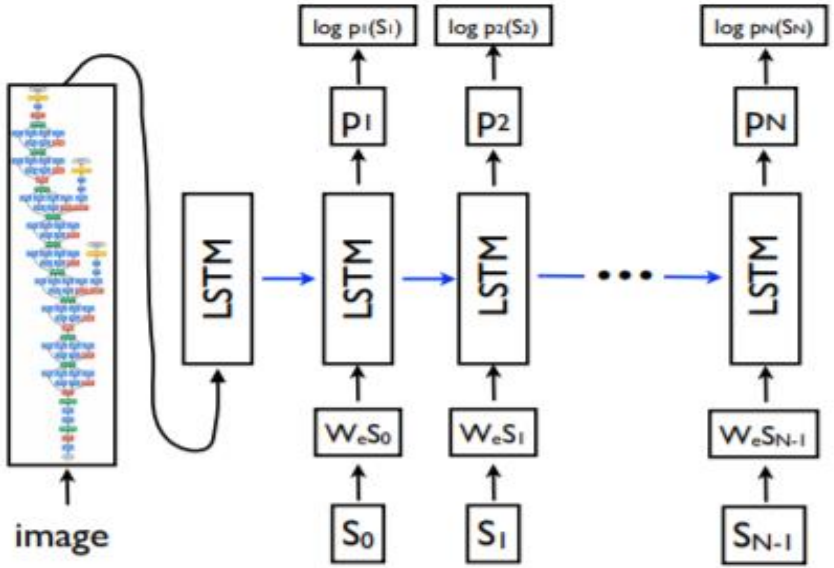


Fig. 3. LSTM model combined with a CNN image embedder (as defined in [24]) and word embeddings. The unrolled connections between the LSTM memories are in blue and they correspond to the recurrent connections in Figure 2. All LSTMs share the same parameters.

Training

$$x_{-1} = \mathrm{CNN}(I) \tag{10}$$

$$x_t = W_e S_t, \quad t \in \{0 \dots N-1\} \tag{11}$$

$$p_{t+1} = \mathrm{LSTM}(x_t), \quad t \in \{0 \dots N-1\} \tag{12}$$

Training

Hàm loss là tổng negative log likelihood của từ đúng ở mỗi bước có công thức như sau:

$$L(I, S) = -\sum_{t=1}^{N} \log p_t(S_t) \ . \qquad (13)$$

EXPERIMENTS

Evaluation Metrics

Đánh giá output của mô hình dựa trên các độ đo:

- 4 bậc độ liên quan của output so với hình
- Dùng BLEU score (Bilingual Evaluation Understudy)
- CIDER
- METEOR và ROUGE

Dataset

Dataset name	size		
	train	valid.	test
Pascal VOC 2008 [2]	-	-	1000
Flickr8k [42]	6000	1000	1000
Flickr30k [43]	28000	1000	1000
MSCOCO [44]	82783	40504	40775
SBU [18]	1M	-	-

Training details

- Dùng các pre-trained CNN model (ImageNet) để hạn chế overfiting
- Khắc phục overfitting bằng các kĩ thuật có hiệu quả như dropout và ensemble
- Huấn luyện bằng hàm cực tiểu SGD với fixed learning rate và không dùng momentum
- LSTM để decode và embedding

Generation Results

TABLE 1
Scores on the MSCOCO development set for two models: NIC, which was the model which we developed in [46], and NICv2, which was the model after we tuned and refined our system for the MSCOCO competition.

Metric	BLEU-4	METEOR	CIDER
NIC	27.7	23.7	85.5
NICv2	**32.1**	**25.7**	**99.8**
Random	4.6	9.0	5.1
Nearest Neighbor	9.9	15.7	36.5
Human	21.7	25.2	85.4

Generation Results

TABLE 2
BLEU-1 scores. We only report previous work results when available.
SOTA stands for the current state-of-the-art.

Approach	PASCAL (xfer)	Flickr 30k	Flickr 8k	SBU
Im2Text [18]				11
TreeTalk [14]				19
BabyTalk [3]	25			
Tri5Sem [16]			48	
m-RNN [27]		55	58	
MNLM [29][5]		56	51	
SOTA	25	56	58	19
NIC	**59**	**66**	**63**	**28**
Human	69	68	70	

Transfer Learning, Data Size and Label Quality

TABLE 2
BLEU-1 scores. We only report previous work results when available.
SOTA stands for the current state-of-the-art.

Approach	PASCAL (xfer)	Flickr 30k	Flickr 8k	SBU
Im2Text [18]				11
TreeTalk [14]				19
BabyTalk [3]	25			
Tri5Sem [16]			48	
m-RNN [27]		55	58	
MNLM [29][5]		56	51	
SOTA	25	56	58	19
NIC	**59**	**66**	**63**	**28**
Human	69	68	70	

Generation Diversity

TABLE 3
N-best examples from the MSCOCO test set. Bold lines indicate a novel sentence not present in the training set.

A man throwing a frisbee in a park. **A man holding a frisbee in his hand.** **A man standing in the grass with a frisbee.**
A close up of a sandwich on a plate. A close up of a plate of food with french fries. A white plate topped with a cut in half sandwich.
A display case filled with lots of donuts. **A display case filled with lots of cakes.** **A bakery display case filled with lots of donuts.**

Sheduled sampling

Các LSTM được huấn luyện bằng cách cố gắng dự đoán từng từ của chú thích với trạng thái hiện tại của mô hình và từ trước đó trong chú thích. Đối với một hình ảnh mới, từ trước đó hiển nhiên là không xác định và do đó được thay thế bằng từ do chính mô hình tạo ra ở bước trước đó.

Ranking Results

TABLE 4
Recall@k and median rank on Flickr8k.

Approach	Image Annotation			Image Search		
	R@1	R@10	Med r	R@1	R@10	Med r
DeFrag [22]	13	44	14	10	43	15
m-RNN [27]	15	49	11	12	42	15
MNLM [29]	18	55	8	13	52	10
NIC	**20**	**61**	**6**	**19**	**64**	**5**

TABLE 5
Recall@k and median rank on Flickr30k.

Approach	Image Annotation			Image Search		
	R@1	R@10	Med r	R@1	R@10	Med r
DeFrag [22]	16	55	8	10	45	13
m-RNN [27]	18	51	10	13	42	16
MNLM [29]	**23**	**63**	**5**	**17**	**57**	**8**
NIC	17	56	7	**17**	**57**	**7**

Analysis of Embeddings

TABLE 6
Nearest neighbors of a few example words

Word	Neighbors
car	van, cab, suv, vehicule, jeep
boy	toddler, gentleman, daughter, son
street	road, streets, highway, freeway
horse	pony, donkey, pig, goat, mule
computer	computers, pc, crt, chip, compute

Những cải tiến

Technique	BLEU-4 Improvement
Better Image Model [24]	2
Beam Size Reduction	2
Fine-tuning Image Model	1
Scheduled Sampling [48]	1.5
Ensembles	1.5

Phần 5.3 trình bày tóm tắt kết quả trên cả số liệu tự động và số liệu con người từ cuộc thi MSCOCO.

Better Image Model

Tác giả sử dụng CNN cụ thể sử dụng mô hình GoogleLeNet, mô hình 22 lớp. Sau đó, người ta đã sử dụng cách tiếp cận tốt hơn đã được đề xuất, mô hình sử dụng thêm Batch Normalization, để chuẩn hóa tốt hơn từng lớp của mạng neural, nó sẽ mạnh mẽ hơn đối với tính phi tuyến. Cách tiếp cận mới đã cải thiện đáng kể tác vụ ImageNet (giảm từ 6,67% xuống 4,8% lỗi) và tác vụ phụ đề hình ảnh MSCOCO, cải thiện BLEU-4 2 điểm tuyệt đối.

Fine-tuning Image Model

- Đối với cuộc thi, tác giả cũng đã xem xét thêm một số tinh chỉnh của mô hình trong khi đào tạo LSTM, giúp mô hình tập trung hơn vào loại hình ảnh được cung cấp trong bộ đào tạo MS COCO và cuối cùng đã cải thiện hiệu suất trong nhiệm vụ phụ đề.
- Những cải tiến đã đạt được 1 điểm BLEU-4.

Ensembling

Tạo ra một ensembles gồm 5 model train scheduled sampling và 10 model train với việc tinh chỉnh mô hình. Mô hình kết quả đã được gửi đến cuộc thi và nó tiếp tục cải thiện kết quả thêm 1,5 điểm BLEU-4.

Beam size reduction

- Tác giả đã thử nhiều kích thước chùm khác và chọn kích thước tạo ra chuỗi từ tốt nhất theo chỉ số CIDER.
- Khi kích thước chùm tăng lên, tác giả cho nhiều từ hơn và chọn câu tốt nhất theo khả năng thu được. Do đó, nếu mô hình được training tốt và khả năng xảy ra phù hợp với phán đoán của con người, thì việc tăng kích thước chùm tia sẽ luôn mang lại các câu tốt hơn. Thực tế, đã thu được hiệu suất tốt nhất với kích thước chùm tia tương đối nhỏ là một dấu hiệu cho thấy mô hình đã bị overfit được sử dụng để đào tạo nó (likelihood) không phù hợp với đánh giá của con người.

Đánh giá tự động

	CIDER	METEOR	ROUGE	BLEU-4	Rank
Google [46]	0.943	0.254	0.53	0.309	1st
MSR Captivator [34]	0.931	0.248	0.526	0.308	2nd
m-RNN [28]	0.917	0.242	0.521	0.299	3rd
MSR [23]	0.912	0.247	0.519	0.291	4th
m-RNN (2) [28]	0.886	0.238	0.524	0.302	5th
Human	0.854	0.252	0.484	0.217	8th

Hình. 5 bài nộp hàng đầu theo các chỉ số tự động trên bộ thử nghiệm (được sắp xếp theo CIDER).

Đánh giá con người

	M1	M2	M3	M4	M5	Rank
Google [46]	0.273	0.317	4.107	2.742	0.233	1st
MSR [23]	0.268	0.322	4.137	2.662	0.234	1st
MSR Captivator [34]	0.250	0.301	4.149	2.565	0.233	3rd
Montreal/Toronto [31]	0.262	0.272	3.932	2.832	0.197	3rd
Berkeley LRCN [30]	0.246	0.268	3.924	2.786	0.204	5th
Human	0.638	0.675	4.836	3.428	0.352	1st

Hình. 5 bài dự thi hàng đầu theo các chỉ số này (được sắp xếp theo M1 + M2).

Kết luận

Nhóm vừa giới thiệu NIC, một hệ thống hoàn chỉnh có thể sinh mô tả từ ảnh bằng tiếng anh đơn giản. NIC dựa trên một mạng cnn để mã hóa hình ảnh và mạng rnn sử dụng kết quả của mạng cnn để sinh mô tả ảnh. Mô hình được huấn luyện để tối đa hóa likelihood của câu dựa vào ảnh. Thử nghiệm trên một số bộ dataset đã cho thấy sự ưu việt hơn của NIC. Dựa vào những kết quả ban đầu này, nhóm tác giả đã tham gia cuộc thi MS COCO 2015 và so sánh với các phương pháp khác về điểm số.