Which Settings Give Better Results?

- Experiments on image captioning in Chinese for Flickr8k and Flickr30k images

Xi Chen

Goal

NOT to propose any model for image captioning in Chinese which can compete with the state-of-the-art research

conduct several experiments on generating Chinese captions for both Flickr8K and Flickr30K images

train models with different settings

Evaluate and compare their performance

Data

	Flickr8k-CN			Flickr30k-CN		
	train	val	test	train	val	test
Images	6000	1000	1000	29783	1000	1000
Human-annotated Chinese sentences	30000	5000	5000			
Machine-translated Chinese sentences	30000	5000		148915	5000	
Human-translated Chinese sentences			5000			5000

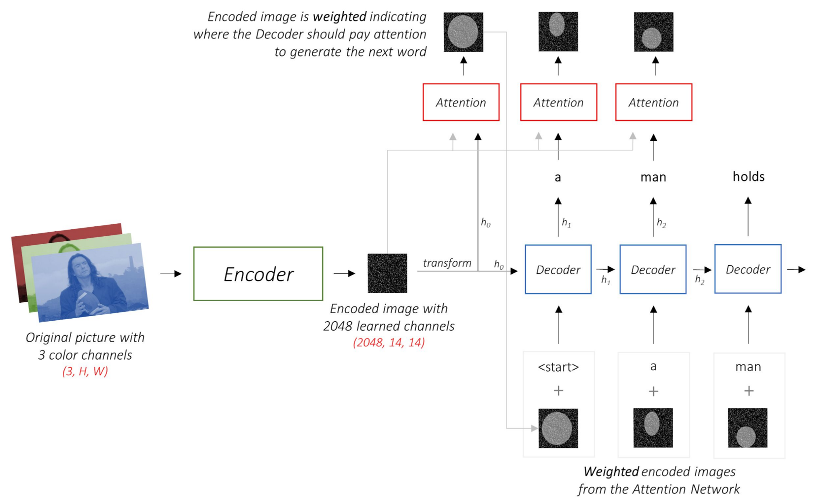
Method

Code:

https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning

Model

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044 (2015)



Method: Experiment Settings

Segmentation-Based and character-Based

ResNet101 and VGG19

Feature Extraction and Fine-Tuning

Models with and without Best BLEU Scores

Train

Epochs: 100 (All of them early-stopped at around 30-40 epochs)

Learning rate for encoder: 1e-4 (Fine-tune)

Learning rate for decoder: 4e-4

Validation after each epoch, if BLEU has not improved in 8 epochs, the learning rate will be decreased by 0.8

Early-stop if BLEU score has not improved in 20 epochs.

6 Different Settings x 2 (Flickr8k and 30k) x 2 (Best BLEU and 20 more epochs) = 24 saved checkpoints

	Character-based	Segmenta-tion-based	Pre-trained VGG19	Pre-trained ResNet101	Feature Extraction	Fine-tuning
Setting 1	x			x	x	
Setting 2		x		x	x	
Setting 3	x			x		x
Setting 4		x		x		x
Setting 5	x		x		x	
Setting 6		x	x		x	

Evaluation Criteras

BLEU-1, BLEU-2, BLEU-3, BLEU-4,

ROUGE_L

CIDEr

Not Meteor (No stemming, No synonym vocab)

Evaluation Results

models trained on Flickr8k-CN got much better scores than models trained on Flickr30k-CN

(the ground truth captions in the train and val split of Flickr30k-CN were machine-translated, while captions in test split were human-translated)

Evaluation Results

Which Beam Size Performs Better?

- No fixed pattern, depends on the settings
- In most cases, beam size between 3 and 5 gave higher scores

Evaluation Results

Character-Based vs Segmentation-Based

- almost all character-based models have gotten higher scores

Captioning Examples 10

BB	30k	seg	FE	RN	'一', '只', '白色', '的', '狗', '嘴里', '叼', '着', '一个', '玩具'
					a white dog has a toy in its mouth
BB	30k	char	FE	RN	'一', '只', '白', '色', '的', '狗', '在', '它', '的', '嘴', '里', '叼', '着', '一', '个', '紫', '色', '的', '飞', '盘'
					a white dog has a purple frisbee in its mouth
BB	8k	seg	FE	RN	'一', '只', '白色', '的', '狗', '带', '着', '紫色', '的', '飞盘'
					a white dog holds a purple frisbee
BB	8k	char	FE	RN	'一', '只', '白', '色', '的', '狗', '叼', '着', '一', '个', '紫', '色', '的', '飞', '盘'
					a white dog has a purple frisbee in its mouth
BB	30k	seg	FT	RN	一', '只', '白色', '的', '狗', '正在', '玩', '一个', '紫色', '的', '<unk>'
					a white a dog is playing a purple <unk>
BB	30k	char	FT	RN	一', '只', '白', '色', '的', '狗', '拿', '着', '一', '个', '紫', '色', '的', '飞', '盘'
					a white dog holds a purple frisbee
BB	8k	seg	FT	RN	'一', '只', '白色', '的', '狗', '在', '草地', '上', '玩耍'
					a white dog is playing on the grass

飞 (fly) + 盘 (plate) = 飞盘 (frisbee)

Evaluation Results

ResNet101 vs VGG19

- almost all models using pre-trained ResNet101 have achieved better BLEU, ROUGE_L.
- one exception: “BB-30k-seg-FE” using VGG19 got better BLEU and ROUGE_L scores than using ResNet101
- All using ResNet101 got better CIDEr scores

Evaluation Results

Feature Extraction vs Fine-Tuning

- almost all “BB” models, fine-tuning the encoder gave better BLEU and ROUGE_L scores
- NB models (models trained for 20 more epochs): not fine-tuning the encoder gave better BLEU and ROUGE_L scores
- almost all of the models without fine-tuning got better CIDEr scores, except one marked as “BB-30k-seg-RN”

Evaluation Results

Checkpoints with Best BLEU Score vs Continue Training (20 More Epochs)

- models marked by “BB” achieved higher scores
- early-stopping when BLEU-4 score does not increase more seems to be a good approach, avoiding overfitting

Discussion

Evaluation Criterias

- Different aspects of a language (e.g. culture)
- Not always reliable
- some BLEU scores or ROUGE_L scores showed that some models are better, while the CIDEr showed the opposite result.
- Human evaluation should be considered if available

Discussion

Training Data Bias

- models trained on Flickr30k-CN got much lower scores (machine-translated ground truth)
- When the pictures have nothing to do with humans or animals, the results were bad. (most of the captions in the train set contain people or animals)
- a picture of zebras, was generated as “dogs” and “birds” instead. (The tokens for zebra probably do not exist in the training data)

Captioning Examples 7

BB	30k	seg	FE	RN	'一', '只', '棕色', '的', '狗', '正', '站', '在', '一个', '装满', '水果', '的', '篮子', '里'
					a brown dog is standing in a basket full of fruits
BB	30k	char	FE	RN	'一', '个', '女', '人', '坐', '在', '一', '个', '水', '果', '摊', '旁', '边'
					a woman is beside a fruit stall
BB	8k	seg	FE	RN	'一个', '小', '男孩', '坐', '在', '一', '张', '桌子', '上'
					a little is sitting on a table
BB	8k	char	FE	RN	'一', '个', '特', '写', '镜', '头', '的', '男', '孩'

Captioning Examples 9

BB	30k	seg	FE	RN	'一个', '人', '在', '码头', '上', '散步'
					a person walking on a shipside
BB	30k	char	FE	RN	'一', '个', '人', '站', '在', '一', '座', '桥', '上', '一', '座', '桥'
					a person stands on a bridge a bridge
BB	8k	seg	FE	RN	'两', '只', '人', '站', '在', '一', '座', '桥', '上'
					two piece of persons standing on a bridge
BB	8k	char	FE	RN	'一', '个', '人', '站', '在', '一', '座', '桥', '上', '俯', '瞰', '着', '水'
					a person standing on a bridge looking down at the water

Discussion

Gender Bias

- a picture: a man is holding a baby -> “a woman is holding a boy”, even though the features of the man were very clear. (probability that a woman appears together with a baby is much higher than a man with a baby)

Captioning Examples 1

BB	30k	seg	FE	RN	'一个', '穿', '着', '蓝色', '衬衫', '的', '年轻', '男孩', '在', '看', '他', '的', '手机'
					A young boy in a blue shirt is looking at his mobile phone
BB	30k	char	FE	RN	'一', '个', '年', '轻', '的', '女', '人', '看', '着', '她', '的', '电', '话'
					A young woman is looking at her telephone
BB	8k	seg	FE	RN	'一个', '穿', '着', '蓝色', '衬衫', '的', '男人', '在', '看', '着', '他', '的', '手机'
					A man in a blue shirt is looking at his mobile phone
BB	8k	char	FE	RN	'一', '个', '小', '女', '孩', '看', '了', '看', '相', '机'
					A young girl looks at camera

Discussion

Quantity Errors

['<start>', '一', '群', '人', '在', '一', '辆', '摩托车', '上', '<end>']

A group of people on one motorbike

['<start>', '一个', '骑', '摩托车', '的', '人', '<end>']

A person riding a motorbike

Future Works

MS COCO

Gender bias

Quantity errors

Other evaluation criteria - human evaluation

Thanks!