Dense Captioning with Joint Inference and Visual Context

Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li

Snap Inc.

Introduction

Dense captioning is a newly emerging computer vision topic for understanding images with dense language descriptions.

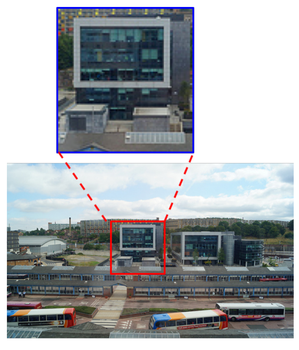
Goals

1. densely detect visual concepts
2. label each with a short description

Two key challenges

1. Highly overlapping target regions

2. Ambiguous objects without context



Building or desktop?

Contribution

We design network structures that incorporate two novel ideas:

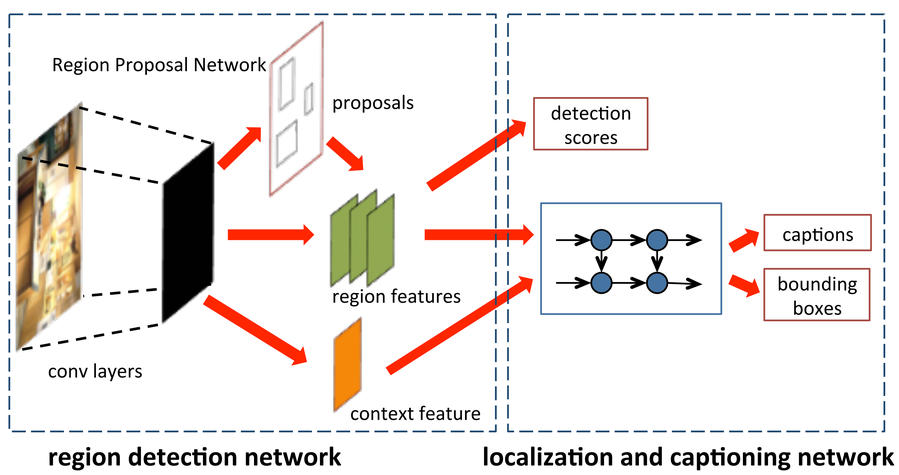
joint inference Localization guided by semantic information

context fusion Description guided by visual context

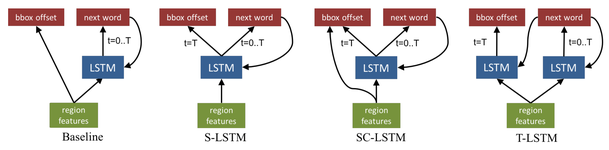
Our best model achieves a 73% relative gain over the previous state-of-the-art.

Approach

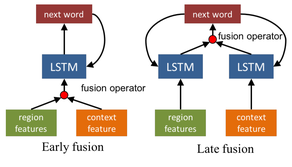
End-to-end network with two stages: a region detection network and a localization and captioning network



Joint inference structure

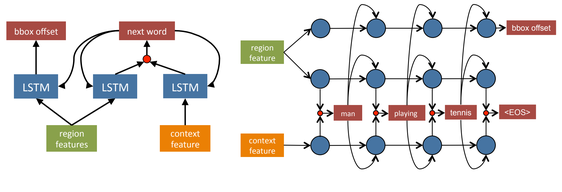


Context fusion structure



Integrated model

We can combine any joint inference structure with any context fusion structure. Combination of T-LSTM + Late fusion looks like



Experiments

Visual Genome V1.0

- Baseline and joint inference models

model	Johnson et.al.	baseline	S-LSTM	SC-LSTM	T-LSTM
fixed-CNN&RPN	-	5.26	5.15	5.57	5.64
end-to-end	5.39	6.85	6.47	6.83	**8.03**

- Integrated models

model		S-LSTM	SC-LSTM	T-LSTM
early-fusion	$[\cdot,\cdot]$	6.74	7.18	8.24
	$\oplus$	6.54	7.29	8.16
	$\otimes$	6.69	7.04	8.19
late-fusion	$[\cdot,\cdot]$	7.50	7.72	8.49
	$\oplus$	7.19	7.47	8.53
	$\otimes$	7.57	7.64	**8.60**

Visual Genome V1.2

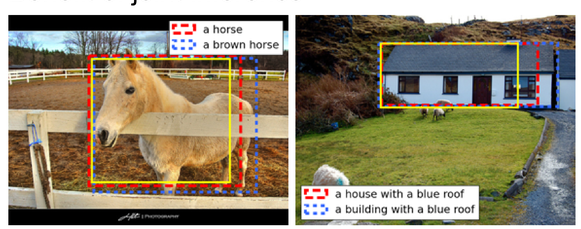
model		baseline	S-LSTM	T-LSTM
no context		6.98	6.44	8.16
late-fusion	$[\cdot,\cdot]$		7.76	**9.03**
	$\oplus$		7.06	8.71
	$\otimes$		7.63	8.52

Best practice: hyper-parameters

	#proposal	NMS_r1	NMS_r2	mAP
V1.0	100	0.5	0.4	8.67
	300	0.6	0.5	**9.31**
V1.2	100	0.5	0.5	9.47
	300	0.6	0.5	**9.96**

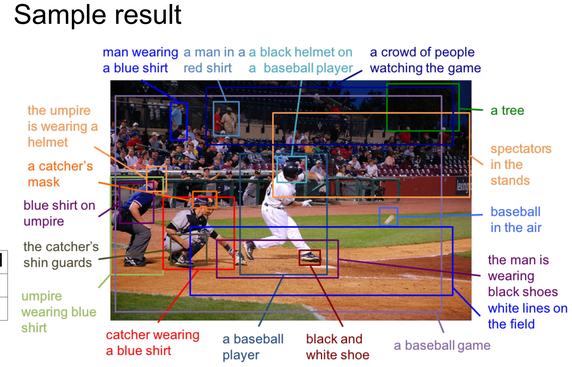
Visualization

Benefit of joint inference



Benefit of context fusion

Sample result



Conclusions

- Joint inference and context fusion increase the capacity of the model
- The flow of language and visual information in the network should be carefully designed
- The learned representation can potentially benefit other computer vision tasks, such as object detection and semantic segmentation.

Code Released!

- http://github.com/linjieyangsc/densecap
- Contact email: linjie.yang@snap.com