Areas of Attention in Image Captioning

Marco Pedersoli*, Thomas Lucas
Cordelia Schmid, Jakob Verbeek

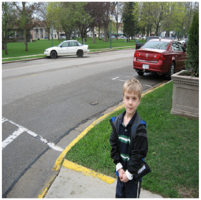
INRIA Grenoble Rhône-Alpes, France
* Now at École de Technologie Supérieure Montreal, Canada

Image captioning

- Given an image, generate a natural language description

Figure taken from [Kiros et al., 2015]

a black and white photo of a window .



a young boy standing on a parking lot next to cars .

a wooden table and chairs arranged in a room .

a car is parked in the middle of nowhere .

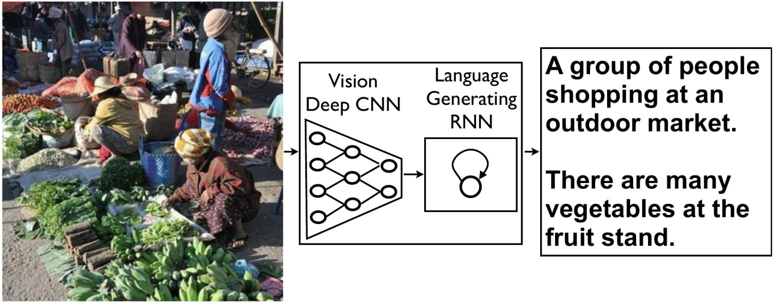
a ferry boat on a marina with a group of people .

a little boy with a bunch of friends on the street .

Encoder-decoder models for captioning

- State of the art based on encoder-decoder approach [Kiros et al., 2014]
  - Inspired from encoder-decoder models in machine translation, see *e.g.* [Sutskever et al., 2014]
- Encoder transforms input to a internal representation
- Decoder maps internal representation to output

Figure taken from [Vinyals et al., 2015]



Limitations

- Only discriminative training
  - Pure-text corpus to better learn language?
  - Image-only data to learn image parser?
- Limited to a fixed vocabulary
  - How to generalize better from few examples?
  - Character-level prediction?
- Single image parse into a vector representation
  - Global image representation, how to get compositionality?
  - How to offload visual content from memory state?

Figure taken from [Kiros et al., 2015]

a giraffe is standing next to a fence in a field .
(hallucination)

the two birds are trying to be seen in the water .
(counting)

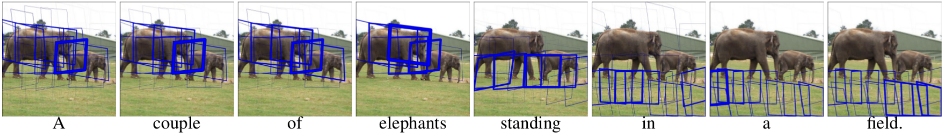
a parked car while driving down the road .
(contradiction)

the handlebars are trying to ride a bike rack .
(nonsensical)

a woman and a bottle of wine in a garden .
(gender)

Leveraging locality and compositionality with attention

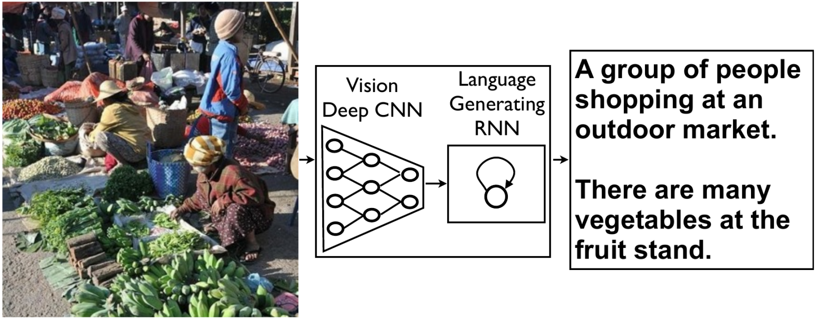
- **Sequentially** attend to different **parts** of the input
- Associate local image evidence with words in caption
- Also used in speech recognition and machine translation



- **Which areas** to consider?
- **Which mechanism** to exploit these areas?

Baseline: "vanilla" captioning system

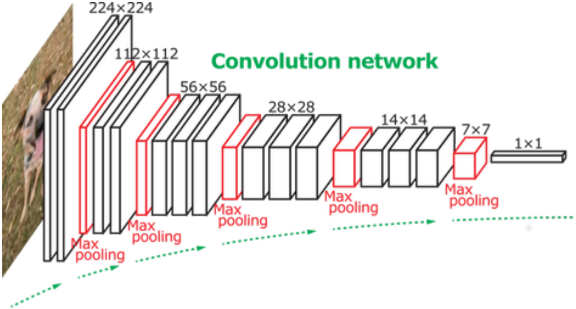
Figure taken from [Vinyals et al., 2015]



Encoder

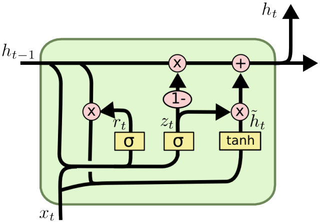
- CNN with VGG-16 architecture [Simonyan and Zisserman, 2015]
  - 16 layers with trainable weights, 138M parameters
  - Penultimate layer of ImageNet pre-trained model

Figure taken from [Noh et al., 2015]



Decoder

- GRU-based RNN [Chung et al., 2014]
  - State initialized with CNN code
  - Previous word used as input: "output feedback"

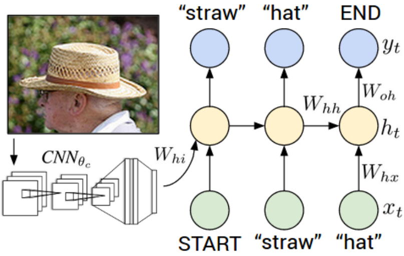


$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$



Figures taken from [Karpathy and Fei-Fei, 2015] and `http://colah.github.io`

Baseline model: word prediction

- Baseline RNN is based on state-word interactions

$$p(w_t|h_t) \propto \exp\left(w_t^\top W \theta_{wh} h_t\right) \quad (1)$$

  - w_t: 1-hot coding of word at time t
  - W: contains word-embedding vectors in rows
  - θ_{wh}: parameter matrix to score word-state combination
  - Think: "a logistic discriminant word-classifier given state"

- Train: maximum-likelihood using ground-truth inputs for state evolution ("teacher forced")

- Test: Generate approximate maximum-likelihood sentences with beam-search

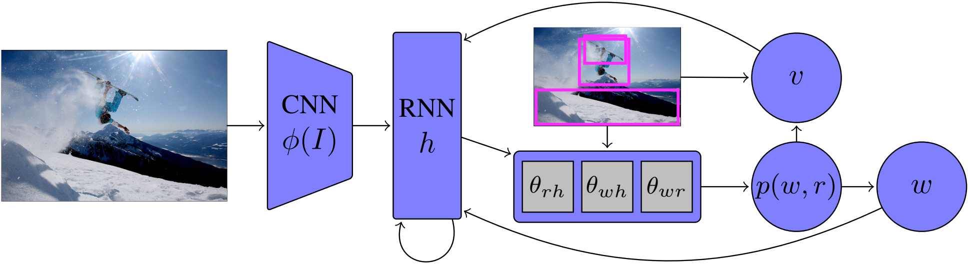
Our “Areas of Attention” model

- Based on scoring state-word-region combinations
  - Which region-word pair “stands out” given the current state?

$$
\begin{aligned}
p(w_t, r_t | h_t) &\propto \exp s(w_t, r_t, h_t), & (2)\\
s(w_t, r_t, h_t) &= w_t^\top W \theta_{wh} h_t + w_t^\top W \theta_{wr} R^\top r_t \\
&\quad + r_t^\top R \theta_{rh} h_t + w_t^\top W \theta_w + r_t^\top R \theta_r, & (3)
\end{aligned}
$$

- w_t: 1-hot coding of word at time t
- W: contains word-embedding vectors in rows
- r_t: 1-hot coding of region at time t
- R: contains region feature vectors in rows
- $\theta_{wh}, \theta_{wr}, \theta_{rh}$: region-word-state interaction matrices
- θ_w, θ_r: region and word bias vectors

Our "Areas of Attention" model



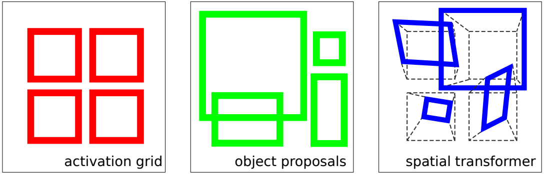
- Predict words using $p(w_t|h_t) = \sum_{r_t} p(w_t, r_t|h_t)$
- Use appearance of attended regions for state update

$$v_t = \sum_{r_t} p(r_t|h_t) r_t^\top R, \tag{4}$$

$$h_{t+1} = \mathrm{GRU}(h_t, [w_t^\top W \quad v_t^\top]^\top). \tag{5}$$

And how about the regions?

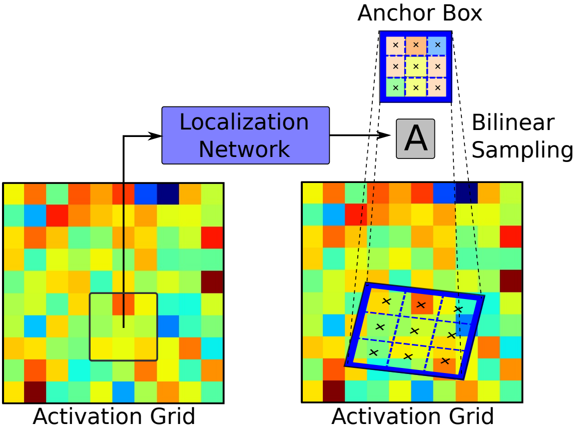
- Our AoA model is agnostic to type of image region, experimentally we compare three different region types



- Activation grid: take positions of conv5 layer as regions, descriptor is "column" of activations across feature channels
- Object proposals: using EdgeBox object proposals [Zitnick and Dollár, 2014], average conv5 features over box
- Spatial transformer: predict region from each conv4 position, compute conv5 features over warped 3×3 area

Spatial transformer regions

- Localization network regresses affine transformations for all feature map positions
- Transformations are applied to the anchor boxes that are used to locally re-sample the feature map, before convolution
- Reverts to “Activation grid” for identity transformation



Microsoft Common Objects in Context (MSCOCO)

- 80k train, 40 development images, 5 sentences per image

1. A woman kneeling down next to a dog on a snow covered slope.
2. A boy and his dog are playing in the snow.
3. A snowboarder in a blue jacket and a black and brown dog.
4. Snowboarder sitting next to a dog in the snow.
5. A snowboarder sits in snow beside a dog.

Evaluation of model components

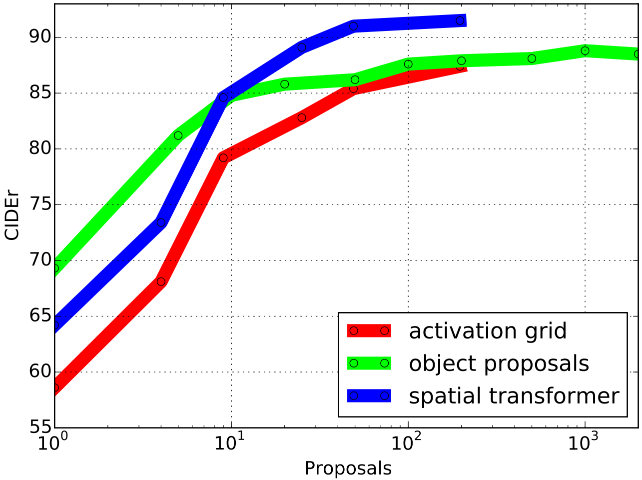
- Using activation grid as attention areas

Method	B1	B4	Meteor	CIDEr
Baseline: θ_{wh}	66.3	26.4	22.2	78.9
Ours: θ_{wh}, θ_{wr}	68.0	28.0	22.9	83.6
Ours: $\theta_{wh}, \theta_{wr}, \theta_{rh}$	68.2	28.4	23.3	85.5
Ours: conditional feedback	68.3	28.7	**23.7**	86.8
Ours: full model	**69.1**	**28.8**	**23.7**	**87.4**

- Local word-region interaction improves
- Local region-state interaction improves
- Word-conditioning visual feedback, *i.e.* using $p(r_t|w_t, h_t)$ instead of $p(r_t|h_t)$, degrades w.r.t. full model

Evaluation of attention areas

- Object proposals: top regions by “objectness”
- Grids + transformers: regular sampling

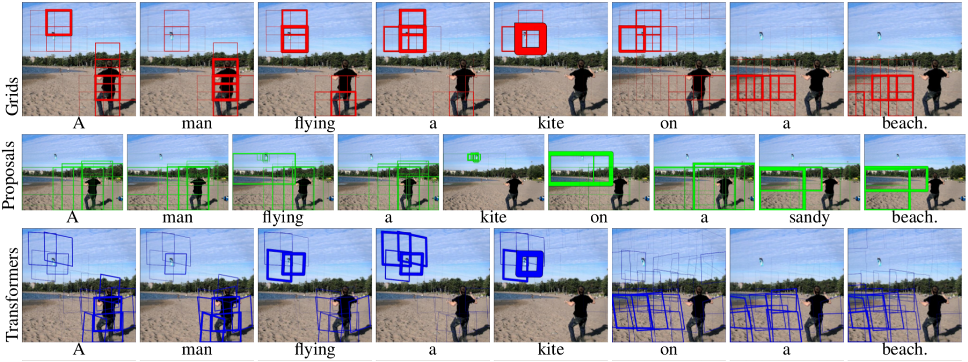


Effect of CNN fine-tuning

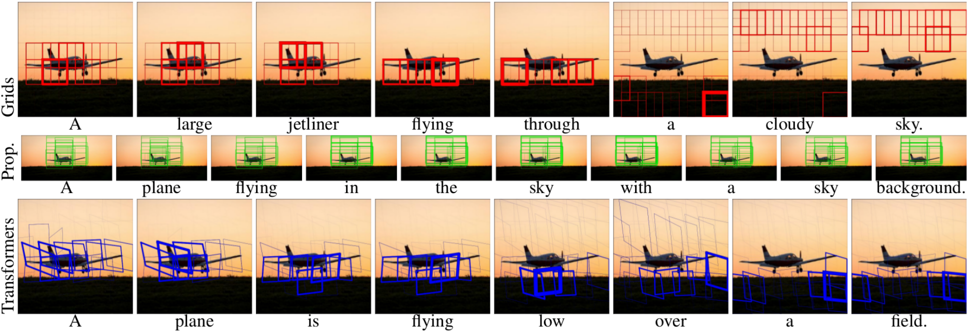
- RNN training only: fixed pre-trained CNN
- CNN-RNN fine-tuning: second stage trains all

Method	B1	B4	Meteor	CIDEr
	RNN training only			
Baseline	66.3	26.4	22.2	78.9
Spatial transformers	**70.2**	**30.2**	**24.2**	**91.1**
	CNN-RNN fine-tuning			
Baseline	68.6	28.7	23.5	87.1
Spatial transformers	**70.8**	**30.7**	**24.5**	**93.8**

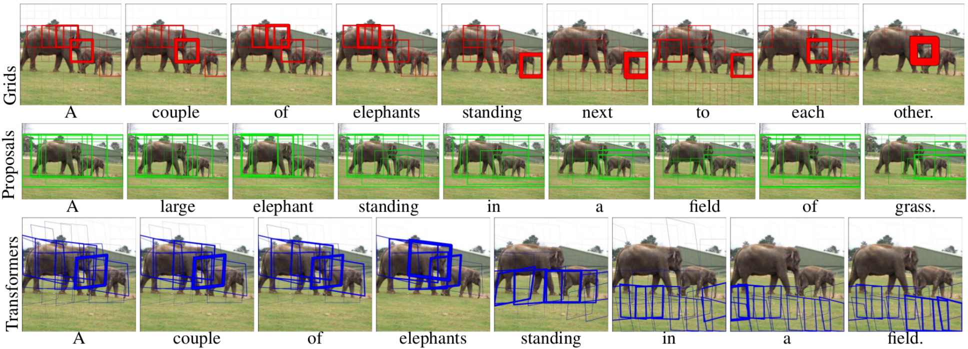
Comparison of attention areas



Comparison of attention areas



Comparison of attention areas

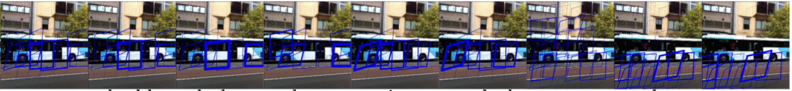


Comparison to the state of the art

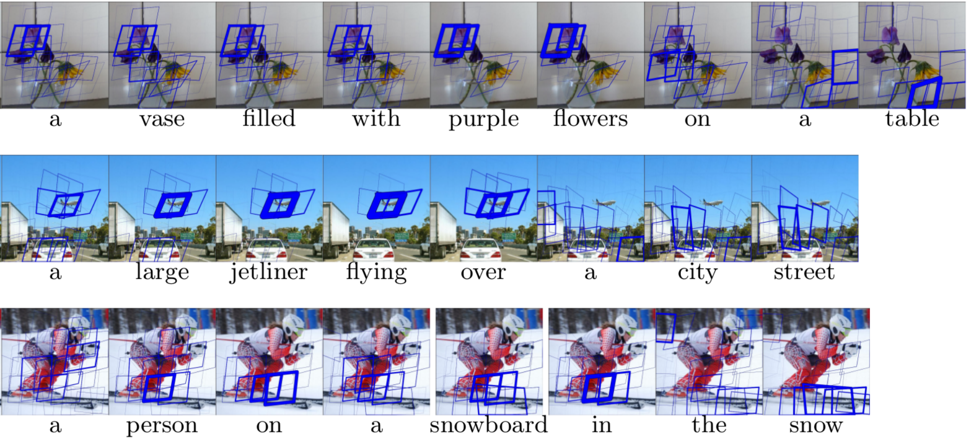
- Competitive with state-of-the-art methods
- More data (80k+30k) improves performance
- Ensemble of training with different seeds expected to improve

Method	B1	B4	Meteor	CIDEr
Vinyals *et al*. [Vinyals et al., 2015]	-	27.7	23.7	85.5
Xu *et al*. [Xu et al., 2015], soft	70.9	24.3	23.9	-
Xu *et al*. [Xu et al., 2015], hard	**71.8**	25.0	23.0	-
Yang *et al*. [Yang et al., 2016]	-	29.0	23.7	88.6
Jin *et al*. [Jin et al., 2015]	69.7	28.2	23.5	83.8
Donahue *et al*. [Donahue et al., 2015]	71.1	30.0	24.2	89.6
Ranzato *et al*. [Ranzato et al., 2016]	-	29.2	-	-
Bengio *et al*. [Bengio et al., 2015]	-	30.6	24.3	92.1
Areas of Attention (ours)	70.8	**30.7**	**24.5**	**93.8**
AoA, data augmentation	72.1	31.1	25.0	95.6

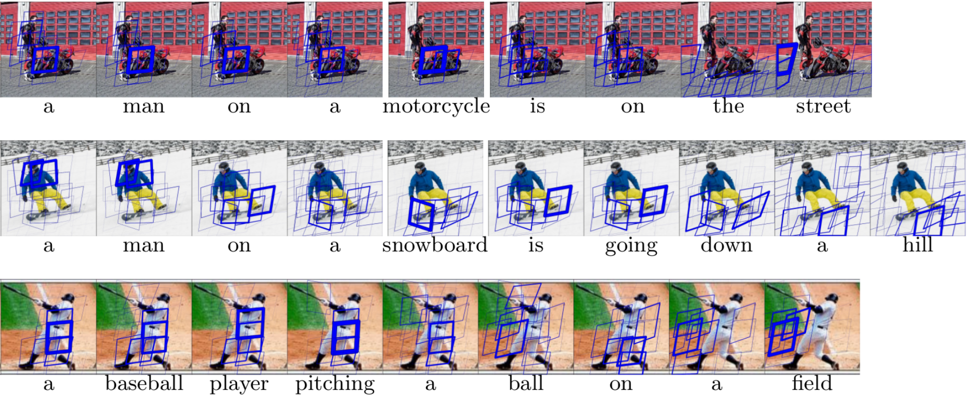
More examples



More examples



More examples



Areas of Attention in Image Captioning

Marco Pedersoli*, Thomas Lucas
Cordelia Schmid, Jakob Verbeek

INRIA Grenoble Rhône-Alpes, France
* Now at École de Technologie Supérieure Montreal, Canada

References I

[Bengio et al., 2015] Bengio, S., Vinyals, O., Jaitly, N., and Shazeer, N. (2015).
Scheduled sampling for sequence prediction with recurrent neural networks.
In *NIPS*.

[Chung et al., 2014] Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014).
Empirical evaluation of gated recurrent neural networks on sequence modeling.
In *NIPS Deep Learning Workshop*.

[Donahue et al., 2015] Donahue, J., Hendricks, L., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., and Darrell, T. (2015).
Long-term recurrent convolutional networks for visual recognition and description.
In *CVPR*.

[Jin et al., 2015] Jin, J., Fu, K., Cui, R., Sha, F., and Zhang, C. (2015).
Aligning where to see and what to tell: image caption with region-based attention and scene factorization.
arXiv:1506.06272.

[Karpathy and Fei-Fei, 2015] Karpathy, A. and Fei-Fei, L. (2015).
Deep visual-semantic alignments for generating image descriptions.
In *CVPR*.

[Kiros et al., 2014] Kiros, R., Salakhutdinov, R., and Zemel, R. (2014).
Multimodal neural language models.
In *ICML*.

[Kiros et al., 2015] Kiros, R., Salakhutdinov, R., and Zemel, R. (2015).
Unifying visual-semantic embeddings with multimodal neural language models.
TACL.
to appear.

[Noh et al., 2015] Noh, H., Hong, S., and Han, B. (2015).
Learning deconvolution network for semantic segmentation.
In *ICCV*.

References II

[Ranzato et al., 2016] Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. (2016).
Sequence level training with recurrent neural networks.
In *ICLR*.

[Simonyan and Zisserman, 2015] Simonyan, K. and Zisserman, A. (2015).
Very deep convolutional networks for large-scale image recognition.
In *ICLR*.

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. (2014).
Sequence to sequence learning with neural networks.
In *NIPS*.

[Vinyals et al., 2015] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015).
Show and tell: A neural image caption generator.
In *CVPR*.

[Xu et al., 2015] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2015).
Show, attend and tell: Neural image caption generation with visual attention.
In *ICML*.

[Yang et al., 2016] Yang, Z., Yuan, Y., Wu, Y., Salakhutdinov, R., and Cohen, W. (2016).
Encode, review, and decode: Reviewer module for caption generation.
In *NIPS*.

[Zitnick and Dollár, 2014] Zitnick, C. and Dollár, P. (2014).
Edge boxes: locating object proposals from edges.
In *ECCV*.