Generating Images from Captions with Attention

Elman Mansimov, Emilio Parisotto, Jimmy Ba and Ruslan Salakhutdinov

UNIVERSITY OF TORONTO

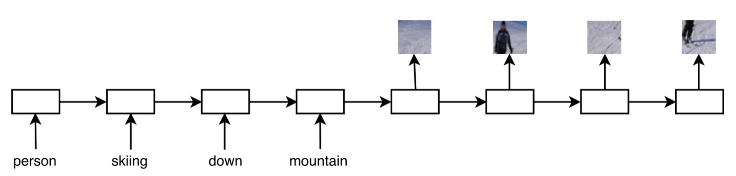
Introduction

Why to condition on captions ?

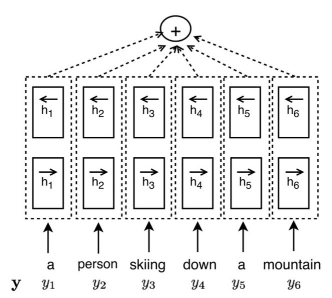
- Captions could be used to simplify image modelling task.
- Generating images conditioned on novel captions helps better understand its generalization.

Key Ideas

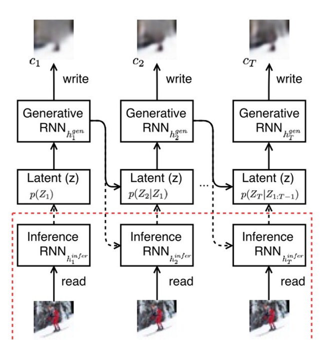
- Treat the problem as part of sequence-to-sequence framework [2,9].



- Caption $\mathbf{y}$ is represented as sequence of words $(y_1, y_2, ..., y_N)$, where N is the length of the sequence.



- Image $\mathbf{x}$ is represented as a sequence of $p \times p$ patches drawn on a $w \times h$ canvas c_t over time $t = 1, ..., T$.



References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[2] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, 2014.

[3] E. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS*, 2015.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

[5] K. Gregor, I. Danihelka, A. Graves, and D. Wierstra. DRAW: A recurrent neural network for image generation. In *ICML*, 2015.

[6] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

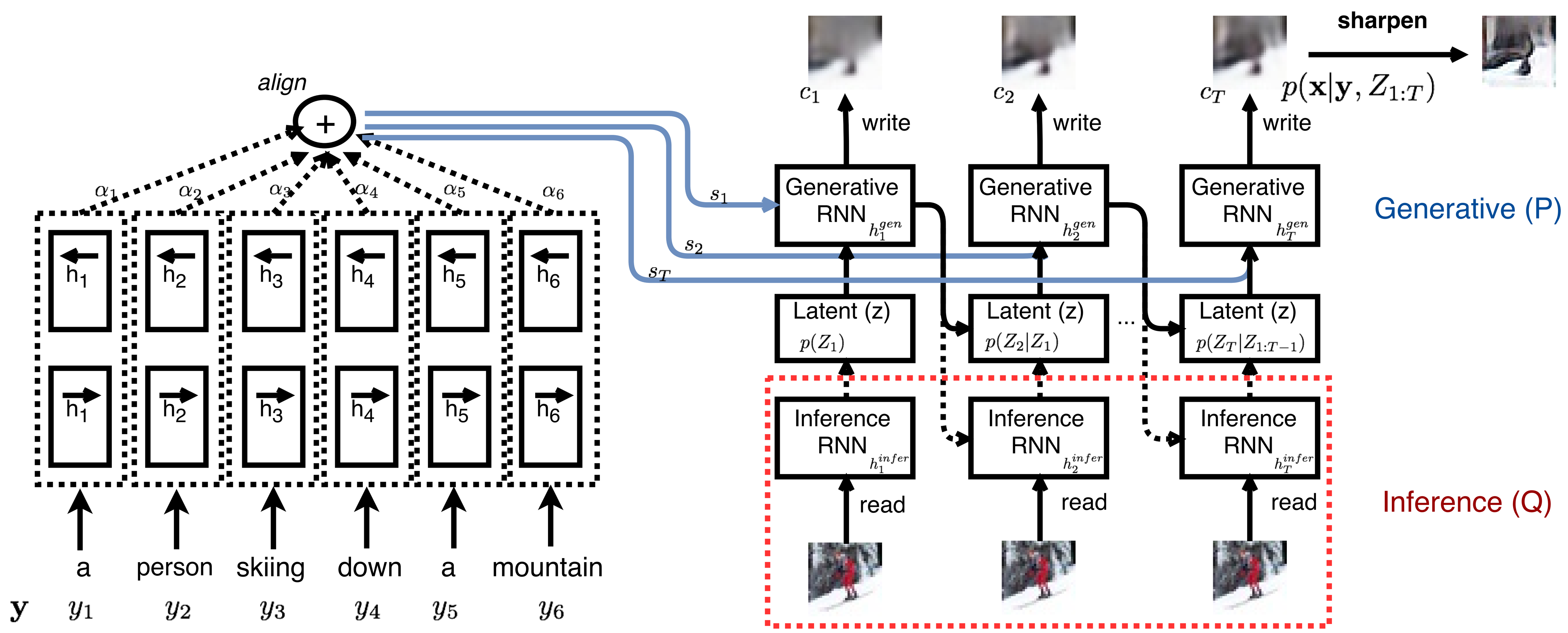
[7] D. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[8] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

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

Model Description

The proposed model consists of four main parts: language model, image model, alignment model and post-processing model.



- Bidirectional LSTM [6] computes a sequence of forward $\overrightarrow{h}^{lang}_{1...N}$ and backward $\overleftarrow{h}^{lang}_{1...N}$ hidden states respectively, which are concatenated together into the sentence representation $h^{lang} = [\overrightarrow{h}^{lang}_{1...N}, \overleftarrow{h}^{lang}_{1...N}]$.
- The image model [5] iteratively computes the following set of equations over time $t = 1, ..., T$:

$$\begin{aligned}
\hat{\mathbf{x}}_t &= \mathbf{x} - \boldsymbol{\sigma}(c_{t-1}),\\
r_t &= read(\mathbf{x}_t, \hat{\mathbf{x}}_t, h^{gen}_{t-1}),\\
h^{infer}_t &= LSTM^{infer}(h^{infer}_{t-1}, [r_t, h^{gen}_{t-1}]),\\
Q(Z_t|\mathbf{x}, \mathbf{y}, Z_{1:t-1}) &= \mathcal{N}\left(\mu(h^{infer}_t), \sigma(h^{infer}_t)\right),
\end{aligned}$$

$$\begin{aligned}
z_t &\sim P(Z_t|Z_{1:t-1}) = \mathcal{N}\left(\mu(h^{gen}_{t-1}), \sigma(h^{gen}_{t-1})\right),\\
s_t &= align(h^{gen}_{t-1}, h^{lang}),\\
h^{gen}_t &= LSTM^{gen}(h^{gen}_{t-1}, [z_t, s_t]),\\
c_t &= c_{t-1} + write(h^{gen}_t),\\
\tilde{\mathbf{x}} &\sim P(\mathbf{x}\,|\,\mathbf{y}, Z_{1:T}) = \prod_i P(x_i\,|\,\mathbf{y}, Z_{1:T}) = \prod_i \mathrm{Bern}(\boldsymbol{\sigma}(c_{T,i})).
\end{aligned}$$

- *align* operator [1] outputs a dynamic sentence representation s_t at each timestep by computing a weighted sum of hidden states of words using alignment probabilities $\alpha^t_{1...N}$:

$$s_t = align(h^{gen}_{t-1}, h^{lang}) = \alpha^t_1 h^{lang}_1 + \alpha^t_2 h^{lang}_2 + ... + \alpha^t_N h^{lang}_N \qquad \alpha^t_k \propto \exp\left(v^\top \tanh(U h^{lang}_k + W h^{gen}_{t-1} + b)\right)$$

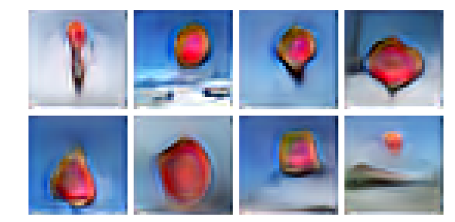
- Images are generated by discarding the inference network and by sampling latent variables $Z_{1:t}$ from prior distribution.
- Finally, images are sharpened using a deterministic adversarial network [3,4] trained on residuals of a Laplacian pyramid.

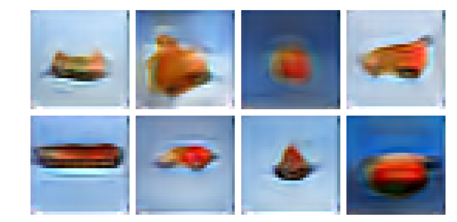
Learning

The model is trained to optimize the variational lower bound $\mathcal{L}$ of image $\mathbf{x}$ given caption $\mathbf{y}$ using the SGVB [7] algorithm.

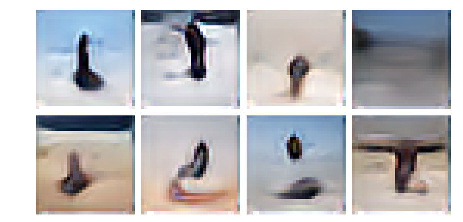
$$\mathcal{L} = \mathbb{E}_{Q(Z_{1:T}\,|\,\mathbf{y},\mathbf{x})}\left[\log p(\mathbf{x}\,|\,\mathbf{y}, Z_{1:T}) - \sum_{t=2}^{T} \mathrm{D_{KL}}\left(Q(Z_t\,|\,Z_{1:t-1}, \mathbf{y}, \mathbf{x}) \,\|\, P(Z_t\,|\,Z_{1:t-1}, \mathbf{y})\right)\right] - \mathrm{D_{KL}}\left(Q(Z_1\,|\,\mathbf{x}) \,\|\, P(Z_1)\right).$$

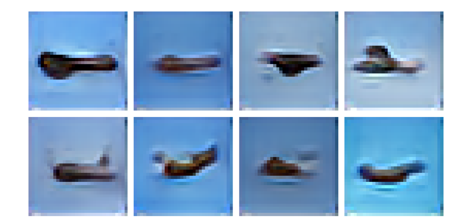
Generated Images

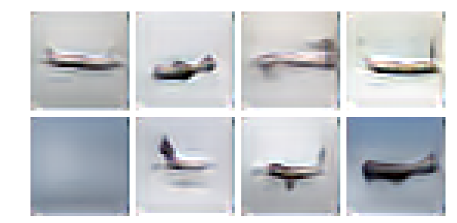

A stop sign is flying in blue skies.

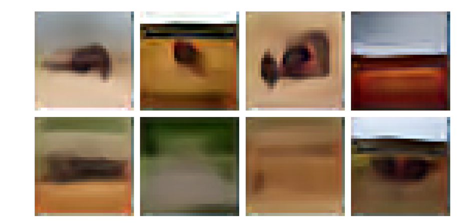

A yellow school bus is flying in blue skies.

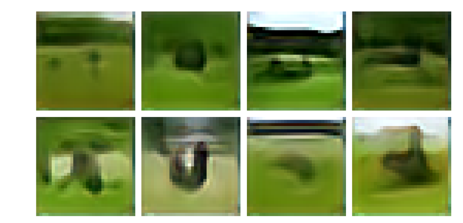

A toilet seat sits open in the grass field.

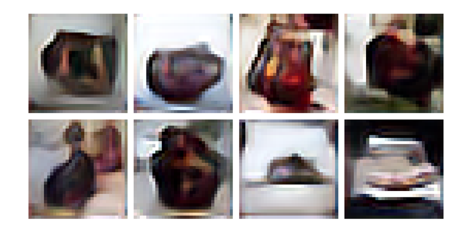

A person skiing on sand clad vast desert.

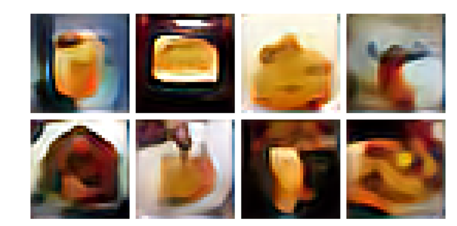

A very large commercial plane flying in blue skies.

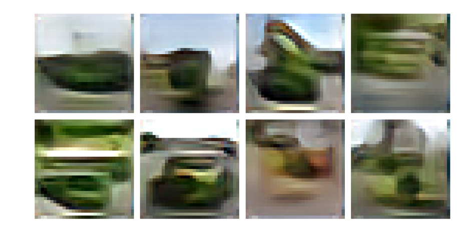

A very large commercial plane flying in rainy skies.

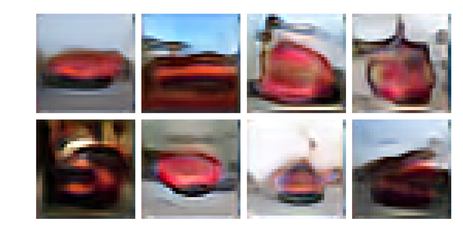

A herd of elephants walking across a dry grass field.


A herd of elephants walking across a green grass field.

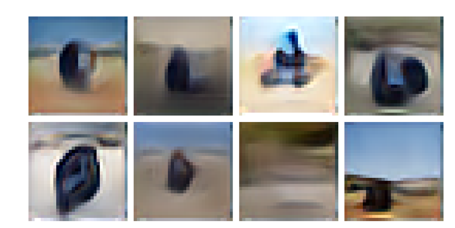

The chocolate desert is on the table.

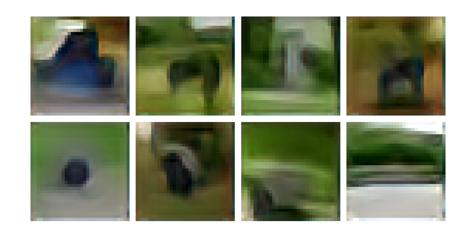

A bowl of bananas is on the table.

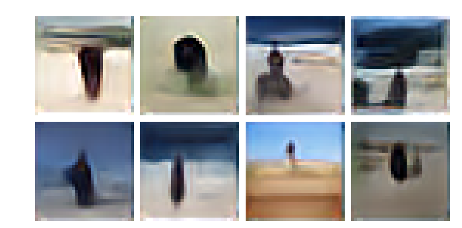

Green bus parked in a parking lot.

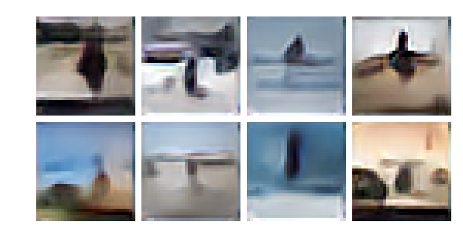

Red bus parked in a parking lot.

Examples of Alignment


A rider on a blue motorcycle in the desert.


A rider on a blue motorcycle in the forest.


A surfer, woman and child walk on beach.


A surfer, woman and child walk on sun.

Microsoft COCO [8] Results

Microsoft COCO (before post-processing)

Model	Retrieval R@1	R@5	R@10	R@50	Med r	Similarity SSI
LAPGAN	-	-	-	-	-	0.08
Fully-Conn VAE	1.0	6.6	12.0	53.4	47	0.156
Conv-Deconv VAE	1.0	6.5	12.0	52.9	48	0.164
skipthoughtDRAW	2.0	11.2	18.9	63.3	36	0.157
noalignDRAW	2.8	14.1	23.1	68.0	31	0.155
alignDRAW	3.0	14.0	22.9	68.5	31	0.156

Top: Image retrieval and similarity results of different models. **R@K** is Recall@K (higher is better). **Med** r is the median rank (lower is better). **SSI** is Structural Similarity Index, which is between -1 and 1 (higher is better).