The University of Texas at Austin
Department of Computer Science
College of Natural Sciences

Hidden State Guidance: Improving Image Captioning Using an Image Conditioned Autoencoder

Jialin Wu, and Raymond J. Mooney The University of Texas at Austin

Introduction

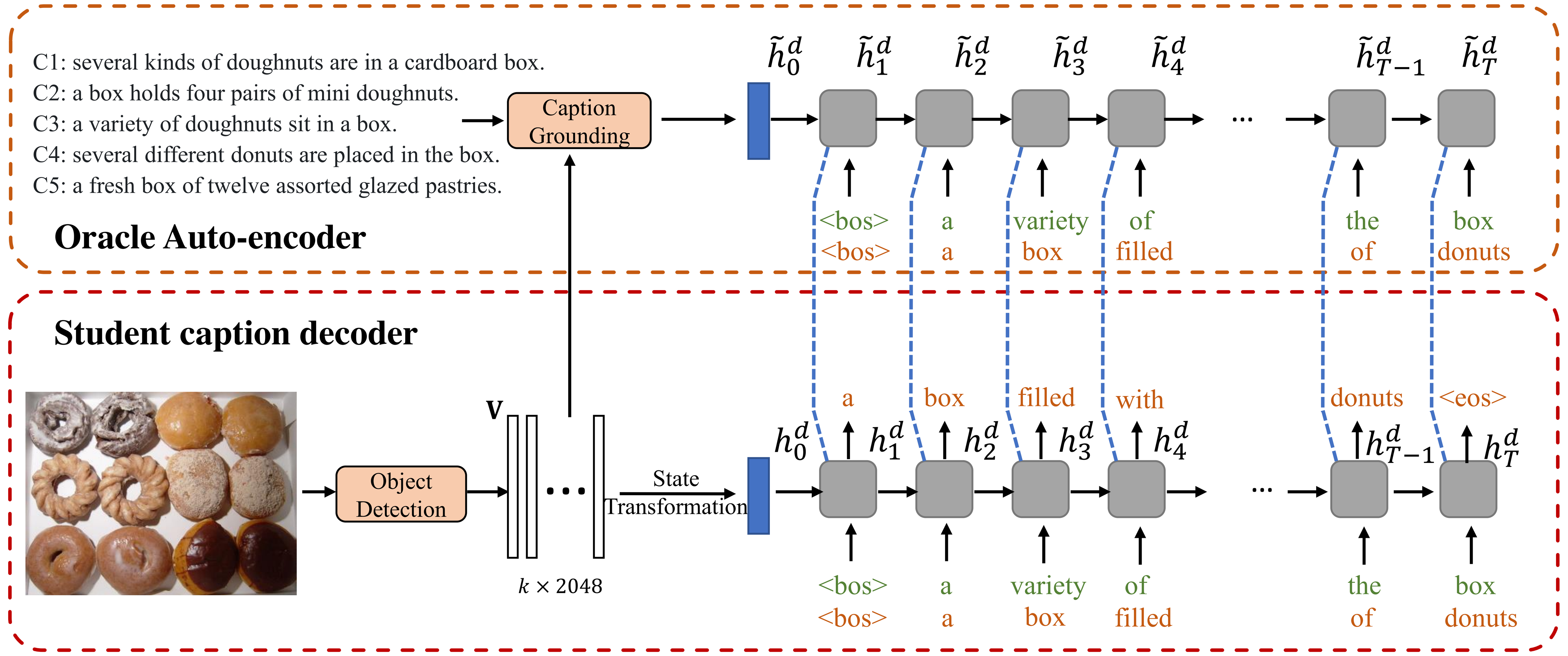
- Most RNN-based image captioning models receive supervision on the output words to mimic human captions.
- During self-critical training, sparse rewards are delayed till the end and equally distributed to each word in the generated caption, regardless of whether or not the words are descriptive.
- We present a new framework, called Hidden State Guidance (HSG), to provides a word-level intermediate reward that highlights the important words.

A girl pitching a baseball with a mitt.
A girl wearing a hat and purple outfit.
A girl is waiting for a ball looking at a baseball.
A woman in blue baseball uniform swinging a glove.
A young girl grabs the suitcase 's glove.

A window topped with pizzas with several toppings.
Lots of pizzas are on the window rack.
Baked trays with pizzas displayed in oven window.
Several pizzas displayed in different varieties in a restaurant.
A box covered in pizza and cheese with other pizzas.

- HSG uses a caption autoencoder as the teacher, whose hidden states encode richer representation, to directly guides the hidden-state learning of the original RNN caption.

Model Overview



- Train teacher network using captions and image as input using teacher forcing
- Train student network to jointly learn to generate image captions and mimic teacher's hidden states using teacher forcing
  - Hidden states loss $\mathcal{L}_{s,t} = \|h_t^d - \tilde{h}_t^d\|_2^2$
- Finetune student using self-critical training (REINFORCE):
  - Additional word-level rewards:

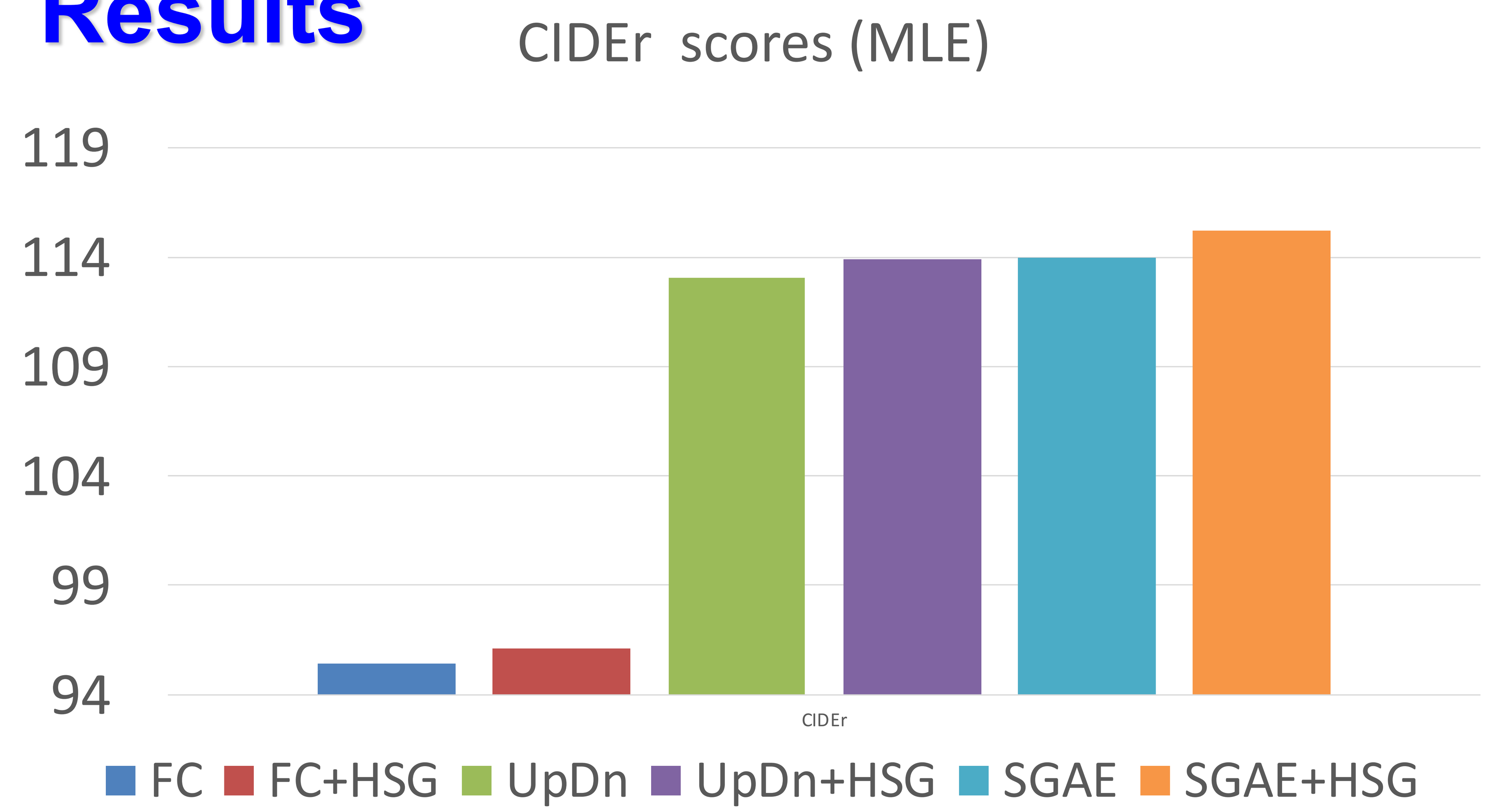
$$\tilde{\mathcal{R}} = -\sum_{t=0}^{T} \mathbb{E}_{\hat{c}_{\leqslant t}\sim p}[\mathcal{L}_{s,t}]$$

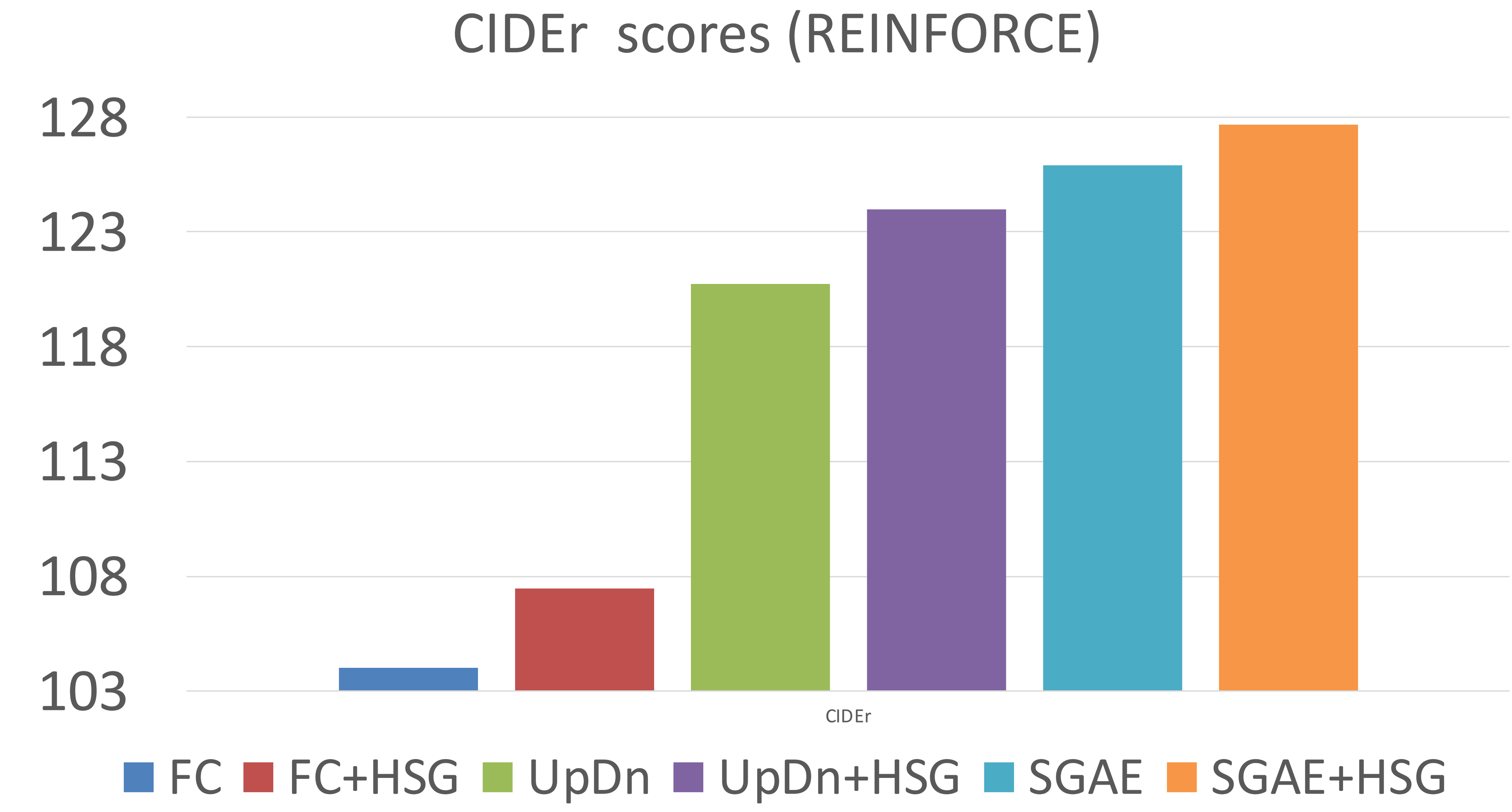
  - Interpretation of the gradients:

$$\nabla_{\theta_g}\tilde{\mathcal{L}} = \nabla_{\theta_g}(\mathcal{L}+\lambda\tilde{\mathcal{R}}) =$$

$$\underbrace{\mathbb{E}_{\hat{c}\sim p}\left[\sum_{\tau=0}^{T}\left(\lambda\sum_{t=\tau}^{T}\mathcal{L}_{s,t} - \tilde{r}(\hat{c})\right)\nabla_{\theta_g}\log p(\hat{c}_\tau|\hat{c}_{<\tau})\right]}_{\text{Reward Term}} + \lambda\,\underbrace{\mathbb{E}_{\hat{c}\sim p}\left[\sum_{\tau=0}^{T}\nabla_{\theta_g}\mathcal{L}_{s,t}\right]}_{\text{Punishing Term}}$$

Results





Conclusion

- Directly supervising hidden states, which provide word-level rewards, is especially helpful during self-critical training