Bidirectional Beam Search: Forward-Backward Inference in Neural Sequence Models for Fill-in-the-Blank Image Captioning - Supplementary

Qing Sun Virginia Tech

Stefan Lee Virginia Tech

Dhruv Batra Georgia Tech

sunging@vt.edu

steflee@vt.edu

dbatra@gatech.edu

BiBS Convergence. As mentioned in the main paper, our proposed algorithm, BiBS, typically converges within in 1 to 2 rounds for the fill-in-the-blank image captioning task. Fig. 1 shows additional qualitative results that demonstrate how the highest ranked sentences change as the BiBS algorithm progress through these meta-iterations.

Unknown Length Blanks for Visual Madlibs. We extend BiBS on the Visual Madlib [1] fill-in-the-blank description generation tasks to the unknown length blanks setting. The basic setting is the same with main paper. We find that BiBS outperforms nearly all baselines on all metrics (narrowly being bested by GSN(ordered) at Blue-2 for type-7).

	type 7		type 12	
	Bleu-1	Bleu-2	Bleu-1	Bleu-2
URNN-f	0.317	0.155	0.285	0.174
URNN-b	0.334	0.184	0.309	0.186
URNN-f+b	0.334	0.181	0.302	0.184
BiRNN-f+b	0.343	0.195	0.291	0.190
GSN [2] (Ordered)	0.348	0.203	0.270	0.184
BiRNN-BiBS	0.351	0.197	0.31	0.190

Table 1: Unknown blank length setting on the Visual Madlibs task using BLEU-1 and BLEU-2. B=5 by default.

References

- [1] L. Yu, E. Park, A. C. Berg, and T. L. Berg, "Visual Madlibs: Fill in the blank Description Generation and Question Answering," ICCV, 2015. 1
- [2] M. Berglund, T. Raiko, M. Honkala, L. Karkkainen, A. Vetek, and J. Karhunen, "Bidirectional recurrent neural networks as generative models," in NIPS, 2015. 1



(GT) Some baseball players are playing a game (Init) Some baseball baseball bat during a game (1st) Some baseball baseball players in a game (2ed) Some baseball players playing in a game



(GT) a woman in a red shirt is holding a blue and orange kite (Init) a woman in a man is holding a blue and orange kite (1st) a woman in a a hat is holding a blue and orange kite (2ed) a woman in a red shirt is holding a blue and orange kite



(GT) this woman is sitting in front of a restaurant smoking a cigarette (Init) this woman is is working on a laptop and smoking a cigarette (1st) this woman is sitting down on a laptop while smoking a cigarette (2ed) this woman is sitting down in a chair and smoking a cigarette



(GT) the skier went down the ramp of fake snow (Init) the skier street filled with lots of fake snow (1st) the skier is covered in the of fake snow (2ed) the skier is sitting in front of fake snow



(GT) a man drinking some wine using a wineglass (Init) a man $\underline{\text{cell phone while holding}}$ a wineglass (1st) a man a smiling while holding a wineglass (2ed) a man is smiling while holding a wineglass



(GT) an adorable little girl holding two ski poles (Init) an adorable riding skis while holding ski poles (1st) an adorable snow skis and two ski poles (2ed) an adorable young girl holding a ski poles



(GT) gourmet pizza cooked and sliced and on a plate (Init) gourmet pizza $\underline{\text{pepperoni pizza is sitting}}$ on a plate (1st) gourmet pizza with pizza and cheese on a plate (2ed) gourmet pizza with pepperoni and cheese on a plate

Figure 1: Performance vs. Iteration. Our model is initialized with right-to-left standard BS (Init) and updated alternatively from left-to-right (1st) and right-to-left (2nd).