IMPROVING CONTEXT MODELLING IN MULTIMODAL DIALOGUE GENERATION

Shubham Agarwal * Ondřej Dušek Ioannis Konstas Verena Rieser
Interaction Lab, Department of Computer Science, Heriot-Watt University, Edinburgh, UK
* Adeptmind Scholar, Adeptmind Inc., Montreal, Canada

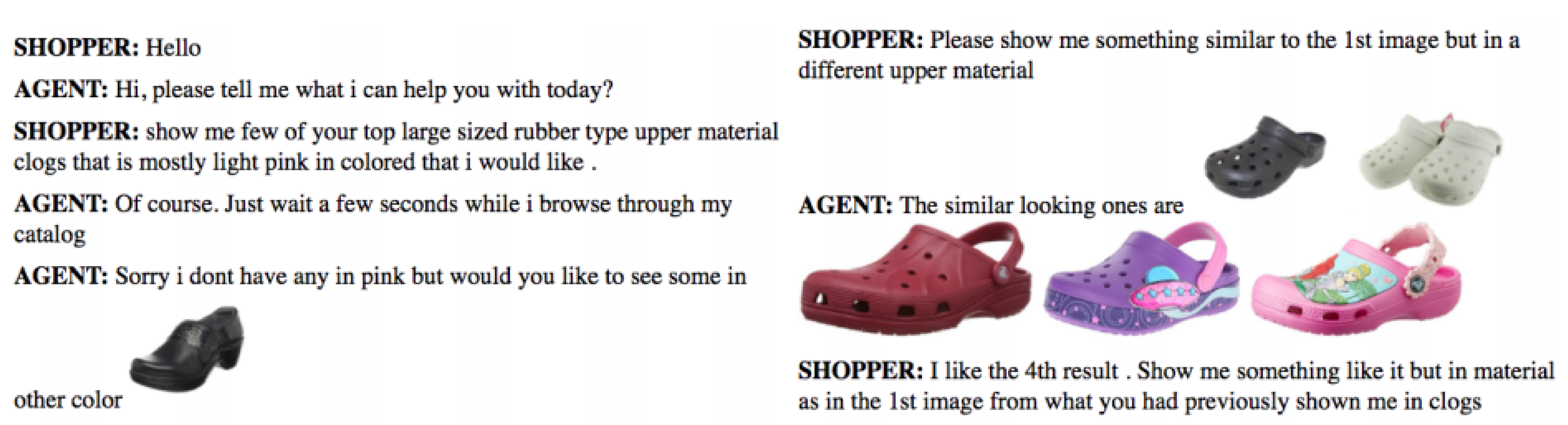
OVERVIEW

Goal: Multimodal dialogue response generation

- Task oriented dialogue system in e-commerce setting
- Based on recently released **MultiModal Dialogue (MMD)** dataset
- **Multimodal HRED with attention** for textual response generation
- Improved context modeling by incorporating multiple images

DATASET

- Raw chatlog of an user-agent interaction in the fashion domain (150k chat sessions with 40 dialogue turns per session)



- Saha et al. 'unroll' multiple images in a single utterance to include only one image per utterance
- Example chatlog and corresponding context for a system response

Our version of the dataset

Text Context: Sorry i don't think i have any 100 % acrylic but i can show you in knit | Show me something similar to the 4th image but with the material different
Image Context: [Img 1, Img 2, Img 3, Img 4, Img 5] | [0, 0, 0, 0, 0]
Target Response: The similar looking ones are

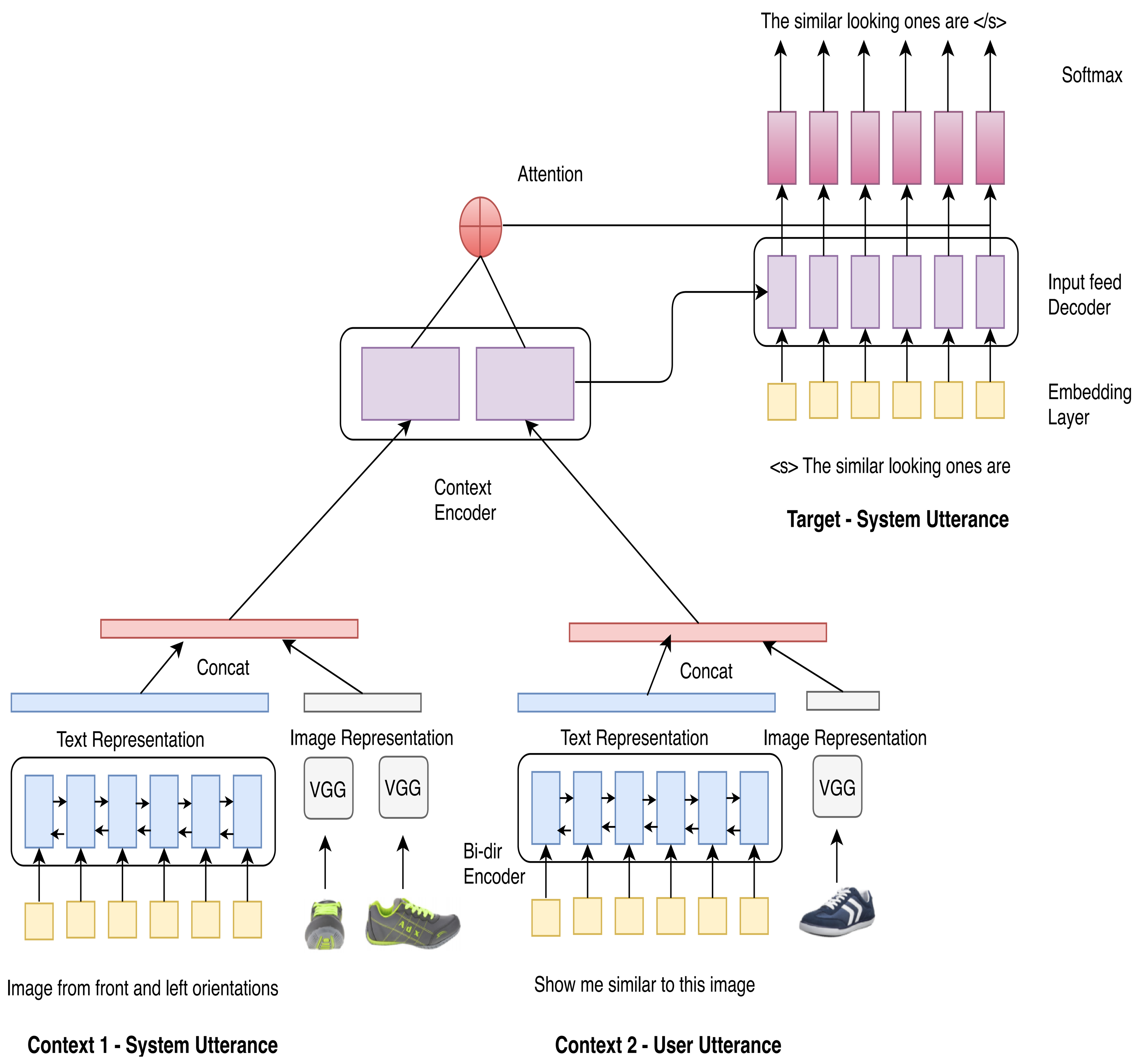
Saha et al.

Text Context: |
Image Context: Img 4 | Img 5
Target Response: The similar looking ones are

EVALUATION AND RESULTS

MODEL

- Multimodal extension to Hierarchical Recurrent Encoder Decoder (HREDs) over multiple images
- Model in encoder-decoder paradigm with four modules
  - Text (Utterance) Encoder
  - Image Encoder
  - Context encoder
  - Input Feeding Decoder



EVALUATION AND RESULTS

Model	Cxt	BLEU-4	METEOR	ROUGE-L
Saha et al. M-HRED	2	0.3767	0.2847	0.6235
T-HRED	2	0.4292	0.3269	0.6692
M-HRED	2	0.4308	0.3288	0.6700
T-HRED–attn	2	0.4331	0.3298	0.6710
M-HRED–attn	2	0.4345	0.3315	0.6712
T-HRED–attn	5	0.4442	0.3374	0.6797
M-HRED–attn	**5**	**0.4451**	**0.3371**	**0.6799**

Table 1: Automatic evaluation based on BLEU-4, METEOR & ROUGE-L

CONCLUSION

- Contrary to Saha et al. , generated outputs improved by adding attention and increasing context size
- Multimodal HRED (M-HRED-attn) *does not* improve significantly over text-only HRED (T-HRED-attn)
- Model learns to handle textual correspondence between the questions and answers, mostly *ignoring* the visual context
- Need better visual models to encode the image representations when we have multiple similar-looking images
- Improvement of **7 BLEU points** over the baseline approach
- Code available at `https://github.com/shubhamagarwal92/mmd`