OVERVIEW

Task: Multimodal search-based dialogue system
- Task oriented dialogue system in e-commerce setting
- Based on recently released MultiModal Dialogue (MMD) dataset
- Multimodal HRED with attention for textual response generation grounded in vision, language and knowledge base input

DATASET

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

KB GROUNDED MULTIMODAL CONVERSATIONAL MODEL

- Hierarchical Encoder
  - Utterance (Text) encoder
  - Image encoder
  - Context encoder

Fig 1: Early fusion of textual and visual representations at the encoder level.

Predictions for different intents

Intent | Model | Example text

show-similar-to

Gold Target: None of the similar looking ones are

show-similar-to

M-HRED–attn–kb

Predicted: The similar looking ones are the blouse in the 1st image has alfani brand

sort-results

Gold Target: Sort these by best seller ranking

sort-results

M-HRED–attn–kb

Predicted: Sorry, do not have the best seller ranking information for these

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

Table 2: BLEU scores for the entire corpus predictions for specific intents

CONCLUSION

- A novel conversational model grounded in language, vision and KB
- Best performing model (M-HRED–attn–kb) outperforms baseline by 9 BLEU points. Grounding in KB gave stark uplift (M-HRED–attn–kb vs. M-HRED–attn)
- Integrating visual context still remains a bottleneck. Need better visual models to encode the image representations for multiple similar-looking images

- Code available at https://github.com/shubhamagarwal92/mmd