In this paper, we proposed a general knowledge-aware multimodal dialog system, which can analyze both textual and visual information, and can respond in time will be rather helpful!

**Motivation**

- There are so many times we hope to find a similar cloth which we see on street or wearing by model. We want to change certain attributes or find a matching item for it.
- Also, when we travel to a foreign place with language barrier, we might still want to explore some new food, tourist sites or shopping venues.
- Under such scenarios, multimodal dialogue system which can analyze both text and image, and can respond in time will be rather helpful!

**Style Tips Domain Knowledge Incorporation**

- With the HRED backbone and multimodal knowledge incorporated, the agent manages to take in multimodal utterances and generate responses turn by turn.
  - The EITree model handles the understanding of product images (w/o possible text descriptions in backend).
  - The multimodal knowledge memory component handles style tips and helps to enrich the systems’ intelligence.

**Results**

- TK captures the taxonomy knowledge thus makes it learn more informative representations for fashion products.
- EK makes it generate responses not only based on the conversation context but also the external knowledge.
- RL fine-tunes the MHRED backbone network and directly optimize the BLEU score or similarity rewards.

**Table 3: Performance of different knowledge components**

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>Diversity (unigram)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MHRED</td>
<td>0.8478</td>
<td>0.004344</td>
</tr>
<tr>
<td>MHRED+TR</td>
<td>0.8573</td>
<td>0.003144</td>
</tr>
<tr>
<td>MHRED+TR-EK</td>
<td>0.8679</td>
<td>0.003144</td>
</tr>
<tr>
<td>MHRED+TR+EK</td>
<td>0.8726</td>
<td>0.003144</td>
</tr>
<tr>
<td>KMD</td>
<td>0.8731</td>
<td>0.003144</td>
</tr>
</tbody>
</table>

**Results**

- We first map the clothing images and text descriptions into a joint visual semantic embedding space via bi-directional ranking loss.
- We then apply the EITree to guide the learning procedure and obtain meaningful representations where each dimension corresponds to a concrete fashion concept.
- Each concept is traced from the root to itself along the EITree and a probability is generated based on the tracing path, which mimics the general to specific recognition procedure.