Figure 1. Illustration of principal objects predicted by the entity module. See Section 1 for more details.
Figure 2. Examples of generated captions using two variants of our model. For comparison convenience, we also present the ground truth and the result of our model.

1. Illustration of Principal Objects

We show four MSR-VTT examples of what our entity module can learn in Figure 1. Since $\tilde{E} = \{\tilde{e}_i\}_{i=1}^N$ are predicted linguistic embeddings of $N$ principal objects, we compute the cosine similarity between $\tilde{e}_i$ and each entity words in vocabulary. The top-5 results are presented.

In Figure 1 (a), $\tilde{e}_0$ captures the entity “woman” in the video, which serves as the subject in the generated caption; $\tilde{e}_1$ and $\tilde{e}_5$ capture the entity “girl” and “child”, which are discarded by other modules when generating the final caption; $\tilde{e}_3$, $\tilde{e}_4$ and $\tilde{e}_5$ capture the “stroller”, which serves as the object in the generated caption. In Figure 1 (b), $\tilde{e}_0$ captures the subject “woman”; $\tilde{e}_3$, $\tilde{e}_4$ and $\tilde{e}_5$ capture the object “dish” on the cutting board; $\tilde{e}_6$ captures the adverbial modifier “kitchen”. Although many other video objects, such as knives, paintings, cutting broad, and gas stove, also appear in the video, our entity module does not adopt them as principal objects. This demonstrates that our proposed entity module has the ability to select those video objects which are likely to be mentioned in captions.

2. Examples of Using Different Supervisions

We show the captions generated by two model variants, i.e., “noun supervision” and “verb supervision”, on eight MSR-VTT videos in Figure 2, where the “noun supervision” replaces the entity with broader nouns to supervise our entity module, and the “verb supervision” replaces the predicate with verb to supervise our predicate module. We also present the generated captions of our model for comparison.

We observe that the captions generated by our model contain richer and more accurate content than the two variants. For instance, in Figure 2 (a), “noun supervision” misses the “motorcycle”, and “verb supervision” generates predicate “playing video game” rather than more accurate “riding motorcycle”. Similarly, in Figure 2 (b), “noun supervision” focuses on the girl’s hair rather than her face. “verb supervision” incorrectly generates “showing her hair”, without realizing that the action is “applying makeup”. Our model outperforms “noun supervision” by generating more accurate entities. This is because abstract nouns in the ground truth, such as “hunger” and “satisfaction”, have no corresponding video objects, and thus introduce noise for generating captions. Meanwhile, our model can predict more accurate video actions than “verb supervision” because using the predicate as supervision can keep the agreement between verbs and verbs’ recipients.