Supplementary for Hierarchical Modular Network for Video Captioning

Ethy Gtanoom	CIII CIII CIII			ci Baly Girow			Peppe				
Object#0	woman 0.992	lady 0.889	female 0.861	feminie 0.704	wife 0.682	Object#0	woman 0.995	lady 0.901	female 0.864	feminie 0.710	wife 0.686
	girl	daughter	female	schoolgirl	girlfriend						
Object#1	0.945	0.792	0.718	0.704	0.680	Object#1	girl 0.926	female 0.810	woman 0.778	lady 0.756	girlfriend 0.752
Object#2	person	human	someone	somebody	personnel	Object#2	person	human	someone	somebody	personnel
0.0]00(#2	0.930	0.835	0.792	0.732	0.721	Object#2	0.878	0.775	0.751	0.718	0.683
Object#3	stroller	baby	pram	infant	cradle	Object#3	cooking	food	dish	dishware	cook
	0.899	0.696	0.691	0.660	0.608		0.786	0.762	0.760	0.731	0.729
Object#4 Object#5 Object#6 Object#7	stroller	pram	baby	infant	toddler	Object#4 Object#5 Object#6 Object#7	cooking	dish	food	dishware	cookware
	0.899	0.701	0.666	0.630	0.589		0.782	0.758	0.752	0.733	0.727
	stroller	baby	infant	pram	toddler		cooking	dish	food	dishware	cookware
	0.886	0.772	0.688	0.717	0.709		0.782	0.758	0.752	0.733	0.727
	child	kid	baby	youngster	infant		kitchen	pot	dish	bowl	cookware
	0.821	0.773	0.737	0.717	0.709		0.657	0.638	0.636	0.635	0.627
	cradle	item	gear	stuff	outfit		container	pot	bowl	dish	item
	0.638	0.614	0.608	0.606	0.598		0.653	0.652	0.645	0.626	0.625
Ground trut Ours: a won	nan is demor	nstrating a str				Ours: a wor	nan in a kitch	ien is talking a (I	about how to) = = = = = = = = = = = = = = = = = = =	prepare a dis	sh =======
				Roote car		Ours: a wor	nan in a kitch	ien is talking a	about how to b)	prepare a dis	sh
Ours: a won	nan is demor			feminie	wife		nan in a kitch	ien is talking a (t male	about how to b) gentleman	prepare a dis	sh
	rooth.com			feminie 0.710	wife 0.681	Ours: a wor		()			
Ours: a won	woman	(a lady	a) Ender female			Object#0	man	(I male	p) gentleman	men	guy
Ours: a won	woman 0.993	lady 0.892	a) female 0.868	0.710	0.681		man 0.992	male 0.765	gentleman 0.752	men 0.690	guy 0.659
Ours: a won Ours: a won Object#0 Object#1	woman 0.993 man	lady 0.892 male	a) female 0.868 gentleman	0.710 men	0.681 guy	Object#0 Object#1	man 0.992 person	male 0.765 human	gentleman 0.752 someone	men 0.690 somebody	guy 0.659 personnel
Ours: a won	woman 0.993 man 0.992	lady 0.892 male 0.776	a) female 0.868 gentleman 0.758	0.710 men 0.715	0.681 guy 0.663	Object#0	man 0.992 person 0.965	male 0.765 human 0.844	gentleman 0.752 someone 0.818	men 0.690 somebody 0.760	guy 0.659 personnel 0.739
Ours: a won Object#0 Object#1 Object#2	woman 0.993 man 0.992 girl	lady 0.892 male 0.776 daughter	female 0.868 gentleman 0.758 female	0.710 men 0.715 schoolgirl	0.681 guy 0.663 girlfriend	Object#0 Object#1 Object#2	man 0.992 person 0.965 car	male 0.765 human 0.844 automobile	gentleman 0.752 someone 0.818 vehicle	men 0.690 somebody 0.760 sedan	guy 0.659 personnel 0.739 suv
Ours: a won Ours: a won Object#0 Object#1	woman 0.993 man 0.992 girl 0.945	lady 0.892 male 0.776 daughter 0.758	female 0.868 gentleman 0.758 female 0.744	0.710 men 0.715 schoolgirl 0.728	0.681 guy 0.663 girlfriend 0.707	Object#0 Object#1	man 0.992 person 0.965 car 0.929	male 0.765 human 0.844 automobile 0.854	gentleman 0.752 someone 0.818 vehicle 0.835	men 0.690 somebody 0.760 sedan 0.767	guy 0.659 personnel 0.739 suv 0.682
Ours: a won Object#0 Object#1 Object#2 Object#3	woman 0.993 man 0.992 girl 0.945 food	lady 0.892 male 0.776 daughter 0.758 meal	female 0.868 gentleman 0.758 female 0.744 cuisine	0.710 men 0.715 schoolgirl 0.728 snack	0.681 guy 0.663 girlfriend 0.707 dish	Object#0 Object#1 Object#2 Object#3	man 0.992 person 0.965 car 0.929 car	male 0.765 human 0.844 automobile 0.854 automobile	gentleman 0.752 someone 0.818 vehicle 0.835 vehicle	men 0.690 somebody 0.760 sedan 0.767 sedan	guy 0.659 personnel 0.739 suv 0.682 coupe
Ours: a won Object#0 Object#1 Object#2	woman 0.993 man 0.992 girl 0.945 food 0.925	lady 0.892 male 0.776 daughter 0.758 meal 0.835	female 0.868 gentleman 0.758 female 0.744 cuisine 0.783	0.710 men 0.715 schoolgirl 0.728 snack 0.699	0.681 guy 0.663 girlfriend 0.707 dish 0.692	Object#0 Object#1 Object#2	man 0.992 person 0.965 car 0.929 car 0.904	male 0.765 human 0.844 automobile 0.854 automobile 0.833	gentleman 0.752 someone 0.818 vehicle 0.835 vehicle 0.823	men 0.690 somebody 0.760 sedan 0.767 sedan 0.749	guy 0.659 personnel 0.739 suv 0.682 coupe 0.686
Ours: a won Object#0 Object#1 Object#2 Object#3 Object#4	woman 0.993 man 0.992 girl 0.945 food 0.925 food	lady 0.892 male 0.776 daughter 0.758 meal 0.835 meal	female 0.868 gentleman 0.758 female 0.744 cuisine 0.783 cuisine	0.710 men 0.715 schoolgirl 0.728 snack 0.699 snack	0.681 guy 0.663 girlfriend 0.707 dish 0.692 lunch	Object#0 Object#1 Object#2 Object#3 Object#4	man 0.992 person 0.965 car 0.929 car 0.904 car	male 0.765 human 0.844 automobile 0.854 automobile 0.833 vehicle	gentleman 0.752 someone 0.818 vehicle 0.835 vehicle 0.823 automobile	men 0.690 somebody 0.760 sedan 0.767 sedan 0.749 sedan	guy 0.659 personnel 0.739 suv 0.682 coupe 0.686 coupe
Ours: a won Object#0 Object#1 Object#2 Object#3	woman 0.993 man 0.992 girl 0.945 food 0.925 food 0.918	lady 0.892 male 0.776 daughter 0.758 meal 0.835 meal 0.835 meal 0.816	female 0.868 gentleman 0.758 female 0.744 cuisine 0.783 cuisine 0.767	0.710 men 0.715 schoolgirl 0.728 snack 0.699 snack 0.675	0.681 guy 0.663 girlfriend 0.707 dish 0.692 lunch 0.666	Object#0 Object#1 Object#2 Object#3	man 0.992 person 0.965 car 0.929 car 0.904 car 0.879	male 0.765 human 0.844 automobile 0.854 automobile 0.833 vehicle 0.813	gentleman 0.752 someone 0.818 vehicle 0.835 vehicle 0.823 automobile 0.811	men 0.690 somebody 0.760 sedan 0.767 sedan 0.749 sedan 0.729	guy 0.659 personnel 0.739 suv 0.682 coupe 0.686 coupe 0.695
Ours: a won Object#0 Object#1 Object#2 Object#3 Object#4 Object#5	woman 0.993 man 0.992 girl 0.945 food 0.925 food 0.918 person	lady 0.892 male 0.776 daughter 0.758 meal 0.835 meal 0.835 meal 0.836 human	female 0.868 gentleman 0.758 female 0.744 cuisine 0.783 cuisine 0.767 someone	0.710 men 0.715 schoolgirl 0.728 snack 0.699 snack 0.675 somebody	0.681 guy 0.663 girlfriend 0.707 dish 0.692 lunch 0.666 people	Object#0 Object#1 Object#2 Object#3 Object#4 Object#5	man 0.992 person 0.965 car 0.929 car 0.904 car 0.879 car	male 0.765 human 0.844 automobile 0.854 automobile 0.833 vehicle 0.813 vehicle	gentleman 0.752 someone 0.818 vehicle 0.835 vehicle 0.823 automobile 0.811 automobile	men 0.690 somebody 0.760 sedan 0.767 sedan 0.749 sedan 0.729 coupe	guy 0.659 personnel 0.739 suv 0.682 coupe 0.686 coupe 0.695 sedan
Ours: a won Object#0 Object#1 Object#2 Object#3 Object#4	woman 0.993 man 0.992 girl 0.945 food 0.925 food 0.918 person 0.880	lady 0.892 male 0.776 daughter 0.758 meal 0.835 meal 0.835 meal 0.816 human 0.818	female 0.868 gentleman 0.758 female 0.744 cuisine 0.783 cuisine 0.767 someone 0.754	0.710 men 0.715 schoolgirl 0.728 snack 0.699 snack 0.675 somebody 0.723	0.681 guy 0.663 girlfriend 0.707 dish 0.692 lunch 0.666 people 0.707	Object#0 Object#1 Object#2 Object#3 Object#4	man 0.992 person 0.965 car 0.929 car 0.904 car 0.879 car 0.879 car 0.788	male 0.765 human 0.844 automobile 0.854 automobile 0.833 vehicle 0.813 vehicle 0.764	gentleman 0.752 someone 0.818 vehicle 0.835 vehicle 0.823 automobile 0.811 automobile 0.736	men 0.690 somebody 0.760 sedan 0.767 sedan 0.749 sedan 0.729 coupe 0.690	guy 0.659 personnel 0.739 suv 0.682 coupe 0.686 coupe 0.695 sedan 0.672
Ours: a won Object#0 Object#1 Object#2 Object#3 Object#4 Object#5 Object#6	woman 0.993 man 0.992 girl 0.945 food 0.925 food 0.918 person 0.880 food	lady 0.892 male 0.776 daughter 0.758 meal 0.835 meal 0.835 meal 0.816 human 0.818 meal	female 0.868 gentleman 0.758 female 0.744 cuisine 0.783 cuisine 0.767 someone 0.754 cuisine	0.710 men 0.715 schoolgirl 0.728 snack 0.699 snack 0.675 somebody 0.723 snack	0.681 guy 0.663 girlfriend 0.707 dish 0.692 lunch 0.666 people 0.707 lunch	Object#0 Object#1 Object#2 Object#3 Object#4 Object#5 Object#6	man 0.992 person 0.965 car 0.929 car 0.904 car 0.879 car 0.879 car 0.788 person	male 0.765 human 0.844 automobile 0.854 automobile 0.833 vehicle 0.813 vehicle 0.764 woman	gentleman 0.752 someone 0.818 vehicle 0.835 vehicle 0.823 automobile 0.811 automobile 0.736 lady	men 0.690 somebody 0.760 sedan 0.767 sedan 0.749 sedan 0.729 coupe 0.690 human	guy 0.659 personnel 0.739 suv 0.682 coupe 0.686 coupe 0.686 coupe 0.695 sedan 0.672 female
Ours: a won Object#0 Object#1 Object#2 Object#3 Object#4 Object#5	woman 0.993 man 0.992 girl 0.945 food 0.925 food 0.925 food 0.918 person 0.880 food 0.880	lady 0.892 male 0.776 daughter 0.758 meal 0.835 meal 0.835 meal 0.816 human 0.818 meal 0.818	female 0.868 gentleman 0.758 female 0.744 cuisine 0.783 cuisine 0.767 someone 0.754 cuisine 0.754 cuisine	0.710 men 0.715 schoolgirl 0.728 snack 0.699 snack 0.675 somebody 0.723 snack 0.671	0.681 guy 0.663 girlfriend 0.707 dish 0.692 lunch 0.666 people 0.707 lunch 0.667	Object#0 Object#1 Object#2 Object#3 Object#4 Object#5	man 0.992 person 0.965 car 0.929 car 0.904 car 0.879 car 0.879 car 0.788 person 0.764	male 0.765 human 0.844 automobile 0.854 automobile 0.833 vehicle 0.813 vehicle 0.764 woman 0.760	gentleman 0.752 someone 0.818 vehicle 0.835 vehicle 0.823 automobile 0.811 automobile 0.736 lady 0.749	men 0.690 somebody 0.760 sedan 0.767 sedan 0.749 sedan 0.729 coupe 0.690 human 0.721	guy 0.659 personnel 0.739 suv 0.682 coupe 0.686 coupe 0.685 sedan 0.672 female 0.694
Ours: a won Object#0 Object#1 Object#2 Object#3 Object#4 Object#5 Object#7 Ground truth	woman 0.993 man 0.992 girl 0.945 food 0.925 food 0.918 person 0.880 food 0.867 item 0.630 : a man and	lady 0.892 male 0.776 daughter 0.758 meal 0.835 meal 0.816 human 0.818 meal 0.818 human 0.818	a) female 0.868 gentleman 0.758 female 0.744 cuisine 0.783 cuisine 0.767 someone 0.754 cuisine 0.754 cuisine 0.754 cuisine 0.754 cuisine 0.754 cuisine 0.756 stuff 0.622 ng food	0.710 men 0.715 schoolgirl 0.728 snack 0.699 snack 0.675 somebody 0.723 snack 0.671 meal	0.681 guy 0.663 girlfriend 0.707 dish 0.692 lunch 0.666 people 0.707 lunch 0.667 dish	Object#0 Object#1 Object#2 Object#3 Object#4 Object#5 Object#6 Object#7 Ground truth	man 0.992 person 0.965 car 0.929 car 0.904 car 0.879 car 0.879 car 0.788 person 0.764 gear	male 0.765 human 0.844 automobile 0.854 automobile 0.833 vehicle 0.813 vehicle 0.764 woman 0.760 equipment 0.676 fixing a car	gentleman 0.752 someone 0.818 vehicle 0.835 vehicle 0.823 automobile 0.821 automobile 0.736 lady 0.749 vehicle	men 0.690 somebody 0.760 sedan 0.767 sedan 0.749 sedan 0.729 coupe 0.690 human 0.721 apparatus	guy 0.659 personnel 0.739 suv 0.682 coupe 0.686 coupe 0.695 sedan 0.672 female 0.694 coupe

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 $\overline{\mathcal{E}} = {\{\overline{e}_i\}_{i=1}^N}$ are predicted linguistic embeddings of N principal objects, we compute the cosine similarity between \overline{e}_i and each *entity* words in vocabulary. The top-5 results are presented.

In Figure 1 (a), \bar{e}_0 captures the entity "woman" in the video, which serves as the *subject* in the generated caption; \bar{e}_1 and \bar{e}_6 capture the entity "girl" and "child", which are discarded by other modules when generating the final caption; \bar{e}_3 , \bar{e}_4 and \bar{e}_5 capture the "stroller", which serves as the *object* in the generated caption. In Figure 1 (b), \bar{e}_0 captures the *subject* "woman"; \bar{e}_3 , \bar{e}_4 and \bar{e}_5 capture the object "dish" on the cutting broad; \bar{e}_6 captures the adverbial modifier "kitchen". Although many other video objects, such as knifes, 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.