A Implementation Details

All our models are trained until convergence using early stopping with a patience of 20 epochs and a maximum of 50 epochs. We use the Adamax optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.002, which we decay with a factor of 0.8 after 8 epochs without improvements on the validation set. Dropout regularisation with a probability of 50% is applied on word embeddings and on the hidden state of the second LSTM layer $h_2^{(t)}$ before it is projected to compute the next word probabilities. We use a beam size of 5 during evaluation. All hidden layers and embedding sizes are set to 1024. Models are all trained on a single 12GB NVIDIA GPU.

We use a fixed number of $n = 36$ objects extracted with our pretrained Faster R-CNN. The number of objects and relations extracted with the pretrained Iterative Message Passing model varies according to the input image, i.e. a maximum of $o = 100$ objects and of $r = 2500$ relations .

B Using Ground-Truth Graphs

	SPICE			Captioning		
	All	Obj	Rel	B4	C	R
FA	18.3	34.3	**5.3**	30.5	95.8	53.1
HA-IM	**19.0**	**35.2**	4.8	<u>33.5</u>	**106.0**	**54.8**
HA-SG	18.8	34.7	4.9	32.9	<u>104.5</u>	54.3
+ GAT	18.6	34.7	<u>5.2</u>	32.5	100.9	53.9
+ C-GAT	<u>18.9</u>	<u>34.8</u>	5.1	**33.6**	104.3	<u>54.5</u>

Table 4: Results for the full VG-COCO validation set using features extracted for ground-truth scene graphs. Models and acronyms are described in Sections 5. Metrics reported are: the overall SPICE F1 score (All) and it object (Obj) and relation (Rel) F1 score components, BLEU-4 (B4), CIDEr (C), and ROUGE-L (R). We bold-face the best and underline the second-best overall scores per metric.

In this small-scale experiment, we generate features for ground-truth scene graphs to determine if more a positive transfer can be achieved on image captioning models. For the VG-COCO dataset, we take all the ground-truth object and relation boxes and pass these through the pretrained Iterative Message Passing model, instead of the RPN and RelPN proposed boxes.

Wang et al. (2019) also did a similar experiment, however, they also trained their models using features from gold-standard scene graphs, whereas we only use them to evaluate models previously trained on predicted scene graph features. In Table 4 we show that when using ground-truth scene graphs results are worse than those obtained using predicted ones (Table 3a). One obvious explanation is the mismatch between training and testing data, with regards to quality and number of features. Models are trained on predicted scene graphs, which have an average of 34 object and 48 relation features per image (probably noisy, as seen in Section 5.4), whereas ground-truth graphs have an average of 21 objects and 18 relations per image.