Supplemental Material – SemStyle: Learning to Generate Stylised Image Captions using Unaligned Text

Contents

1 Baseline Implementation Details

Our evaluations includes 5 state-of-the-art baselines.

CNN+RNN-coco is based on the Show+Tell model [11] and trained on only the MSCOCO dataset. We use a GRU cell in place of an LSTM cell for a fairer comparison with our model. In fact, this baseline is just the *term generator* component of SemStyle trained to output full sentences rather than sequences of terms. All hyper-parameter settings are the same as for the *term generator*.

TermRetrieval uses the *term generator* to generate a list of terms – in this case the term vocabulary is words rather than lemmas with POS tags. These terms are used in an OR query of the Romance text corpus and scored with BM25 [4] using hyper-parameters $b = 0.75, k_1 = 1.2$. Our query engine is Whoosh[1], which includes a tokenizer, lower-case filter, and porter stem filter. This model cannot generate caption that are not part of the romance text corpus and the same set of terms always gives the same sentence – ie it is deterministic and only dependent on terms.

StyleNet is our re-implemented of the method proposed by Gan et al. [2] – the original code was not released at the time of writing. We train it on the MSCOCO dataset and the Romantic text dataset. Our implementation follows Gan et al. [2] with the following implementation choices to ensure a fair comparison with other baselines. Rather than ResNet152 [3] features we use Inception-v3 [10] features and a batch size of 128 for both datasets. When training on styled text *StyleNet* requires random input noise from some unspecified distribution, we tried a few variations and found Gaussian noise with $\mu = 0$ and $\sigma = 0.01$ worked reasonably well. Gan et al.suggested a training scheme where the training set alternates between descriptive and styled at the end of every epoch. We found this fails to converge, perhaps because our datasets are larger and more diverse compared with the *FlickrStyle10k* dataset used in the original implementation. *FlickrStyle10k*, which is not publicly released at the time of writing, contains styled captions rather than sentences sampled from novels. To ensure *StyleNet* converges on our dataset we alternate between the MSCOCO dataset and Romantic text dataset after every mini-batch – a strategy suggested by Luong et al. [8] for multi-task sequence-to-sequence learning.

neural-storyteller consists of pre-trained models released by Kiros [5] for generating styled image captions – see Figure 1. This model, first retrieves descriptive captions using an multi-modal space [6] trained on MSCOCO with a VGG-19 [9] CNN image encoder and a GRU caption encoder. Retrieved captions are encoded into skip-thought vectors [7], averaged, and then style shifted. This style shift is performed by subtracting off the mean skip-though vector for captions and adding the mean skip-thought vector of text in the target style. The style shifted vector is decoded by a conditional RNN language model trained on text in the target style. The skip-though vectors are trained on the entirety of bookcorpus [12], while the skip-thought vector decoder is trained on the romance genre subset of bookcorpus (the same subset we have used for our models). *neural-storyteller* generates passages by repeatedly sampling the decoder, we use only the first sentence because long passages would be disadvantaged by the evaluation criteria.

JointEmbedding, shown in Figure 2, uses a learnt multi-modal vector space as the intermediate representation. The image embedder is a projection of pre-trained Inception-v3 [10] features

[1] `https://pypi.python.org/pypi/Whoosh/`

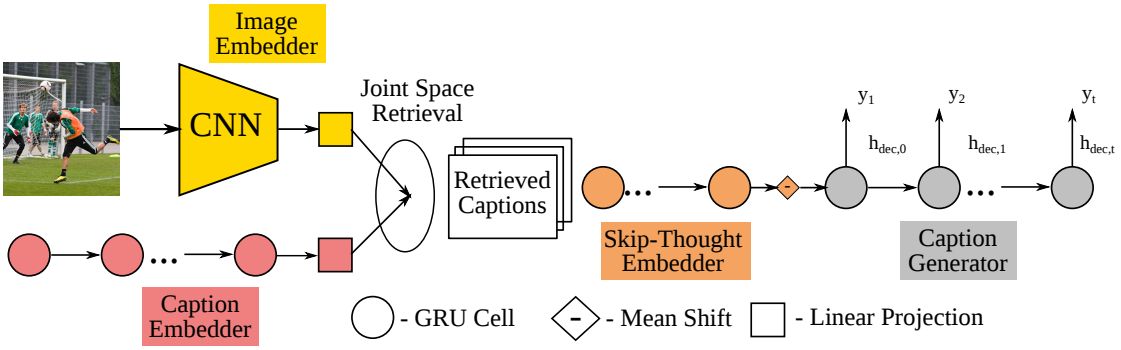

Figure 1: The neural-storyteller model [5], for generating short styled stories about images. The mean shift block subtracts off the mean skip-though vector for captions and adds on the mean skip-thought vector for the target style.

h_I, while the *sentence embedder* is a projection of the last hidden state of an RNN with GRU units h_{enc}. Formally the projections are:

$$\begin{aligned} v_I &= \tanh(W_I.h_I) \\ v_s &= \tanh(W_s.h_{enc}) \end{aligned}$$

Denoting the projections as, v_I for images and v_s for captions, and the learnt projection weights as W_I for images and W_s for captions. Agreement between image and caption embedding is defined as the cosine similarity:

$$g(v_I, v_s) = \frac{v_I.v_s}{|v_I||v_s|}$$

To construct the space we use a noise contrastive pair-wise ranking loss suggested by Kiros et al [6]. Intuitively, this loss function encourages greater similarity between embeddings for paired image-captions than for un-paired images and captions.

$$\mathcal{L} = \max(0, m - g(v_I, v_s) + g(v_{I'}, v_s)) + \max(0, m - g(v_I, v_s) + g(v_I, v_{s'}))$$

Where s is the input caption pared with image I, while s' is a randomly sampled noise contrastive caption and I' the noise contrastive image. The margin m is fixed to 0.1 in our experiments.

The *sentence generator* is an RNN with GRU units, that decodes from the joint vector space. The loss function is categorical cross entropy given in Equation 1.

$$\mathcal{L} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j \in V^m} \log p(y_i = j | I, y_{i-1} ... y_1)^{\mathbb{I}[y_i = j]} \tag{1}$$

Training is a two stage process, first we define the joint space by learning the image embedder and the *sentence embedder* on MSCOCO caption-image pairs. From here on the parameters of image embedder and the *sentence embedder* are fixed. The *sentence generator* is learnt separately by embedding styled sentences from the romantic novel dataset with the *sentence embedder* into the multi-modal space and then attempting to recover the original sentence. This model has not been published previously, but is based on existing techniques for descriptive captioning [6].

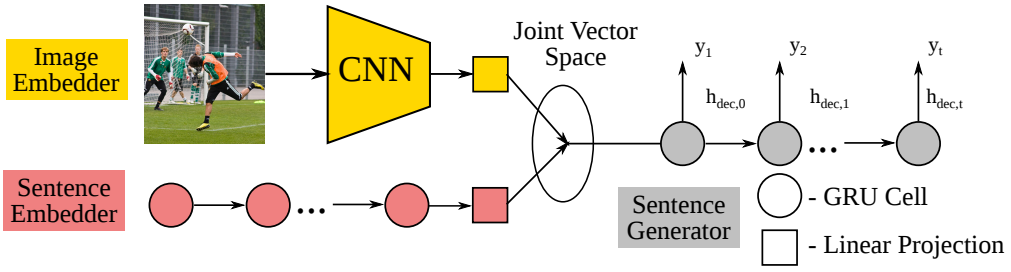

Figure 2: An overview of the *JointEmbedding* model. The two embedding components image embedder (in yellow) and *sentence embedder* (in red) are shown on the left while the *sentence generator* (in grey) is on the right.

2 Model Variants

Our full model is denoted *SemStyle*. We use the following variants to assess several modelling choices.

SemStyle-coco is the *SemStyle* model trained jointly on MSCOCO and the romance corpus with dataset indicator set to MSCOCO at test time. The output of this model should be purely descriptive.

SemStyle-cocoonly is the *SemStyle* model trained only on MSCOCO. The output of this model should be purely descriptive.

SemStyle-unordered is a variant of *SemStyle* with a randomised semantic term ordering. This model helps us to quantify the effect of ordering in the term space.

SemStyle-words is a variant where the semantic terms are raw words – they are not POS tagged, lemmatized or mapped to FrameNet frames.

SemStyle-lempos is a variant where the semantic terms are lemmatized and POS tagged, but verbs are not mapped to FrameNet frames. This helps us to quantify the degree to which verb abstraction effects the model performance.

SemStyle-romonly is *SemStyle* without joint training – the language generator was trained only on the romantic novel dataset. This model helps to quantify the effect of joint training.

3 Human Evaluation

3.1 Crowd-sourcing Task Setup

We performed two human evaluation tasks using the CrowdFlower[2] platform. The first was a relevance task, asking how well a caption describes an image on a four point scale. We provide screen-shots of the instructions given to workers, Figure 3, and an example question, Figure 4. The second task evaluates conformity to the romantic novel style, by asking workers if the caption is from a story about the image, from someone trying to describe the image or completely unrelated to the image. We provide screen-shots of the instructions given to workers, Figure 5, and an example question, Figure 6.

[2] `https://www.crowdflower.com`

Overview

Help us decide how well image captions relate to each image.

Steps

1. Examine the image.
2. Decide how well the caption relates to the image
3. Select the appropriate option

Rules & Tips

Rules:

- There are four possible choices for level of descriptiveness:
  - "Completely unrelated": the caption does not describe the image at all, nor does it have any of the right words for describing the image. This also includes captions that are not specific to any image.
  - "A few of the right words": the caption has some of the right words but they may be in the wrong order or used in a way that doesn't relate to the image
  - "Almost there, a few mistakes": has some of the main objects and/or actions, clearly relates to the image but has some mistakes such as: confusing male and female, using the wrong colors or adjectives, using the wrong action or missing some of the contents.
  - "A clear and accurate caption perhaps with extra non-visual information.": a caption related to this image that may have additional non-visual or contextual information. It doesn't have to describe everything only the main object/s and or actions.

Tips:

- "Completely unrelated" also refers to captions that would work with almost any image. (ie they are not specific)
- "A clear and accurate caption" does not need to use perfect grammar but should be understandable and clear.

Figure 3: A screen-shot of the instructions provided to workers evaluating the relevance of a caption to an image.

I had a glass vase filled with flowers, using it to block out all of it.

How well does this caption relate to the image? (required)

- Completely unrelated
- A few of the right words
- Almost there, a few mistakes.
- A clear and accurate caption, perhaps with extra non-visual information.

Figure 4: A screen-shot of a single question asked of workers in the relevance evaluation task.

Overview

Help us decide which sentences related to images could come from a story about the image or are more likely to be only a description of the image.

Steps

1. Examine the image.
2. Decide if the sentence is related to the image
3. If it is, then decide if it came from a story or from someone trying to describe the image contents to you.

Rules & Tips

Rules:

- The sentence **does not** have to describe the image perfectly, but should relate to the image
- The sentence **does not** have to use perfect grammar
- If the sentence is **completely unrelated** to the image then select "The sentence is completely unrelated to the image."

Tips:

- Stories may use the first person eg "I went to the store", while a description would not.
- Stories often use more colorful and and emotive language eg "The tranquil lake shimmered in the dawn light."
- Stories might refer to state of mind eg "I thought about eating the donut."
- Descriptions tend to be in present tense, relatively short and direct eg "A dog on some grass", "The pizza is sitting on a table"

Figure 5: A screen-shot of the instructions provided to workers evaluating how well a caption conforms to the desired style.

I had a glass vase filled with flowers, using it to block out all of it.

Is this sentence from a story about the image or from someone trying to describe the image to you? (required)

- Story
- Description
- The sentence is completely unrelated to the image.

Figure 6: A screen-shot of a single question asked of workers in the style evaluation task.

Model	*BLEU-1*	*BLEU-4*	*METEOR*	*CIDEr*	*SPICE*	*CLF*	*LM*	*GRU LM*
CNN+RNN-coco	0.667	0.238	0.224	0.772	0.154	0.001	6.591	6.270
StyleNet-coco	0.643	0.212	0.205	0.664	0.135	0.0	6.349	5.977
SemStyle-cocoonly	0.651	0.235	0.218	0.764	0.159	0.002	6.876	6.507
SemStyle-coco	0.653	0.238	0.219	0.769	0.157	0.003	6.905	6.691

Table 1: Evaluating caption descriptiveness on MSCOCO dataset. For details of metrics see the main text for details of methods see Section 1.

3.2 Crowd-sourcing Quality Control and Rating Aggregation

To ensure reliable results and avoid workers who choose randomly CrowdFlower injects questions with known ground truth into each task, requiring workers to achieve at least 70% accuracy on these questions. We manually labelled a small selection of questions which were judged to be clear exemplars. On a limited number of our ground truth questions, workers consistently made mistakes. We revised or removed these question from the ground-truth. The ground truth was expanded by adding selecting questions to which all three annotators agreed on the answer. This is the method suggested by the CrowdFlower documentation for running large evaluations, because additional ground truth speeds up evaluation as workers may complete more tasks (ground truth is never re-used for the same worker and so acts as a limit on the number of tasks they can complete).

Each image-caption pair is seen by $n \geq 3$ workers. Where $n = 3$ in most cases, typically being greater than 3 when workers have successfully challenged the original ground truth. We aggregate these judgements by assigning each one a weight $1/n$, and calculating the weight normalised sum for each possible answer. The resulting scores are displayed in Figure 3 of the main text. In the case of descriptiveness judgements a further summary statistic is calculated as the average descriptiveness score in the range 1-4.

4 Results

4.1 BLEU, METEOR, and CIDEr for styled captions

Table 1 and Table 2 provide additional automatic results, include BLEU, METEOR, and CIDEr scores – as measured on the MSCOCO results. As we note in the main text these n-gram based measures are less relevant in the style generation case, but are provided here for completeness.

4.2 Tabular Details for Human Evaluation

Table 3 and Table 4 give the full results for the human evaluation tasks. In the main text these are presented in graphical form, for completeness the full numerical results are given here.

Model	*BLEU-1*	*BLEU-4*	*METEOR*	*CIDEr*	*SPICE*	*CLF*	*LM*	*GRU LM*
StyleNet	0.272	0.099	0.064	0.009	0.010	0.415	7.487	6.830
TermRetrieval	0.322	0.037	0.120	0.213	0.088	0.945	3.758	4.438
neural-storyteller	0.265	0.015	0.107	0.089	0.057	0.983	5.349	5.342
JointEmbedding	0.237	0.013	0.086	0.082	0.046	0.99	3.978	3.790
SemStyle-unordered	0.446	0.093	0.166	0.400	0.134	0.501	5.560	5.201
SemStyle-words	0.531	0.137	0.191	0.553	0.146	0.407	5.208	5.096
SemStyle-lempos	0.483	0.099	0.180	0.455	0.148	0.533	5.240	5.090
SemStyle-romonly	0.389	0.057	0.156	0.297	0.138	0.770	4.853	4.699
SemStyle	0.454	0.093	0.173	0.403	0.144	0.589	4.937	4.759

Table 2: Evaluating styled captions with automated metrics. For *SPICE* and *CLF* larger is better, for *LM* & *GRU LM* smaller is better. For metrics see the main text for baselines see Sec. 1.

Method	*Desc 0*	*Desc 1*	*Desc 2*	*Desc 3*
CNN+RNN-coco	15.6	16.7	24.2	43.4
neural-storyteller	42.3	27.3	17.0	13.5
TermRetrieval	24.4	28.5	20.3	26.8
SemStyle-romonly	16.1	24.3	25.0	34.7
SemStyle	12.2	23.2	20.9	43.8

Table 3: Human evaluations of the percentage of captions from each method that were, in regards to the image: 0 – Completely unrelated, 1 – Have a few of the right words, 2 – Almost correct with a few mistakes, 3 – Clear and accurate

Method	*% Unrelated*	*% Desc.*	*% Story*
CNN+RNN-coco	27.8	66.0	6.2
neural-storyteller	44.2	3.2	52.6
TermRetrieval	26.0	18.5	55.5
SemStyle-romonly	21.6	24.5	53.8
SemStyle	22.8	35.3	41.9

Table 4: Human evaluations of the percentage of captions from each method that were judged as: unrelated to the image content, a basic description of the image, or part of a story relating to the image.

4.3 Hypothesis Tests for Human Evaluations

Statistical hypothesis testing (null hypothesis testing) for human story judgements is shown in Table 5, for human descriptiveness judgements it is shown in Table 6. In both cases we have used $\mathcal{X}^2$ tests on method pairs with the beonferroni correction.

	CNN+RNN-coco	neural-storyteller	TermRetrieval	SemStyle-romonly
CNN+RNN-coco	-	-	-	-
neural-storyteller	5.6e-09*	-	-	-
TermRetrieval	1.2e-08*	0.88	-	-
SemStyle-romonly	2.1e-12*	0.18	0.13	-
SemStyle	1.4e-06*	0.27	0.34	0.014

Table 5: $\mathcal{X}^2$ tests on method pairs for **human story judgements**. We combine counts for "unrelated" with "purely descriptive", while "story" is kept as its own class. Those marked with a * indicate rejection of the null hypothesis (H0: the two methods give the same multinomial distribution of scores) at p-value of 0.005 – this is p-value of 0.05 with bonferroni correction of 10 to account for multiple tests.

	CNN+RNN-coco	neural-storyteller	TermRetrieval	SemStyle-romonly
CNN+RNN-coco	-	-	-	-
neural-storyteller	1e-56*	-	-	-
TermRetrieval	4.1e-18*	9.3e-14*	-	-
SemStyle-romonly	0.00032*	2.3e-35*	3.4e-07*	-
SemStyle	0.18	2.1e-48*	1.7e-13*	0.023

Table 6: $\mathcal{X}^2$ tests on method pairs for **human descriptiveness judgements**. We combine counts for "clear and accurate" with "only a few mistakes", and "some correct words" with "unrelated". Those marked with a * indicate rejection of the null hypothesis (H0: the two methods give the same multinomial distribution of scores) at p-value of 0.005 – this is p-value of 0.05 with bonferroni correction of 10 to account for multiple tests.

4.4 Attributes of the Generated Style

The style of the text is difficult to define in its entirety but we can look at a few easily identifiable style attributes to better understand the scope of the style introduced into the captions. First, we randomly sample 4000 captions or sentences from the MSCOCO and romance dataset. We then generate captions for 4000 images using *CNN+RNN-coco* and *CNN+RNN-coco*. On these four datasets we count: the percentage of sentences with past or present tense verbs (to identify the tense used in the captions), the fraction of sentences with first person pronouns (to identify sentences using first person perspective), the number of unique verbs used in the 4000 samples (to identify verb diversity). The results are summarised in Table 7. Parts-of-speech tags are obtained automatically with the spaCy[3] library. When counting verb tenses it is common to have both past an present tense for example "The dog was wearing a vest.", where "was" is past tense and "wearing" is present tense – this is why, in some cases, the sum of the past tense verbs and the present tense verbs is greater than 100%.

Captions generated by *SemStyle* use past-tense verbs in 75.0% of sentences, which is close to the ground-truth level of 72.0% and far greater than the descriptive method (*CNN+RNN-coco*) at 10.6%. This corresponds to a reduction in present tense verbs, consistent with the ground-truth. *SemStyle* includes first person pronouns in 24.4% of captions, compared to 0.0% for *CNN+RNN-coco*. The *romance ground-truth* has personal pronouns in 31.2% of sentences, which is higher than *SemStyle* – we expect that describing images limits the applicability of first person pronouns. *SemStyle* has an effective verb vocabulary almost twice as large (92.3% larger) as *CNN+RNN-coco*, which suggests more interesting verb usage. However, both *SemStyle* and *CNN+RNN-coco* have lower verb diversity than either ground-truth dataset. We expect that some verbs that are not appropriate for image captioning and the RNN with argmax decoding tends to generate more common words. Compared to *CNN+RNN-coco* the *SemStyle* model reflects the ground-truth style by generating more captions in past tense, first person, and with greater verb diversity.

To further explore the differences between styles we include Table 8 that presents the most common lemmas for each dataset, stratified by POS tag. The most common nouns generated by *SemStyle* have a greater overlap with the *MSCOCO ground-truth* than the *romance ground-truth*. This is the desired behaviour since nouns are a key component of image semantics and so nouns generated by the *term generator* should be included in the output sentence. The most common verbs generated by *SemStyle* are also similar to the *MSCOCO ground-truth*; we expect this is a result of a similar set of common verb in both ground-truth datasets. The use of determiners in *SemStyle* more closely matches the *romance ground-truth*, in particular the frequent use of the definite article "the" rather than the indefinite "a". The most common adjectives in all word sources typically relate to colours and size, and vary little across the different sources.

4.5 Precision and Recall in the Semantic Term Space

To evaluate the precision and recall in the term space we match semantic terms in the output sentence with semantic terms in the caption ground truth. The results will depend on the efficacy of the visual concept detection pipeline (eg the *term generator* for *SemStyle*) as well as the

[3] `https://github.com/explosion/spaCy/tree/v1.9.0`

	Sentences with Present Tense Verbs	Sentences with Past Tense Verbs	Sentences with First Person Pronouns	Unique Verbs
MSCOCO ground-truth	73.8%	17.0%	0.2%	497
romance ground-truth	51.4%	72.0%	31.2%	1286
CNN+RNN-coco	70.4%	10.6%	0.0%	181
SemStyle	56.8%	75.0%	24.4%	348

Table 7: Statistics on attributes of style collected from 4000 random samples from two ground-truth datasets and 4000 test captions generated by the descriptive only model (*CNN+RNN-coco*) and our *SemStyle* model. We measure the fraction of sentences or captions with present tense verbs, past tense verbs or first person pronouns. We also count the number of unique verbs used in the sample.

Word Source	Most Common Lemmas
MSCOCO ground-truth	
NOUN	man(3.7%), people(1.9%), woman(1.8%), street(1.5%), table(1.4%)
VERB	be(20.0%), sit(9.3%), stand(6.4%), hold(4.4%), ride(3.1%)
ADJ	white(6.8%), large(5.4%), black(4.1%), young(4.0%), red(3.8%)
DET	a(81.8%), the(14.9%), some(1.7%), each(0.6%), this(0.4%)
romance ground-truth	
NOUN	man(2.7%), hand(1.5%), eye(1.4%), woman(1.3%), room(1.2%)
VERB	be(15.5%), have(4.6%), do(2.5%), would(2.4%), can(1.9%)
ADJ	small(2.3%), other(2.0%), little(2.0%), black(2.0%), white(1.9%)
DET	the(60.5%), a(26.5%), that(3.2%), this(2.8%), no(1.3)%
CNN+RNN-coco	
NOUN	man(6.9%), group(3.0%), people(2.6%), table(2.6%), field(2.3%)
VERB	be(29.4%), sit(15.4%), stand(10.2%), hold(5.6%), ride(4.6%)
ADJ	large(15.0%), white(10.9%), green(4.7%), blue(4.5%), next(4.5%)
DET	a(91.9%), the(7.7%), each(0.2%), some(0.1%), an(0.1%)
SemStyle	
NOUN	man(5.5%), table(2.8%), street(2.7%), woman(2.6%), who(2.4%)
VERB	be(24.5%), sit(10.3%), stand(4.8%), have(3.6%), hold(3.2%)
ADJ	sure(14.7%), little(9.4%), hot(5.6%), single(4.7%), white(3.9%)
DET	the(68.6%), a(30.8%), no(0.2%), any(0.2%), an(0.1%)

Table 8: The most common words per part-of-speech category in the two ground truth datasets and in the sentences generated by the descriptive model (*CNN+RNN-coco*) and *SemStyle*. For each word we display the relative frequency of that word in the POS category – represented as a percentage.

Model	*Precision*	*Recall*
CNN+RNN-coco	0.561	0.517
StyleNet-coco	0.506	0.468
SemStyle-cocoonly	0.636	0.531
SemStyle-coco	0.631	0.532
StyleNet	0.027	0.028
TermRetrieval	0.505	0.336
neural-storyteller	0.234	0.225
JointEmbedding	0.340	0.177
SemStyle-unordered	0.597	0.501
SemStyle-words	0.611	0.517
SemStyle-lempos	0.593	0.504
SemStyle-romonly	0.624	0.511
SemStyle	0.626	0.517

Table 9: Precision (BLEU-1) and recall (ROUGE-1) in our semantic term space.

language generation (eg the *language generator*). While we expect a bias towards methods using our semantic term space, this analysis is useful for confirming *SemStyle* accurately produces captions with term representations similar to the ground truth. Precision is reported as BLUE-1 without length penalty on terms, while recall is reported as ROUGE-1 on terms – in both cases all ground truth reference sentences are used. BLEU-1 and ROUGE-1 are not effected by term ordering as they are uni-gram metrics. Results in Table 9 shows that the four variants of *SemStyle* (*SemStyle-cocoonly*, *SemStyle-coco*, *SemStyle-romonly*, *SemStyle*) which use our semantic term space, perform better than other model variants and baselines not using term space. Demonstrating *SemStyle* focuses on accurate reproduction of the semantic term space. The best performing models are *SemStyle-cocoonly* with the largest BLEU-1 and *SemStyle-coco* with the largest ROUGE-1 – though both models score highly in BLEU-1 and ROUGE-1. This is in line with the other automatic metrics shown in Table 1, though these metrics also show *CNN+RNN-coco* is competitive. Of the baselines the best performing is *TermRetrieval* which retrieves romance sentences using query words from a *term generator* (trained only on raw words in this case).

4.6 Choosing Semantic Terms

We defined the set of semantic terms by incorporating our domain knowledge, e.g. nouns are semantically important while determiners are not. Alternatively, we can learn which word classes carry semantic information.

We would like to know which word classes (adjectives, nouns, verbs , etc.) carry the most visually semantic information. Intuitively, we seek the word classes which, when removed, lead to the largest increase in entropy. One way to quantify this is the perplexity of the ground truth sentence after conditioning on input words belonging to different classes. For example, remove all nouns from the conditioning set of semantic terms and measure the change in perplexity.

Balancing for class frequency is necessary, because removing unimportant words such as determiners could have a large effect on perplexity if they are frequent.

Our approach requires a probabilistic model with a domain including the word classes of interest and a range including possible output sentences. One, computationally expensive, solution is to train the language generation model for each possible word class. Instead we use a single language generation model trained on input sentences with 66% of the input words randomly removed – an approach reminiscent of de-noising auto-encoders. We train this model once and then selectively drop out words during testing.

Our search for the most important word classes, starts with uniform random removal of all words down to the 33% level and thereby establishing a baseline. From there each possible word class is given a rank, higher ranked word classes are always completely removed before lower ranked word classes; removal stops when only 33% of words remain. Words from classes of the same rank are chosen uniformly at random. For example if the input sentence is "the cat on the mat ." and the removal order had nouns ranked 2 and all other parts of speech ranked 1, then nouns "cat" and "mat" would both be removed. Remaining words would be randomly removed until only 2 out of the 6 remain. Using this method we should see the lowest perplexity when the words are ordered from least important to most important.

Our forward selection approach tries to set each word type to the highest non-occupied rank or the lowest non-occupied rank, the selection which minimises the perplexity is then fixed and the search proceeds until all classes are ranked. The final ordering was **adjective, adverb, coordinating conjunction, particle, determiner, preposition or subordinate conjunction, verb, pronoun and noun**. With adjective judged the least useful and noun the most useful. Adjectives lack importance perhaps because they have only a local effect on a sentence and are often poorly detected by the CNN+RNN systems [1, 11]. This ordering is in line with our term space construction rules presented in the main paper.

Specifically we use the average perplexity per word which is equivalent to the categorical cross-entropy loss calculated with $\log_2$ rather than $\log_e$

References

[1] P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE : Semantic Propositional Image Caption Evaluation. *ECCV'16*, 1:382–398, 2016.

[2] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng. StyleNet: Generating Attractive Visual Captions with Styles. *CVPR*, 2017.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[4] K. S. Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. *Information Processing and Management: an International Journal*, 36(6):779–808, nov 2000.

[5] R. Kiros. neural-storyteller, 2015. `https://github.com/ryankiros/neural-storyteller`.

[6] R. Kiros, R. Salakhutdinov, and R. Zemel. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *arXiv preprint arXiv:1411.2539*, pages 1–13, 2014.

[7] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler. Skip-Thought Vectors. *NIPS*, (786):1–9, 2015.

[8] M.-T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser. Multi-task Sequence to Sequence Learning. *ICLR*, 2016.

[9] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ICLR'15*, 2015.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. *CVPR*, 2016.

[11] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *CVPR'15*, 2015.

[12] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. 2015.