同濟大學
TONGJI UNIVERSITY

To be an Artist: Automatic Generation on Food Image Aesthetic Captioning

Xiaohan Zou, Cheng Lin, Yinjia Zhang, Qinpei Zhao

School of Software Engineering, Tongji University

Background

Good

Score: 9

Good

Score: 7

Bad

Score: 3

Image Aesthetic Captioning

Perfect exposure on this shot, but the composition is weak, I think that you've included too much space at the bottom.

Image Captioning?

Image Captioning

(Flickr8k)

similar

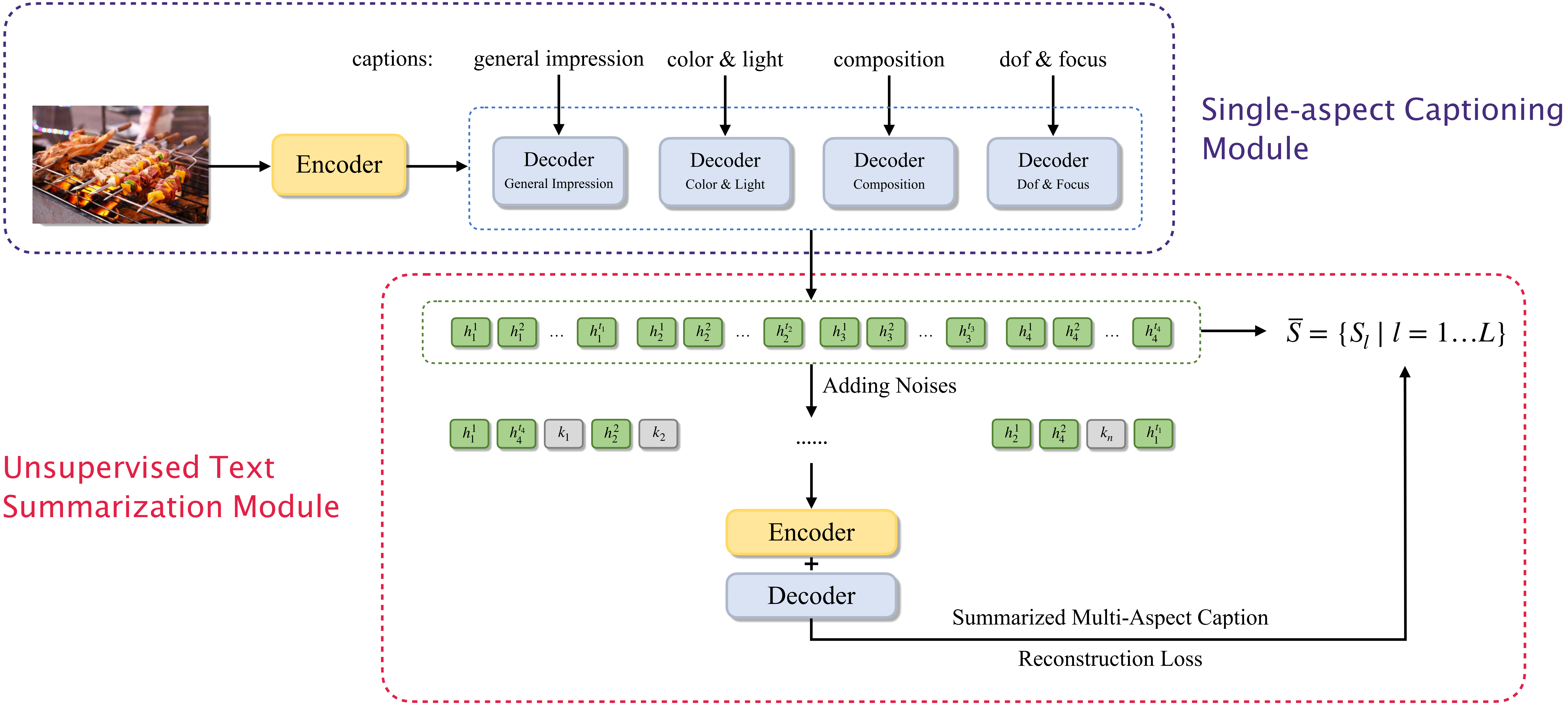
- A brown dog in the snow holding a pink hat
- A brown dog in the snow has something hot pink in its mouth
- A dog is carrying something pink in its mouth while walking through the snow

Image Aesthetic Captioning

totally different among aesthetic attributes

- **Color and Light**: Really like this, the color and contrast of the blinds really accents the subject.
- **Composition**: I like the idea and setup, but think you maybe could have cropped off a little more of the right side.

The Proposed Model

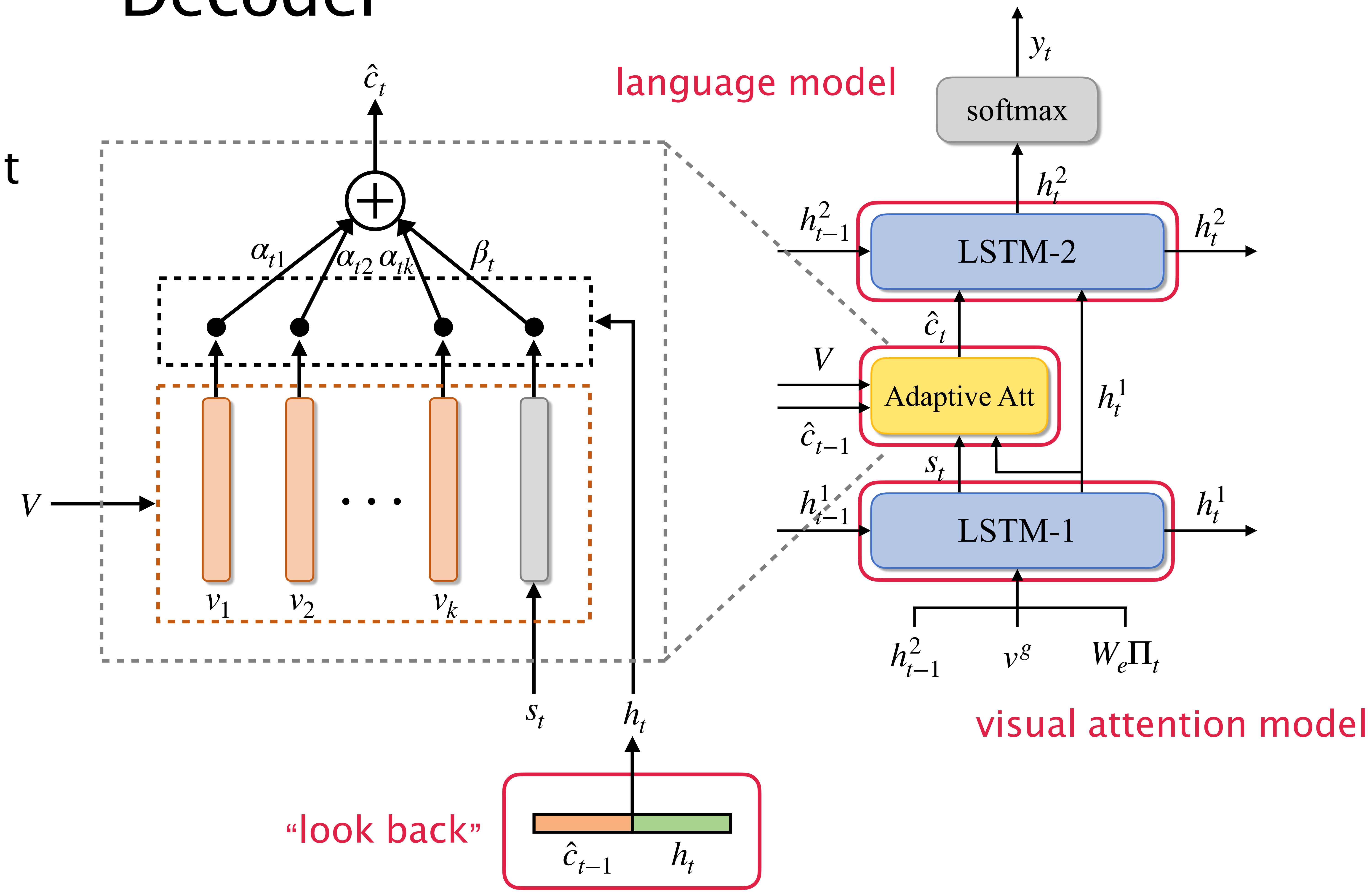


Single-Aspect Captioning (SAC) Module

Encoder

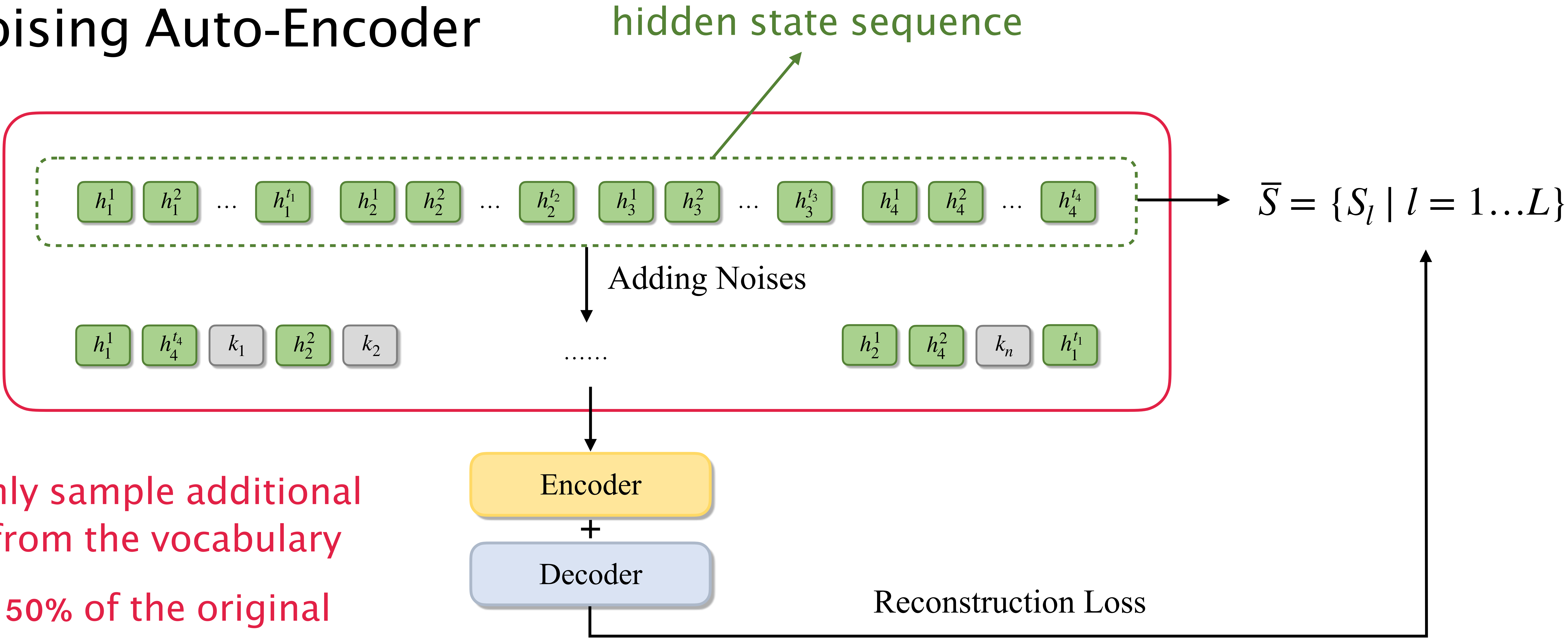
- ResNet-101 pretrained on ImageNet

Decoder



Unsupervised Text Summarization Module

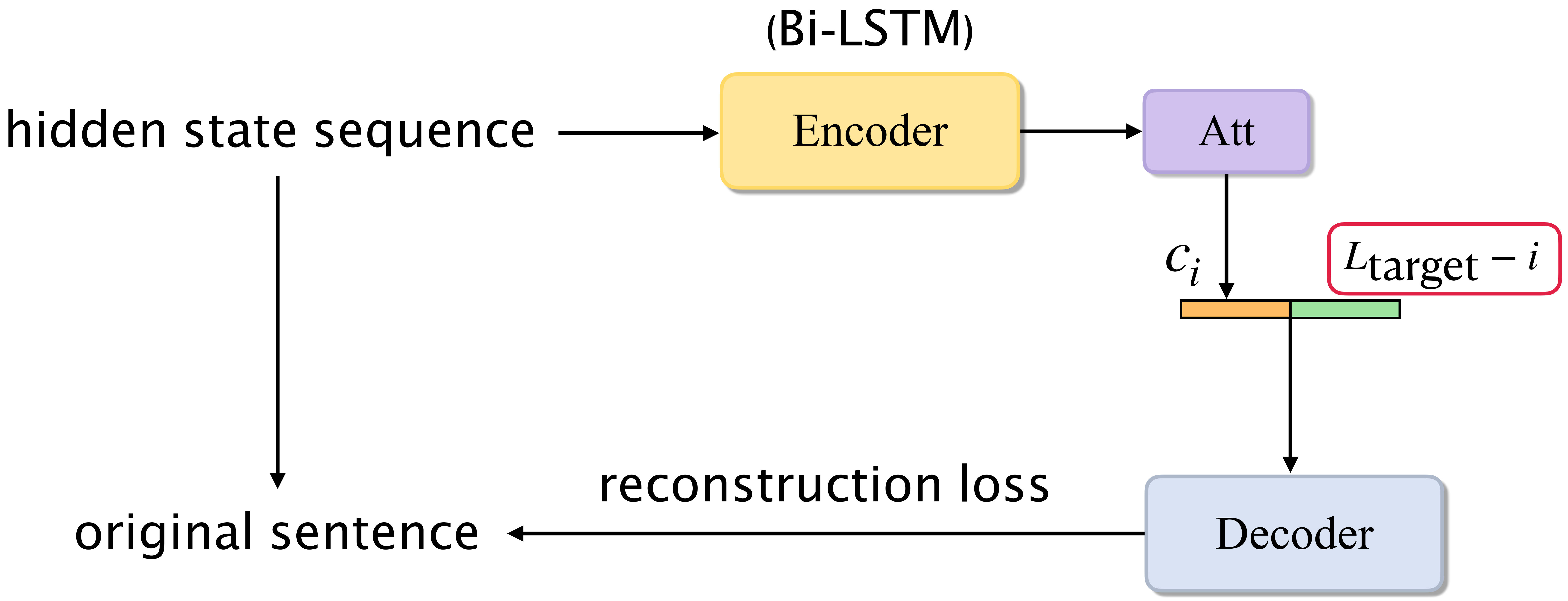
Denoising Auto-Encoder



- randomly sample additional words from the vocabulary
- extend 50% of the original sequence
- shuffle the extended sequence

Unsupervised Text Summarization Module

Encoder-Decoder



Dataset

TABLE I
STATISTICS OF OUR FOOD IMAGE AESTHETIC CAPTIONING DATASET

Aesthetic Aspect	# Images	# Captions	Vocabulary Size
Color & Light	4,708	9,402	9,425
Composition	2,778	4,092	5,802
Dof & Focus	3,843	6,758	7,069
Subject	4,104	7,706	9,031
Use of Camera	1,188	1,487	3,766
General Impression	5,835	17,040	12,859
Total	7,172	46,485	20,987

Criteria

- BLEU
- ROUGE-L
- CIDEr
- METEOR
- SPICE

Additional Criteria

Diversity

proportion of the «different» sentences

“love the color”?

$$\mathrm{sim}_n(a,b) = \mathrm{Jaccard}(g_n(a), g_n(b)) = \frac{|g_n(a) \cap g_n(b)|}{|g_n(a)| + |g_n(b)| - |g_n(a) \cap g_n(b)|} < 30\,\% \Rightarrow$$ sentence a and b are different

Novelty

difference between the generated captions and training data

$$\mathrm{nov}_n(c_i, S_i) = \mathrm{Jaccard}(g_n(c_i), g_n(S_i))$$

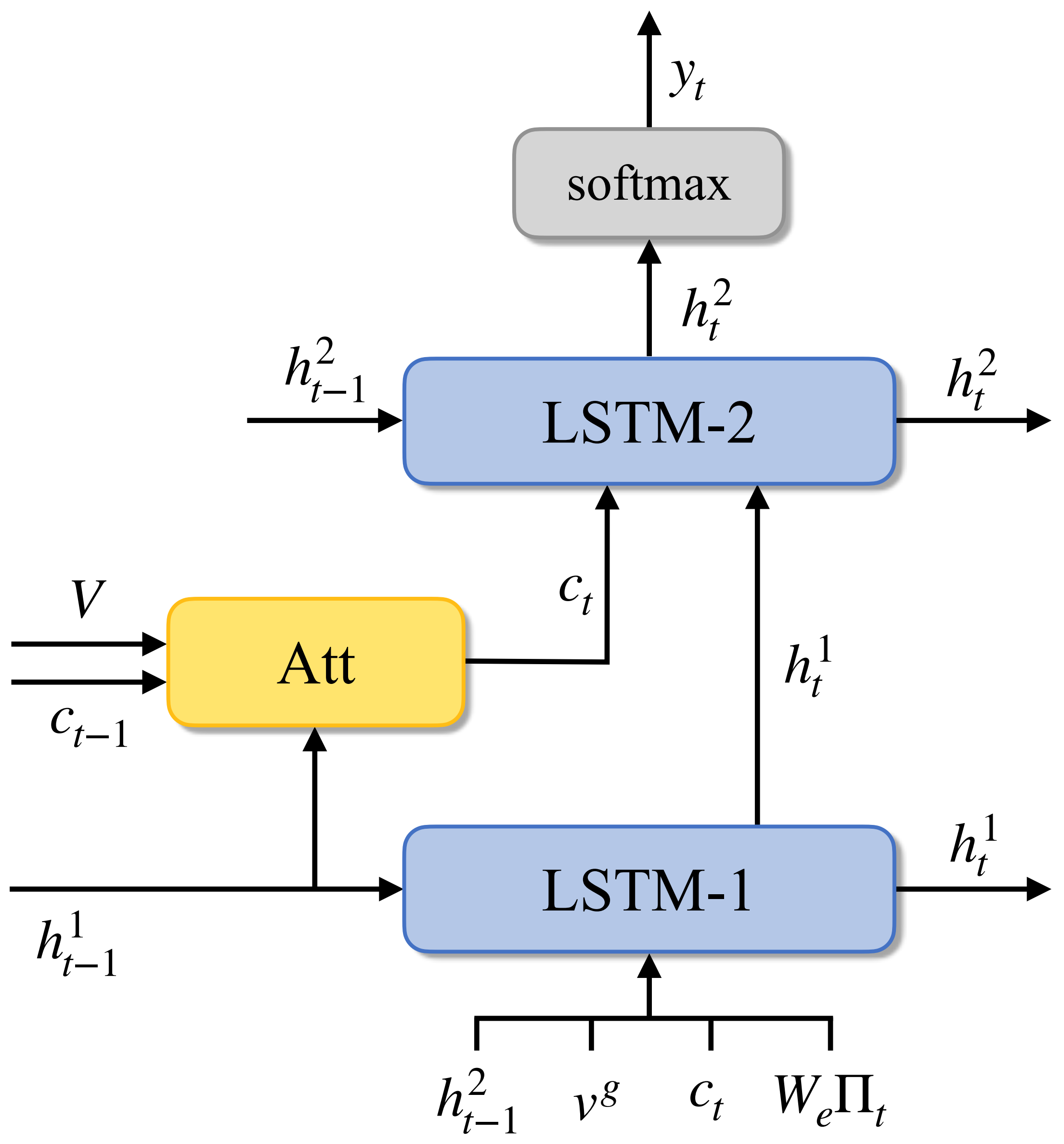
generated caption

each corresponding caption in training data

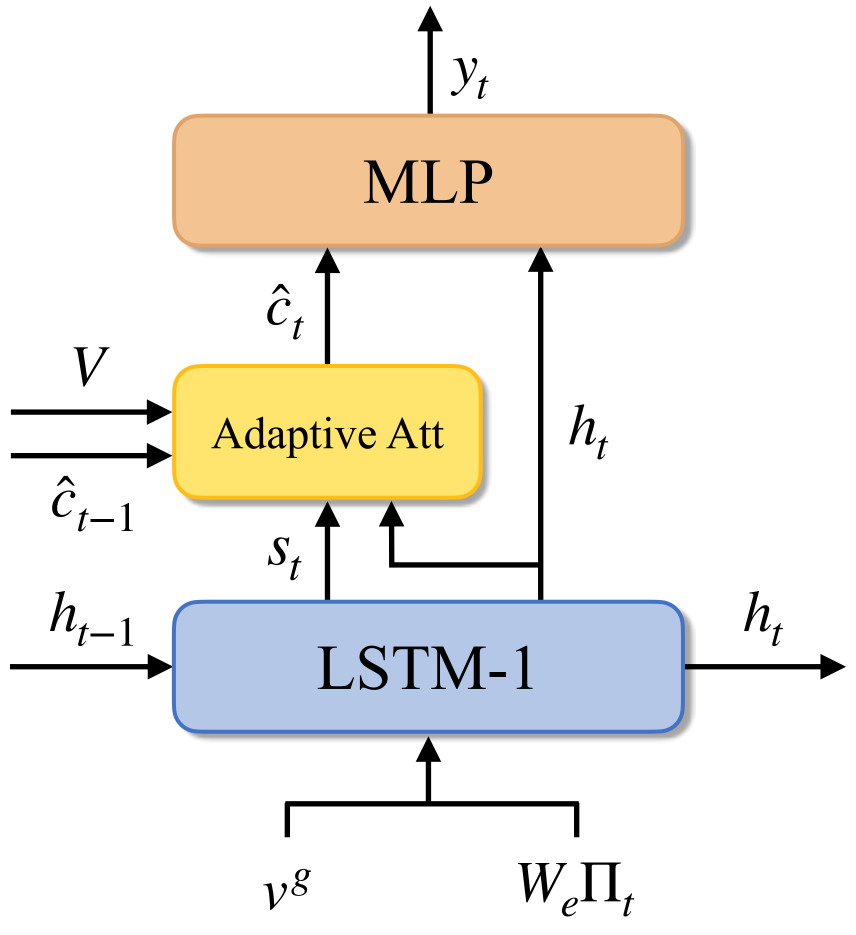
Experiments: Single-Aspect Captioning

- Baseline
  - img2txt ([Vinyals et al.])
  - Soft Attention: ([Xu et al.])
  - Adaptive Attention ([Lu et al.])

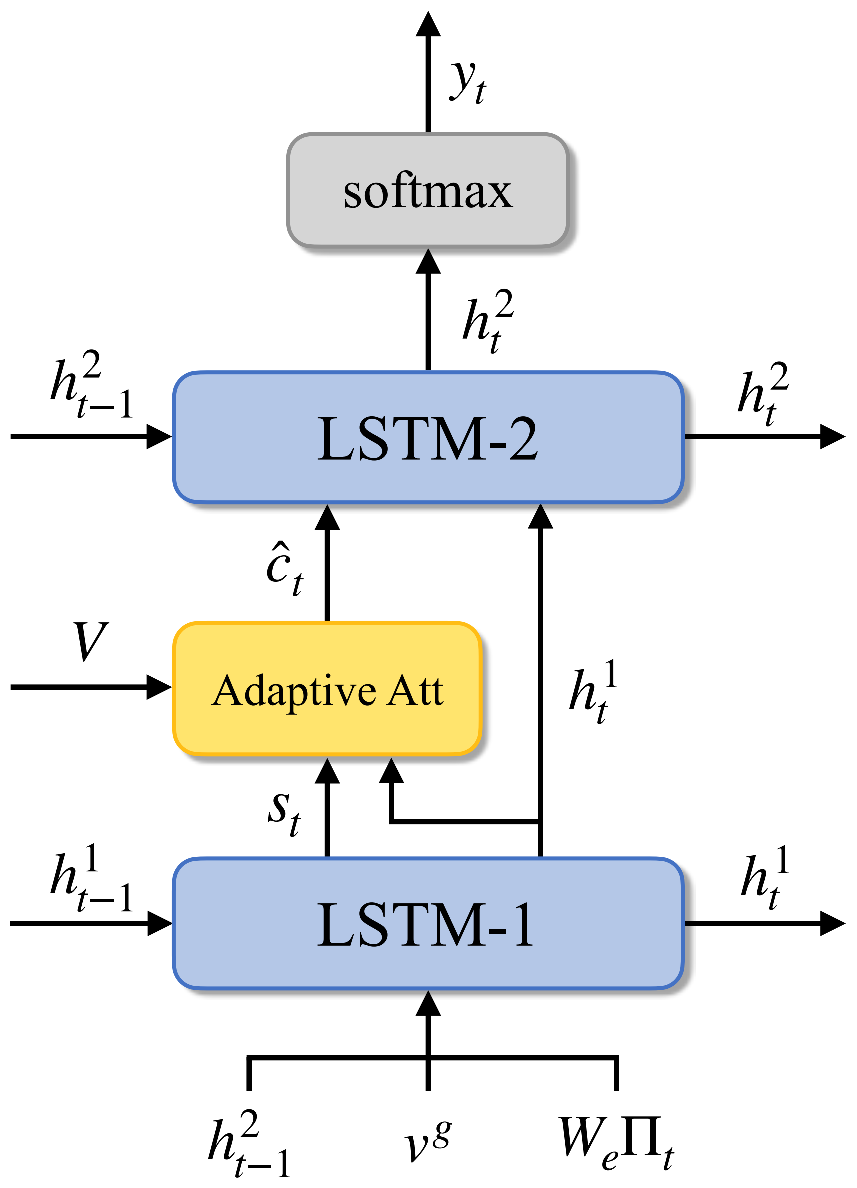
- Ablation Study



A: Soft Attention



B: Single LSTM



C: no Look Back

Experiments: Single-Aspect Captioning

Quantitative Results

TABLE II
THE COMPARISONS ON THE PERFORMANCE OF DIFFERENT MODELS ON EACH AESTHETIC ASPECT.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr	METEOR	SPICE
img2txt (General Impression)	47.3	27.1	14.0	6.8	23.8	6.0	11.7	17.9
Soft Attention (General Impression)	47.7	27.5	14.7	7.1	26.4	6.4	12.1	18.4
Adaptive Attention (General Impression)	47.9	27.8	14.8	7.1	24.3	6.5	12.4	18.7
SAC (General Impression)	**49.4**	**28.3**	**15.9**	**7.6**	**25.4**	**7.0**	**13.5**	**19.6**
img2txt (Color & Light)	44.3	24.2	14.5	6.2	24.8	5.9	10.6	14.6
Soft Attention (Color & Light)	45.0	24.6	14.7	6.8	25.3	6.1	11.1	15.5
Adaptive Attention (Color & Light)	46.1	25.1	15.0	6.9	25.6	6.0	11.4	15.8
SAC (Color & Light)	**49.7**	**27.9**	**15.6**	**7.2**	**26.4**	**6.4**	**12.7**	**17.7**
img2txt (Composition)	45.2	23.5	13.4	6.3	24.3	6.0	11.5	17.2
Soft Attention (Composition)	46.0	23.9	13.8	6.4	24.9	6.3	11.8	17.6
Adaptive Attention (Composition)	46.4	24.1	14.0	6.6	24.8	6.4	12.0	18.0
SAC (Composition)	**48.6**	**25.6**	**14.9**	**7.0**	**25.9**	**6.8**	**13.0**	**18.2**
img2txt (Dof & Focus)	44.8	23.4	13.2	6.0	24.8	5.2	10.3	14.9
Soft Attention (Dof & Focus)	45.7	24.0	13.7	6.5	25.3	5.8	10.8	15.4
Adaptive Attention (Dof & Focus)	45.4	23.8	13.5	6.3	25.6	5.6	11.0	15.3
SAC (Dof & Focus)	**46.8**	**24.9**	**14.3**	**6.7**	**26.4**	**6.2**	**12.3**	**17.0**

The BLEU-1,2,3,4, ROUGE-L, CIDEr, METEOR and SPICE are reported. All values refer to percentage (%). The proposed model and the best performance is highlighted in bold.

Experiments: Single-Aspect Captioning

Quantitative Results

TABLE III
ABLATION STUDY ON EACH AESTHETIC ASPECT

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr	METEOR	SPICE
SAC: Soft Attention (General Impression)	48.7	27.9	15.3	7.4	25.0	6.7	12.7	18.9
SAC: Single LSTM (General Impression)	48.2	27.6	15.1	7.2	24.7	6.7	12.4	19.4
SAC: No Look Back (General Impression)	49.0	28.0	15.6	7.5	25.2	6.9	13.2	19.2
SAC: Full Model (General Impression)	**49.4**	**28.3**	**15.9**	**7.6**	**25.4**	**7.0**	**13.5**	**19.6**
SAC: Soft Attention (Color & Light)	47.3	26.8	15.1	7.1	25.8	6.2	12.1	16.2
SAC: Single LSTM (Color & Light)	45.5	25.6	14.8	6.9	25.5	6.2	11.8	17.4
SAC: No Look Back (Color & Light)	48.8	27.4	15.3	7.2	26.2	**6.5**	12.4	16.8
SAC: Full Model (Color & Light)	**49.7**	**27.9**	**15.6**	**7.2**	**26.4**	6.4	**12.7**	**17.7**
SAC: Soft Attention (Composition)	47.6	24.7	14.5	6.8	25.2	6.7	11.9	17.9
SAC: Single LSTM (Composition)	46.9	24.3	14.2	6.6	25.0	6.4	12.1	**18.7**
SAC: No Look Back (Composition)	48.1	25.1	14.6	6.9	25.5	6.2	12.6	18.4
SAC: Full Model (Composition)	**48.6**	**25.6**	**14.9**	**7.0**	**25.9**	**6.8**	**13.0**	18.2
SAC: Soft Attention (Dof & Focus)	46.1	24.5	14.0	6.6	25.8	5.7	11.4	16.0
SAC: Single LSTM (Dof & Focus)	45.9	24.2	13.8	6.5	25.5	5.8	11.7	16.5
SAC: No Look Back (Dof & Focus)	46.3	24.6	14.1	6.7	26.2	6.0	12.1	**17.2**
SAC: Full Model (Dof & Focus)	**46.8**	**24.9**	**14.3**	**6.7**	**26.4**	**6.2**	**12.3**	17.0

The BLEU-1,2,3,4, ROUGE-L, CIDEr, METEOR and SPICE are reported. All values refer to percentage (%). The full model and the best performance is highlighted in bold.

Experiments: Single-Aspect Captioning

Qualitative Results

Images	Captions
	**Color & Light:** I think this would have been more effective if the lighting was a bit more.
	**Composition:** I really like the idea, but I think it would have been better if the glass was in the background.
	**General Impression:** I really like the idea of this shot.
	**Dof & Focus:** I think the dof is a little bit shallow and the background is a little bit distracting.

Fig. 4. Examples of the captions generated by our single-aspect captioning module for different aspects.

Experiments: Multi-Aspect Captioning

Experiment Settings

- Baseline
  - img2txt: apply the img2txt model (Vinyals et al.) to the whole dataset
  - SAC: apply our SAC model to the whole dataset
- Ablation Study
  - input word embeddings
  - input hidden states

Experiments: Multi-Aspect Captioning

Quantitative Results

TABLE IV
PERFORMANCE ON MULTI-ASPECT CAPTIONING

Method	D-1	D-4	N-1	N-4	S	B-4
img2txt	69.9	81.7	53.4	62.6	—	—
SAC	71.8	84.5	58.7	67.7	—	—
SACTC: Word Embeddings	93.6	98.2	76.13	**86.8**	**9.8**	4.9
SACTC: Hidden States	**94.2**	**98.7**	**77.71**	84.9	9.4	**5.1**

D-n evaluates the diversity and N-n evaluates the novelty of the generated captions in n-grams. All values refer to percentage (%). The proposed model and the best performance is highlighted in bold.

Experiments: Multi-Aspect Captioning

Qualitative Results

Image	Aesthetic Captions
	**General Impression:** the idea is cool i like the great detail and color
	Color & Light: i like the lighting on the cup but I think the background is too bright
	Composition: nice idea but i would like to see more of the cup in the center
	Dof & Focus: this is a great shot but I think the focus is a little bit soft and the background is distracting
	img2txt: nice idea but i think the light from right is a little bit harsh
	SACTC: cool idea and great detail and lighting on the cup but the background is bright and distracting like more cup in center the focus is soft

Fig. 5. An example of multi-aspect captions generated by the img2txt and the proposed model. We also report the captions generated by our single-aspect captioning models in four aesthetic aspects for reference.

Our method:

- comments on all of the four aesthetic aspects
- captures the semantic associations between captions of different aspects
- excludes unimportant phrases

同濟大學
TONGJI UNIVERSITY

Thanks!