Hierarchically-Attentive RNN for Album Summarization and Storytelling

Licheng Yu and **Mohit Bansal** and **Tamara L. Berg**
UNC Chapel Hill
{licheng, mbansal, tlberg}@cs.unc.edu

Abstract

We address the problem of end-to-end visual storytelling. Given a photo album, our model first selects the most representative (summary) photos, and then composes a natural language story for the album. For this task, we make use of the Visual Storytelling dataset and a model composed of three hierarchically-attentive Recurrent Neural Nets (RNNs) to: encode the album photos, select representative (summary) photos, and compose the story. Automatic and human evaluations show our model achieves better performance on selection, generation, and retrieval than baselines.

1 Introduction

Since we first developed language, humans have always told stories. Fashioning a good story is an act of creativity and developing algorithms to replicate this has been a long running challenge. Adding pictures as input can provide information for guiding story construction by offering visual illustrations of the storyline. In the related task of image captioning, most methods try to generate descriptions only for individual images or for short videos depicting a single activity. Very recently, datasets have been introduced that extend this task to longer temporal sequences such as movies or photo albums (Rohrbach et al., 2016; Pan et al., 2016; Lu and Grauman, 2013; Huang et al., 2016).

The type of data we consider in this paper provides input illustrations for story generation in the form of photo albums, sampled over a few minutes to a few days of time. For this type of data, generating textual descriptions involves telling a temporally consistent story about the depicted visual information, where stories must be coherent and take into account the temporal context of the images. Applications of this include constructing visual and textual summaries of albums, or even enabling search through personal photo collections to find photos of life events.

Previous visual storytelling works can be classified into two types, vision-based and language-based, where image or language stories are constructed respectively. Among the vision-based approaches, unsupervised learning is commonly applied: e.g., (Sigurdsson et al., 2016) learns the latent temporal dynamics given a large amount of albums, and (Kim and Xing, 2014) formulate the photo selection as a sparse time-varying directed graph. However, these visual summaries tend to be difficult to evaluate and selected photos may not agree with human selections. For language-based approaches, a sequence of natural language sentences are generated to describe a set of photos. To drive this work (Park and Kim, 2015) collected a dataset mined from Blog Posts. However, this kind of data often contains contextual information or loosely related language. A more direct dataset was recently released (Huang et al., 2016), where multi-sentence stories are collected describing photo albums via Amazon Mechanical Turk.

In this paper, we make use of the Visual Storytelling Dataset (Huang et al., 2016). While the authors provide a seq2seq baseline, they only deal with the task of generating stories given 5-representative (summary) photos hand-selected by people from an album. Instead, we focus on the more challenging and realistic problem of end-to-end generation of stories from entire albums. This requires us to either generate a story from all of the album's photos or to learn selection mechanisms to identify representative photos and then generate stories from those summary photos. We evaluate each type of approach.

Ultimately, we propose a model of hierarchically-attentive recurrent neural nets,

consisting of three RNN stages. The first RNN encodes the whole album context and each photo's content, the second RNN provides weights for photo selection, and the third RNN takes the weighted representation and decodes to the resulting sentences. Note that during training, we are only given the full input albums and the output stories, and our model needs to learn the summary photo selections latently.

We show that our model achieves better performance over baselines under both automatic metrics and human evaluations. As a side product, we show that the latent photo selection also reasonably mimics human selections. Additionally, we propose an album retrieval task that can reliably pick the correct photo album given a sequence of sentences, and find that our model also outperforms the baselines on this task.

2 Related work

Recent years have witnessed an explosion of interest in vision and language tasks, reviewed below.

Visual Captioning: Most recent approaches to image captioning (Vinyals et al., 2015b; Xu et al., 2015) have used CNN-LSTM structures to generate descriptions. For captioning video or movie content (Venugopalan et al., 2015; Pan et al., 2016), sequence-to-sequence models are widely applied, where the first sequence encodes video frames and the second sequence decodes the description. Attention techniques (Xu et al., 2015; Yu et al., 2016; Yao et al., 2015) are commonly incorporated for both tasks to localize salient temporal or spatial information.

Video Summarization: Similar to documentation summarization (Rush et al., 2015; Cheng and Lapata, 2016; Mei et al., 2016; Woodsend and Lapata, 2010) which extracts key sentences and words, video summarization selects key frames or shots. While some approaches use unsupervised learning (Lu and Grauman, 2013; Khosla et al., 2013) or intuitive criteria to pick salient frames, recent models learn from human-created summaries (Gygli et al., 2015; Zhang et al., 2016b,a; Gong et al., 2014). Recently, to better exploit semantics, (Choi et al., 2017) proposed textually customized summaries.

Visual Storytelling: Visual storytelling tries to tell a coherent visual or textual story about an image set. Previous works include storyline graph modeling (Kim and Xing, 2014), unsupervised mining (Sigurdsson et al., 2016), blog-photo alignment (Kim et al., 2015), and language retelling (Huang et al., 2016; Park and Kim, 2015). While (Park and Kim, 2015) collects data by mining Blog Posts, (Huang et al., 2016) collects stories using Mechanical Turk, providing more directly relevant stories.

3 Model

Our model (Fig. 1) is composed of three modules: Album Encoder, Photo Selector, and Story Generator, jointly learned during training.

3.1 Album Encoder

Given an album $A = \{a_1, a_2, ..., a_n\}$, composed of a set of photos, we use a bi-directional RNN to encode the local album context for each photo. We first extract the 2048-dimensional visual representation $f_i \in R^k$ for each photo using ResNet101 (He et al., 2016), then a bi-directional RNN is applied to encode the full album. Following (Huang et al., 2016), we choose a Gated Recurrent Unit (GRU) as the RNN unit to encode the photo sequence. The sequence output at each time step encodes the local album context for each photo (from both directions). Fused with the visual representation followed by ReLU, our final photo representation is (top module in Fig. 1):

$$
\begin{aligned}
f_i &= \text{ResNet}(a_i) \\
\vec{h}_i &= \vec{\text{GRU}}_{album}(f_i, \vec{h}_{i-1}) \\
\overleftarrow{h}_i &= \overleftarrow{\text{GRU}}_{album}(f_i, \overleftarrow{h}_{i+1}) \\
v_i &= \text{ReLU}([\vec{h}_i, \overleftarrow{h}_i] + f_i).
\end{aligned}
$$

3.2 Photo Selector

The Photo Selector (illustrated in the middle yellow part of Fig. 1) identifies representative photos to summarize an album's content. As discussed, we do not assume that we are given the ground-truth album summaries during training, instead regarding selection as a latent variable in the end-to-end learning. Inspired by Pointer Networks (Vinyals et al., 2015a), we use another GRU-RNN to perform this task [1].

Given the album representation $V^{n\times k}$, the photo selector outputs probabilities $p_t \in R^n$ (likelihood of selection as t-th summary image) for all photos using soft attention.

$$
\begin{aligned}
\bar{h}_t &= \text{GRU}_{select}(p_{t-1}, \bar{h}_{t-1}), \\
p(y_{a_i}(t) = 1) &= \sigma(\text{MLP}([\bar{h}_t, v_i])),
\end{aligned}
$$

[1] While the pointer network requires grounding labels, we regard the labels as latent variables

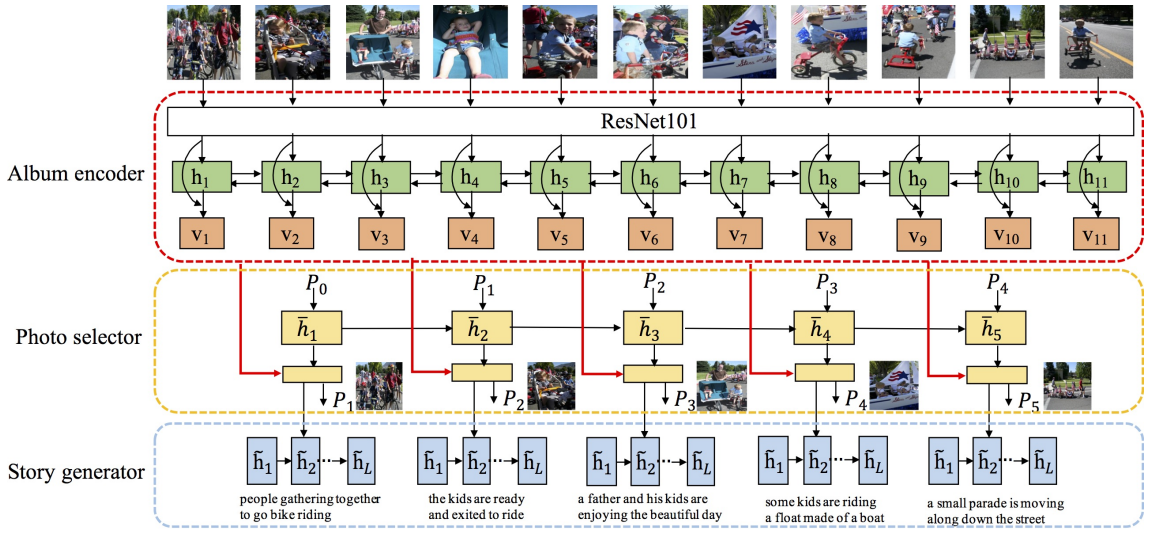


Figure 1: Model: the *album encoder* is a bi-directional GRU-RNN that encodes all album photos; the *photo selector* computes the probability of each photo being the tth album-summary photo; and finally, the *story generator* outputs a sequence of sentences that combine to tell a story for the album.

At each summarization step, t, the GRU takes the previous p_{t-1} and previous hidden state as input, and outputs the next hidden state $\bar{h}_t$. $\bar{h}_t$ is fused with each photo representation v_i to compute the i^{th} photo's attention $p_t^i = p(y_{a_i}(t) = 1)$. At test time, we simply pick the photo with the highest probability to be the summary photo at step t.

3.3 Story Generator

To generate an album's story, given the album representation matrix V and photo summary probabilities p_t from the first two modules, we compute the visual summary representation $g_t \in R^k$ (for the t-th summary step). This is a weighted sum of the album representations, i.e., $g_t = p_t^T V$. Each of these 5 g_t embeddings (for $t = 1$ to 5) is then used to decode 1 of the 5 story sentences respectively, as shown in the blue part of Fig. 1.

Given a story $S = \{s_t\}$, where s_t is t-th summary sentence. Following Donahue et al. (2015), the l-th word probability of the t-th sentence is:

$$\begin{aligned} w_{t,l-1} &= W_e s_{t,l-1}, \\ \tilde{h}_{t,l} &= \text{GRU}_{story}(w_{t,l-1}, g_t, \tilde{h}_{t,l-1}), \\ p(s_{t,l}) &= \text{softmax}(\text{MLP}(\tilde{h}_{t,l})), \end{aligned} \quad (1)$$

where W_e is the word embedding. The GRU takes the joint input of visual summarization g_t, the previous word embedding $w_{t,l}$, and the previous hidden state, then outputs the next hidden state. The generation loss is then the sum of the negative log likelihoods of the correct words: $L_{gen}(S) = -\sum_{t=1}^{T}\sum_{l=1}^{L_t} \log p_{t,l}(s_{t,l})$.

To further exploit the notion of temporal coherence in a story, we add an order-preserving constraint to order the sequence of sentences within a story (related to the story-sorting idea in Agrawal et al. (2016)). For each story S we randomly shuffle its 5 sentences to generate negative story instances S'. We then apply a max-margin ranking loss to encourage correctly-ordered stories: $L_{rank}(S, S') = \max(0, m - \log p(S') + \log p(S))$. The final loss is then a combination of the generation and ranking losses:

$$L = L_{gen}(S) + \lambda L_{rank}(S, S'). \quad (2)$$

4 Experiments

We use the Visual Storytelling Dataset (Huang et al., 2016), consisting of 10,000 albums with 200,000 photos. Each album contains 10-50 photos taken within a 48-hour span with two annotations: 1) 2 album summarizations, each with 5 selected representative photos, and 2) 5 stories describing the selected photos.

4.1 Story Generation

This task is to generate a 5-sentence story describing an album. We compare our model with two sequence-to-sequence baselines: 1) an encoder-decoder model (enc-dec), where the sequence of album photos is encoded and the last hidden state is fed into the decoder for story generation, 2) an encoder-attention-decoder model (Xu et al., 2015) (enc-attn-dec) with weights computed using a soft-attention mechanism. At each decoding time step, a weighted sum of hidden states from the encoder is decoded. For fair comparison, we

beam size=3				
	Bleu3	Rouge	Meteor	CIDEr
enc-dec	19.58	29.23	33.02	4.65
enc-attn-dec	19.73	28.94	32.98	4.96
h-attn	20.53	29.82	33.81	6.84
h-attn-rank	**20.78**	**29.82**	**33.94**	**7.38**
h-(gd)attn-rank	21.02	29.53	34.12	7.51

Table 1: Story generation evaluation.

enc-dec (29.50%)	h-attn-rank (70.50%)
enc-attn-dec (30.75%)	h-attn-rank (69.25%)
h-attn-rank (30.50%)	gd-truth (69.50%)

Table 2: Human evaluation showing how often people prefer one model over the other.

use the same album representation (Sec. 3.1) for the baselines.

We test two variants of our model trained with and without ranking regularization by controlling λ in our loss function, denoted as h-attn (without ranking), and h-attn-rank (with ranking). Evaluations of each model are shown in Table 1. The h-attn outperforms both baselines, and h-attn-rank achieves the best performance for all metrics. Note, we use beam-search with beam size=3 during generation for a reasonable performance-speed trade-off (we observe similar improvement trends with beam size = 1).[2] To test performance under optimal image selection, we use one of the two ground-truth human-selected 5-photo-sets as an oracle to hard-code the photo selection, denoted as h-(gd)attn-rank. This achieves only a slightly higher Meteor compared to our end-to-end model.

Additionally, we also run human evaluations in a forced-choice task where people choose between stories generated by different methods. For this evaluation, we select 400 albums, each evaluated by 3 Turkers. Results are shown in Table 2. Experiments find significant preference for our model over both baselines. As a simple Turing test, we also compare our results with human written stories (last row of Table 2), indicating room for improvement of methods.

4.2 Album Summarization

We evaluate the precision and recall of our generated summaries (output by the photo selector) compared to human selections (the combined set of both human-selected 5-photo stories). For comparison, we evaluate enc-attn-dec on the same task by aggregating predicted attention and selecting the 5 photos with highest accumulated attention. Additionally, we also run DPP-based video summarization (Kulesza et al., 2012) using the same album features. Our models have higher performance compared to baselines as shown in Table 3 (though DPP also achieves strong results, indicating that there is still room to improve the pointer network).

	precision	recall
DPP	43.75%	27.41%
enc-attn-dec	38.53%	24.25%
h-attn	42.85%	27.10%
h-attn-rank	**45.51%**	**28.77%**

Table 3: Album summarization evaluation.

	R@1	R@5	R@10	MedR
enc-dec	10.70%	29.30%	41.40%	14.5
enc-attn-dec	11.60%	33.00%	45.50%	11.0
h-attn	18.30%	**44.50%**	**57.60%**	**6.0**
h-attn-rank	**18.40%**	43.30%	55.50%	7.0

Table 4: 1000 album retrieval evaluation.

4.3 Output Example Analysis

Fig. 2 shows several output examples for both summarization and story generation, comparing our model to the baseline and ground-truth. More examples are provided in the supplementary.

4.4 Album Retrieval

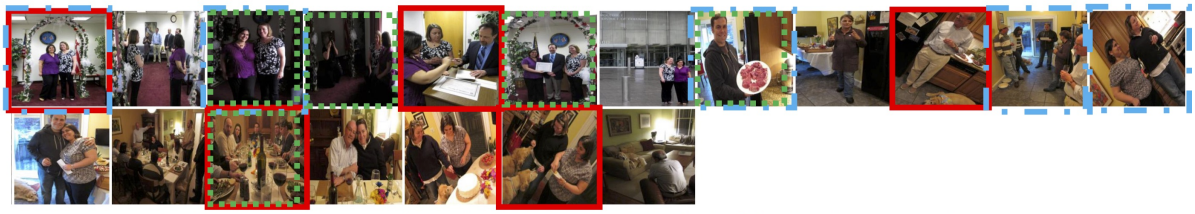
Given a human-written story, we introduce a task to retrieve the album described by that story. We randomly select 1000 albums and one ground-truth story from each for evaluation. Using the generation loss, we compute the likelihood of each album A_m given the query story S and retrieve the album with the highest generation likelihood, $A = \text{argmax}_{A_m} p(S|A_m)$. We use Recall@k and Median Rank for evaluation. As shown in Table 4), we find that our models outperform the baselines, but the ranking term in Eqn.2 does not improve performance significantly.

5 Conclusion

Our proposed hierarchically-attentive RNN based models for end-to-end visual storytelling can jointly summarize and generate relevant stories from full input photo albums effectively. Automatic and human evaluations show that our method outperforms strong sequence-to-sequence

[2] We also compute the p-value of Meteor on 100K samples via the bootstrap test (Efron and Tibshirani, 1994), as Meteor has better agreement with human judgments than Bleu/Rouge (Huang et al., 2016). Our h-attn-rank model has strong statistical significance ($p = 0.01$) over the enc-dec and enc-attn-dec models (and is similar to the h-attn model).

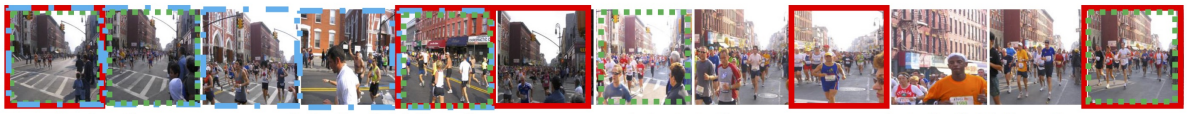
enc-attn-dec: the students were ready for the meeting . the community . they were all ready for the meeting . they were all ready . they had to make sure to make sure had to be done .
hattn-rank: i went to the organization. they were some speakers . they were a lot of the lecture . the students were very happy to see the speaker. the last speaker was the best part of the day .
gd-truth: i walked into the building to give my speech . there were many topics that need to be covered . we stood there and reviewed them . members got to ask questions . we sat and came to an agreement .



enc-attn-dec: the family gathered for dinner for dinner . they had a lot of food . the food and friends and the best together . the best part of the night was a lot of course had a lot of fun .
hattn-rank: the family gathered their friends . they had a great party . They had a lot of people showed up to celebrate the occasion . everyone was so happy to be together . after dinner was over , they had a good drink it was a success .
gd-truth: this couple are going to get married . they took pictures of this event . they got their license . their friend cooked them steaks . they all had a big dinner afterwards .

enc-attn-dec: i went to the fair . the organization was there . there were many people there . the cake was very delicious cake we had a lot of fun .
hattn-rank: the kids were excited for my birthday . the kids were decorated with balloons . the kids were playing with each other. the cake was decorated . We all had a lot of the gifts .
gd-truth: the group had a celebration among themselves . the room was decorated with balloons and ribbons . the cake was enjoyed by all . and there was plenty of food . a few antics were performed to entertain the crowd

enc-attn-dec: today was the soldiers are presenting to the ceremony . they are presenting the award for the men . they all of them . the soldiers were presented . the speech was given to the award .
hattn-rank: today was a great day for the military . the soldiers were very proud of the ceremony . the soldiers were very happy to receive the award . they were very happy to be there . the men shook hands in the event .
gd-truth: i went to the award ceremony yesterday . there were a lot of people there . everyone received an award for their effort . they had a great time . i really enjoyed being there . some of the soldiers started singing .



enc-attn-dec: the runners were ready for the finish line . the runners were ready to start . the runners were ready to the finish line . the runners were ready for the race . the runners .
hattn-rank: the runners were getting ready for the race . the runners were all lined up . the runners were close to the runners . the runners were very tired . the winner the finish line .
gd-truth: the front runners raced through the city . another group followed behind . some waved at supporters . there were many participants in the race . some had to walk to the finish line .

Figure 2: Examples of album summarization and storytelling by enc-attn-dec (blue), h-attn-rank (red), and ground-truth (green). We randomly select 1 out of 2 human album summaries as ground-truth here.

baselines on selection, generation, and retrieval tasks.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This research is supported by NSF Awards #1633295, 1444234, 1445409, 1562098.

References

Harsh Agrawal, Arjun Chandrasekaran, Dhruv Batra, Devi Parikh, and Mohit Bansal. 2016. Sort story: Sorting jumbled images and captions into stories. In *EMNLP*.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *ACL*.

Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. 2017. Textually customized video summaries. *arXiv preprint arXiv:1702.01528*.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. 2014. Diverse sequential subset selection for supervised video summarization. In *NIPS*.

Michael Gygli, Helmut Grabner, and Luc Van Gool. 2015. Video summarization by learning submodular mixtures of objectives. In *CVPR*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *NACCL*.

Aditya Khosla, Raffay Hamid, Chih-Jen Lin, and Neel Sundaresan. 2013. Large-scale video summarization using web-image priors. In *CVPR*.

Gunhee Kim, Seungwhan Moon, and Leonid Sigal. 2015. Joint photo stream and blog post summarization and exploration. In *CVPR*.

Gunhee Kim and Eric P Xing. 2014. Reconstructing storyline graphs for image recommendation from web community photos. In *CVPR*.

Alex Kulesza, Ben Taskar, et al. 2012. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*.

Zheng Lu and Kristen Grauman. 2013. Story-driven summarization for egocentric video. In *CVPR*.

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2016. What to talk about and how? selective generation using lstms with coarse-to-fine alignment. In *NAACL*.

Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. 2016. Jointly modeling embedding and translation to bridge video and language. In *CVPR*.

Cesc C Park and Gunhee Kim. 2015. Expressing an image stream with a sequence of natural sentences. In *NIPS*.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2016. Movie description. *IJCV*.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *EMNLP*.

Gunnar A Sigurdsson, Xinlei Chen, and Abhinav Gupta. 2016. Learning visual storylines with skipping recurrent neural networks. In *ECCV*.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *ICCV*.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015a. Pointer networks. In *NIPS*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015b. Show and tell: A neural image caption generator. In *CVPR*.

Kristian Woodsend and Mirella Lapata. 2010. Automatic generation of story highlights. In *ACL*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *ICCV*.

Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*.

Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016a. Summary transfer: Exemplar-based subset selection for video summarization. In *CVPR*.

Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016b. Video summarization with long short-term memory. In *ECCV*.