Swell-and-Shrink: Decomposing Image Captioning by Transformation and Summarization

Hanzhang Wang , Hanli Wang* , Kaisheng Xu
Department of Computer Science and Technology, Tongji University, Shanghai, P. R. China
{hanzhang.wang, hanliwang, iaalm}tongji.edu.cn

Abstract

Image captioning is currently viewed as a problem analogous to machine translation. However, it always suffers from poor interpretability, coarse or even incorrect descriptions on regional details. Moreover, information abstraction and compression, as essential characteristics of captioning, are always overlooked and seldom discussed. To overcome the shortcomings, a swell-shrink method is proposed to redefine image captioning as a compositional task which consists of two separated modules: modality transformation and text compression. The former is guaranteed to accurately transform adequate visual content into textual form while the latter consists of a hierarchical LSTM which particularly emphasizes on removing the redundancy among multiple phrases and organizing the final abstractive caption. Additionally, the order and quality of region of interest and modality processing are studied to give insights of better understanding the influence of regional visual cues on language forming. Experiments demonstrate the effectiveness of the proposed method.

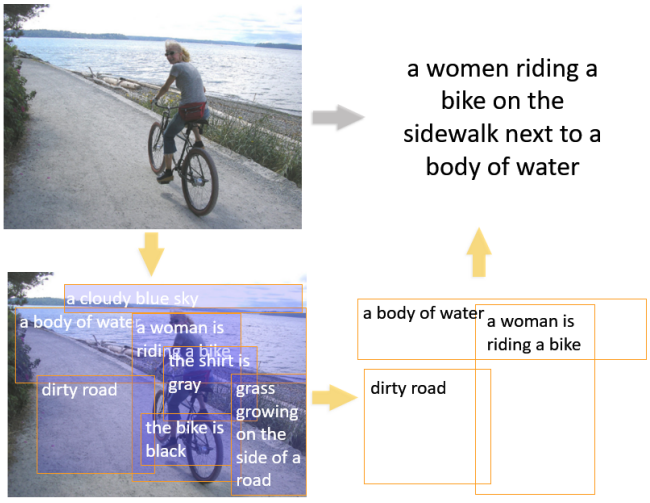


Figure 1: The top part shows traditional image captioning models that directly translate an image into one sentence. The bottom part shows the diagram of the proposed swell-shrink method, which firstly swells one image to multiple smaller regions and then shrinks their corresponding descriptions to one caption.

1 Introduction

To achieve image captioning automatically, many efforts [Karpathy and Fei-Fei, 2015; Vinyals *et al.*, 2015] have been put on optimizing the semantic alignment between vision and language, because it is typically treated as a cross-modality task. Despite these successful models, the problem of information summarization in image captioning is seldom discussed. A picture is worth a thousand words whereas its caption only contains one sentence. Modality transformation is always inseparable from the selective extraction of information. However, unlike visual question answering (VQA) that provides explicit cues to narrow the selection by paired questions, the reduced cues for image captioning are rather vague and subjective.

Another problem is interpretability. Most captioning models encode visual content into implicit representations which are then decoded into texts. Once generating undesired descriptions, it usually relies on the method of trial and error [Zeiler and Fergus, 2014] to diagnose and improve model performance. For example, when the generated caption is ambiguous or even unrelated, it is difficult to decide whether the visual information is incorrectly extracted or the sentence is not well organized. A better interpretable model should be convenient to understand its internal mechanism leading to insights about the association between vision and language.

In this work, image captioning is achieved by the proposed swell-shrink method which consists of two interpretable modules: modality transformation and text summarization, as illustrated in Fig. 1. It enables a better understanding of the mapping between visual spatial regions and language phrases, and its modularity hyalinizes the contribution of each part of the model. First, an image is segmented to several Regions of Interest (RoI) by Faster R-CNN, then the visual features of each region are fed into a repeated long short-term memo-

*Corresponding author: Hanli Wang. This work was supported in part by National Natural Science Foundation of China under Grant 61622115, National Key R&D Program of China (2017YFB1401404), and Shanghai Engineering Research Center of Industrial Vision Perception & Intelligent Computing (17DZ2251600).

ry (LSTM) network to obtain a concise phrase. By this means, image captioning is simplified to several sub-tasks so that visual information can be adequately represented. The second part is an abstractive summarization model. It consists of a hierarchical LSTM that the bottom layer is a repeated bi-directional LSTM to precisely represent the embedding of previous phrases. Then the embedded phrases are ranked and fed into the top layer as the input of each time step. The output from the last time step of the top layer is the representation of final abstractive caption.

The contributions of this paper are summarized as follows. First, image captioning is redefined as a compositional task with modality transformation and text summarization, which shows better interpretability on analyzing the contribution of each parts. Second, to emphasize the summarization for image captioning, a hierarchical LSTM model is particularly designed to fuse and abstract textual descriptions of the corresponding visual regions. Third, the influence of the number and the order of visual cues on caption generation is elaborately discussed with visualized examples. This opens up the possibility to study the relation among visual regions with global language organization.

2 Related Work

2.1 Image Captioning

Recent image captioning models are inspired by neural machine translation [Bahdanau *et al.*, 2015], which follows the sequence-to-sequence pipeline, treating image and corresponding text description as source language and target language, respectively. Specifically, Convolutional Neural Network (CNN) is adopted as the encoder to extract visual features and Recurrent Neural Network (RNN) as the decoder to generate sentences [Karpathy and Fei-Fei, 2015; Vinyals *et al.*, 2015]. Between encoder and decoder, CNN features are projected into the same representation space as word embedding to realize mapping from vision to language. However, coarse description for a whole image is no longer satisfying, more studies turn to detailed and richer descriptions on visual regions. [Johnson *et al.*, 2016] introduces dense captioning which is able to generate a set of captions that describe image details by focusing more on classification, attribute and relationship. Extended on dense captioning, [Krause *et al.*, 2017] proposes a hierarchical RNN model to generate an entire coherent paragraph instead of merely concatenated sentences. Meanwhile, [Liang *et al.*, 2017] proposes a recurrent topic-transition model based on generative adversarial network to generate paragraph in a semi-supervised way, it incorporates regional visual features with language attention mechanism and discriminates the generated samples on both sentence level and topic level. Besides, local features based on object and salient regions are proven to be effective and efficient on both image captioning and VQA [Anderson *et al.*, 2018]. Much implicit information such as location and relation [Li *et al.*, 2017b] can be inferred based on regional contextual features, which further improves language modeling [Yang *et al.*, 2017].

2.2 Text Summarization

Extractive summarization models produce summary by selecting a subset of relevant sentences from the original document. Several heuristic features including term frequency-inverse document frequency, sentence length and position are proved to be robust and commonly used for scoring sentences. [Yin and Pei, 2015] obtains sentence representation by CNN and selects the appropriate sentence by optimizing prestige and dissimilarity criterion. [Cheng and Lapata, 2016] and [Nallapati *et al.*, 2017] use a hierarchical RNN to derive the representation of word, sentence and document. Recently, reinforcement learning is adopted to produce better coherence on both semantic and syntactic [Wu and Hu, 2018].

Abstractive summarization models generate summary by re-organizing words and phrases into different forms. [Rush *et al.*, 2015] formulates abstractive summarization as a sequence-to-sequence learning problem by an attention-based encoder to read the texts and generates the summary. [Miao and Blunsom, 2016] formulates a variational auto-encoder to infer the latent representation for sentence summarization. Common issues in abstractive models such as keyword representation, sentence-to-word hierarchy and unseen words, are detailedly analyzed in [Nallapati *et al.*, 2016]. Besides, data augmentation is also effective to improve the generalization capability of models by generating variants of words, phrases or concepts.

3 Proposed Swell-Shrink Framework

The diagram of the proposed swell-shrink method is shown in Fig. 2, which contains two parts. The left half is the modality transformation module, which transforms an image to a set of textual representations by describing visual regions concisely and precisely. It is achieved by a Faster R-CNN based model followed by a repeated encoder-decoder structure. The right half is the text summarization module to compress the set of descriptions and form an abstractive sentence as the caption. It contains an embedding layer with bidirectional LSTMs and a summarization layer.

3.1 Modality Transformation

Given an image I, its visual content could be represented as a set of RoI regions $\{r_1, r_2, \cdots, r_k\}$, which is then transformed to a textual sequence $\{p_1, p_2, \cdots, p_k\}$ accordingly. The transformation is implemented by dense captioning [Johnson *et al.*, 2016]. Image I is firstly processed by CNN to extract visual features, and a Faster R-CNN based layer is applied to select k RoIs. Then the corresponding k activations are extracted by bilinear interpolation, meanwhile confidential scores and coordinates are generated as the output. The visual features of these k RoIs are flattened and passed through linear layers, and are subsequently fed to LSTM to generate k phrases which describe the corresponding RoI regions.

In implementation, the number of RoIs (*i.e.*, k) and the representation of these k regions significantly affect whether image content can be fully expressed. Although some entities in the image will not be explicitly expressed in the final caption, they still assist language modeling by providing contextual

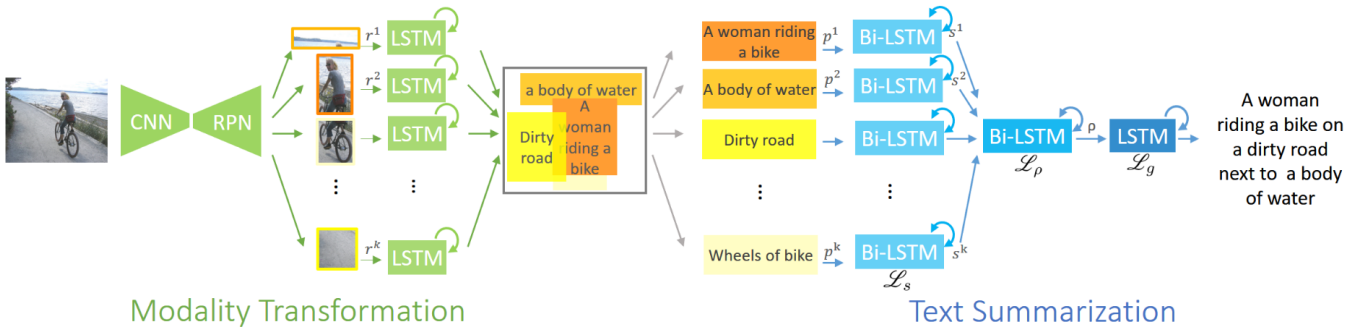


Figure 2: Overview of the proposed swell-shrink method. The modality transformation module divides a given image into multiple regions of interest and meanwhile extracts visual features which are subsequently fed into a repeated LSTM to generate phrases. The text summarization module is comprised of a ranking procedure and a hierarchical LSTM assigned to compress the phrases to an abstractive caption.

information. This observation will be demonstrated by the experiment and discussion in Section 4.3 where a comparison of caption generation is conducted with different numbers of RoIs. In this work, k is finally set to 30; although this number is usually larger than the quantity of objects in an image, it can benefit language generation well.

3.2 Text Summarization

Problem Formulation

After modality transformation, all visual contents are represented as a set of phrases denoted as $\{p_1, p_2, \cdots, p_k\}$, which makes up a document D by concatenation. Then the image captioning task is transformed to be a problem of text summarization on this single document. In order to generate a caption, denoted as a word sequence $\{\pi_1, \pi_2, \cdots, \pi_N\}$, where N is the number of words, the log distribution of language model is approximated as

$$\log p(\pi|D) = \sum_{t=1}^{N} \log p(\pi_t|D, \pi_{1:t-1}), \tag{1}$$

where $\pi_{1:t-1}$ is the slice from 1 to $(t-1)$ elements of π. To model $\log p(\pi_t|D, \pi_{1:t-1})$, LSTM is employed with the output as

$$h_t = f(\pi_t, h_{t-1}), \tag{2}$$

where h_t denotes the output at the t_{th} time step, $f(\cdot)$ indicates a RNN cell which computes as LSTM $\mathscr{L}_s$ or LSTM $\mathscr{L}_g$ in different layers, and the output from LSTM $\mathscr{L}_s$ instead of π is used as the input to LSTM $\mathscr{L}_\rho$ when Eq. (2) is applied to LSTM $\mathscr{L}_\rho$. The bottom layer consists of LSTM $\mathscr{L}_s$ to embed phrases. Its inputs are each words of the phrases, and the hidden state at the last time step of each phrase is subsequently fed to the next top layer to perform summarization. The top layer consists of LSTM $\mathscr{L}_\rho$ to remove phrase redundancy and realize summarization. The detailed formulation of this hierarchy is presented in the following.

Hierarchical LSTM

The bottom layer embeds phrase i with l_i words denoted as $p^i_{l_i}$, with word embedding implemented by the bidirectional LSTM network $\mathscr{L}_s$. Then, two embedding representations are generated, including both of the forward embedding and backward embedding. Finally, the embedding of the phrase $p^i_{l_i}$ is obtained by adding these two embedding representations. Note that the operation of addition is employed instead of concatenation in order to achieve a more balanced gradient distribution for later stochastic descent computation. In a similar manner, all k phrases can be summarized at the top layer by the bidirectional LSTM network $\mathscr{L}_\rho$. Considering the stability of LSTM memory, we only feed hidden states of $\mathscr{L}_s$ to $\mathscr{L}_\rho$ and do not share memory states between them. This hierarchical structure has a upper-bound memory length of $\max(l_i) + k - 1$ which usually varies from 20 to 40, so that model training can be prevented from the problem of gradient vanishing to a great extent.

Moreover, the top LSTM $\mathscr{L}_\rho$ is supposed to be capable of processing structured information, so that the spatial relation in an image can be transformed to be a sequential series in LSTM. This is realized by ranking the phrases based on their locations and confidential scores, which will be further detailed in Section 4.

Caption Generation

After summarizing all k phrases of an image, we employ the common procedure to generate the final image caption. During model training, the summarized image content ρ is treated as a special word embedding and is placed before the whole sentence to form a word embedding vector $\{w_i\} = \{\rho, W_e\pi_0, W_e\pi_1, W_e\pi_2, \cdots\}$, where W_e is the word embedding matrix shared between $\mathscr{L}_s$ and $\mathscr{L}_g$ (*i.e.*, the caption generation LSTM), π_j stands for the j_{th} word. Specially, π_0 is a placeholder token known as Start of Sentence (SoS). A single linear layer is used to predict w_t based on previous words as

$$w_t = \text{softmax}(W_o h_t + b_o), \tag{3}$$

where W_o and b_o denote the output weight matrix and bias, respectively. During inference, beam search is used to select the top-k best sentences at each time step.

4 Experiment and Discussion

4.1 Implementation Detail

Dataset. The proposed method is evaluated on the benchmark dataset MSCOCO [Lin *et al.*, 2014] which contains

123,287 images, and each of them is annotated with 5 captions. We follow the widely adopted train/val/test split as in [Karpathy and Fei-Fei, 2015], *i.e.*, 5000 images for both validation and testing, and the rest for training.

Training. Faster R-CNN with pre-trained on the Visual Genome dataset [Krishna *et al.*, 2017] is used as our detection model which uses VGG16 for visual feature extraction. Non-Maximum Suppression (NMS) is applied and 30 boxes with the highest predicted confidence scores are extracted for subsequent modules. In the language model, the number of hidden units and the number of factors in each LSTM are all set to 512. A gradient will be clipped if its value exceeds 1. The ADAM optimizer is used for training with $\alpha = 0.8$, $\beta = 0.999$ and $\epsilon = 1 \times 10^{-8}$. The initial learning rate is set to 1×10^{-4} and exponential reduction is used which halves the learning rate every 10 epochs. The training of modality transformation follows the implementation of the dense captioning model [Johnson *et al.*, 2016].

Evaluation. Several commonly used metrics are reported to present caption performances, including BLEU [Papineni *et al.*, 2002], METEOR [Lavie and Denkowski, 2009], ROUGE_L [Lin and Och, 2004] and CIDEr [Vedantam *et al.*, 2015], which are denoted as B, M, R and C for short, respectively. Moreover, beam search is used for caption generation, which iteratively considers 2 best sentences up to the current time step t as candidates to generate the sentences at the time step $t + 1$ and only keeps the best 2 results.

4.2 Performance Comparison

Method	C	B3	B4	M	R
DeepVS [Karpathy and Fei-Fei, 2015]	66.0	32.1	23.0	19.5	–
Google NIC [Vinyals *et al.*, 2015]	–	32.9	24.6	–	–
LRCN [Donahue *et al.*, 2015]	–	30.4	21.0	–	–
Soft Attention [Xu *et al.*, 2015]	–	34.4	24.3	23.9	–
Hard Attention [Xu *et al.*, 2015]	–	35.7	25.0	23.0	–
VNet+ft+LSTM [Wu *et al.*, 2016]	73.0	37.0	25.0	22.0	–
RA+SS [Jin *et al.*, 2017]	83.8	38.1	28.2	23.5	–
Global Attention [Li *et al.*, 2017a]	96.4	41.7	31.2	24.9	53.3
ConvCap [Aneja *et al.*, 2018]	95.2	41.8	31.6	25.0	53.1
Bottom-up [Anderson *et al.*, 2018]	97.1	**42.7**	**32.0**	24.9	**53.5**
Proposed Swell-Shrink	**98.2**	41.8	31.6	**25.1**	53.2

Table 1: Comparison with state-of-the-art methods by employing VGG or GoogLeNet to extract visual features on MSCOCO.

Table 1 shows the comparison of the proposed swell-shrink method with other state-of-the-art methods on the MSCOCO dataset. For a fair comparison, visual features are extracted by VGG or GoogLeNet for all the competing methods. The bottom-up method [Anderson *et al.*, 2018] is particularly compared because it also utilizes RoI with extra training on the Visual Genome dataset. The results show that the proposed swell-shrink method is better than most of the methods except the bottom-up method. When compared with the bottom-up method, the proposed swell-shrink method is inferior on the metrics of B and R while superior on the metrics of C and M. As far as model complexity is concerned, the bottom-up method requires over 60 millions parameters, which are much larger than ours which only needs 10 millions parameters; and this enables the proposed swell-shrink method converge faster.

Besides the comparison to state-of-the-art methods, the $\mathscr{L}_\rho$ LSTM layer of the proposed swell-shrink method is replaced by the following adjustments for further comparison. During the comparison, the CNN model and the number of RoIs are all the same. **Baseline Mean** performs summarization by average pooling to embed all phrases. **Baseline Conv** performs summarization by 4 convolutional layers with batch normalization followed by max pooling. The phrases are co-located as the corresponding RoIs to maintain their spatial relations. **Baseline LSTM** and **Baseline LSTM BS2** perform summarization by the proposed swell-shrink method, however with a random ranking of RoIs, and the RNN model is LSTM instead of bi-LSTM for language modeling. BS2 denotes beam search with the beam size of 2 for testing.

Method	C	B_3	B_4	M	R
Baseline Mean	87.8	37.5	26.6	23.8	51.2
Baseline Conv	89.9	38.4	27.5	24.1	51.9
Baseline LSTM	92.2	38.2	27.3	24.2	51.6
Baseline LSTM BS2	95.2	40.4	30.8	24.7	52.5
Proposed Swell-Shrink	98.2	41.8	31.6	25.1	53.2

Table 2: Comparison with varietal summarization baseline models on MSCOCO.

Figure 3: Subjective quality of generated captions compared with NIC and baseline models. GT stands for ground truth sentence.

Table 2 shows the comparison to other summarization baselines on the MSCOCO dataset. Although the mean-pooling baseline model is poorly performed compared with other baselines, it is still comparable to some state-of-the-art methods across all metrics. While incapable of reorganizing the semantics among multiple phrases, a simple pooling operation still provides some non-structure information for caption generation. the method of "Baseline Conv" is designed for two considerations. On one hand, convolutional operations can preserve spatial relation among regions in a better manner, which is very valuable in this task. On the other hand, CNN has been proven to be powerful on processing natural language tasks. However, the method of "Baseline Conv" shows very similar performances compared to "Baseline LSTM" on all metrics except C, which shows that spatial

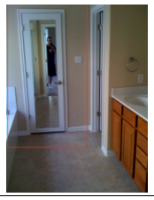
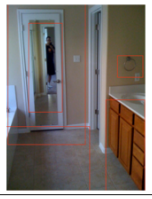
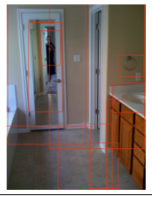
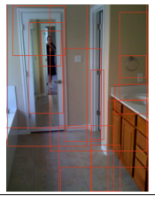
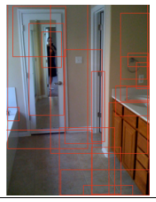
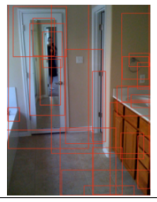
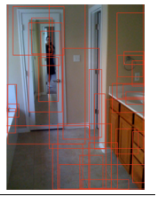
Images with RoIs							
The number of RoIs	1	5	10	15	20	25	30
Coverage area	42%	51%	74%	75%	75%	75%	75%
New phrases	A white door with a door.	a door in a room. a brown tiled floor. wooden cabinets under the sink. a white wall on the wall.	the floor is tiled. a door in a bathroom. white tile on floor. a woman in a black jacket. the sink is white.	white tile on floor. white door with a door. a mirror on the wall. wall is white color. A silver door.	a white sink. a white wall mounted on the wall. white tile on floor. a lamp on the table. a hole in the wall.	the sink is white. a cabinet door. the cabinet is brown. a mirror on the wall. a silver door handle.	a white wall. a white lamp on the wall. a white door. white tile on floor. the floor is tiled.
Sub-caption	a couple of people standing in a room.	a kitchen with a sink and a refrigerator.	a black and white photo of a bathroom.	a bathroom with a sink and a mirror.	a bathroom with a sink and a mirror.	a man standing in a kitchen next to a sink.	a bathroom with a sink and a mirror.
CIDEr metric on all images	43.5	77.8	88.3	91.1	94.0	94.8	95.9

Figure 4: Visualization of caption generation with sub-captions. The correct and incorrect words are denoted in green and red, respectively, and deeper colors indicate the first appearance of these words. The example shows that language model is prone to over-fit the dataset with fewer RoIs while beginning to utilize duplicated phrases when increasing RoIs.

convolutions provide almost no additional boost compared with LSTM based language modeling. It might indicate that spatial information is insignificant for single-sentence captioning, or available spatial information is already contained in the RoI oriented descriptive phrases. Overall, the proposed swell-shrink method achieves competitive results across all metrics, especially on C and M.

A qualitative comparison of generated captions is shown in Fig. 3. From left to right, there are a set of examples with increasing difficulty due to the emergence of obscured, inconspicuous, uncommon objects and scenes. The proposed method generates more precise and specific descriptions at object level, *e.g.*, "bananas and oranges" instead of "fruit", "a couple of people" instead of "a man". However, some attribute information (*e.g.*, yellow, giant) is missing.

4.3 Sub-caption with Partial Region

To make better interpretability of the proposed model, some half-finished "sub-captions" are shown in Fig. 4, which presents the captioning process with increasing regions and the corresponding phrases. Specifically, once training ends, the hidden states derived from the intermediate time steps in the hierarchical LSTM are cutoff and directly fed into the caption generator to accomplish image captioning, then the obtained sentence is considered as sub-caption.

Figure 4 shows an example about a bathroom containing multiple obscured objects and an obfuscated mirror. The first RoI covers more than 40% area of the image, then it reaches 75% with 10 or more RoIs and keeps stable in the following steps despite the number of RoIs is increasing. Surprisingly, there is one quarter of the image area never be covered till the end, which implies a large amount of visual information is totally neglected during language modeling. During the early stages, many new words appear (denoted with deeper color) in the generated phrases along with emerging visual areas. However, these words are not immediately shown in the current sub-caption, but to be postponed. At these early stages, the language model inclines to generate high-frequency words in reference, *e.g.*, kitchen, refrigerator. While at later stages, although the coverage ratio does not increase, the quality of the generated sub-caption (in terms of CIDEr) keeps improving along with the raising number of RoIs. Some phrases seem to be visually redundant and repeatedly describe the same region, which resembles the bottom-up method that assigns weights to different regions in textual form. At these late stages, the language model gains enough phrases and begins to use the semantics from input phrases. Although there are some incorrect contents, *e.g.*, lamp, table and hole, these contents are not employed due to their low frequency.

In general, serried RoIs indicate important regions. Once RoI boxes are sparse, even if covering much area, the language model presents merely poor performances. While the number of RoIs increases, the overlapping area emphasizes important visual regions, and the captioning performance improves consequently. This indicates that serried RoI boxes approximate the mechanism of visual attention for language modeling.

4.4 Impact of RoI Quantity and Order

More phrases convey more contents, however, redundancy and even mistakes with low confidential scores will be brought. Moreover, too many RoIs imply too long dependency in LSTM models, which might result in the problem of

gradient vanishing. Therefore, it is supposed to be a tradeoff between the quantity and the quality of the phrases. Another factor is the order of phrases. LSTM is usually applied to process time series, but the temporal order of phrases in the proposed framework is ambiguous. Different from a sentence or an article whose order is explicitly decided by syntax and logic, the sequential relation between phrases of regions is unclear. To optimize the order of phrases, five different ranking methods are investigated as follows.

SF is short for salient first so that the phrases with higher confidential scores are preferentially fed into LSTM. This arrangement is to verify whether language modeling will benefit from processing visually important regions first, even if they are spatially discrete. **AF** is short for adjacent first so that the order of LSTM time steps is configured according to the ascending distance of regions with each other, while the first input is the region with the highest confidential score. This arrangement is to verify whether the organization of caption conforms to spatial continuity. The third method **Rand** means that phrases are randomly fed into LSTM, which is a simple baseline to eliminate the effect of visual content ordering. Another two methods **Inv_SF** and **Inv_AF** denote the inverse ordering of the corresponding **SF** and **AF**, respectively. It is a trick to relieve the problem of gradient vanishing in LSTM, which is achieved by assigning important information at later time steps so that back propagated gradients can be early received.

Metric	# RoIs	Rand	SF	AF	Inv_SF	Inv_AF
C	10	88.3	89.0	87.9	88.9	89.5
	20	89.8	92.2	90.5	92.1	91.3
	30	91.2	**94.0**	90.9	91.8	92.5
B4	10	26.7	27.1	27.1	27.1	27.5
	20	27.1	27.6	27.4	27.7	27.6
	30	26.9	**28.0**	27.6	27.2	27.8

Table 3: Performance comparison on C and B4 under different number of RoIs and ordering arrangement.

Table 3 shows the CIDEr and BLEU-4 performances of the generated captions under different number of RoIs and ordering arrangement. The captioning performance with 30 RoIs almost outperforms others under all ordering arrangements, except the conditions with 20 RoIs and Inv_SF/random order. The improvement from 20 to 30 RoIs is a little lower than that from 10 to 20. As mentioned before, when the number of RoIs is larger than 10, increasing the number of RoIs will not bring more visual contents, but it produces richer aspects of RoIs and emphasizes the abstractive content by repeated phrases. Generally, quantity beats order to be the dominant factor on caption performance improvement, and order becomes more important when there are more RoIs. Surprisingly, the "Rand" method does not always perform the worst under all circumstances, which proves the significance of order on another side, *i.e.*, incorrect features can be worse than no features. Overall, the "SF" method with 30 RoIs surpasses other arrangements, and it can be concluded that whether visual cues are obtained spatially continuous or discrete, salient regions are desired for sentence generation.

4.5 Transformation-First or Summary-First

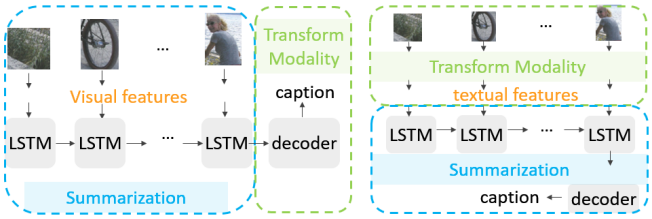


Figure 5: The left and right sub-figures illustrate the summary-first and transformation-first diagrams, respectively.

The proposed swell-shrink method is based on the idea that it is simpler for neural networks like LSTM to summarize texts than visual contents. However, there is an opposite alternative that firstly merges visual regions and then transforms visual representation to textual caption, as illustrated on the left in Fig. 5, while the proposed method is shown on the right. These two arrangements are with almost the same structure and only distinct on the feeding order of the first modality. For implementation, the penultimate layer of VGG16 with 4096 dimensions is treated as the visual representation for each region. Then a fully connected layer is applied to reduce the dimension to 512 which is the same as that of the summary-first arrangement for a fair comparison.

# RoIs	Summary-First		Transformation-First	
	C	B4	C	B4
10	67.6	21.2	89.0	27.1
20	69.6	22.7	92.2	27.6
30	71.0	22.7	94.0	28.0

Table 4: Performance comparison between the summary-first and transformation-first methods with different number of RoIs.

The performance comparison between the summary-first and transformation-first methods is presented in Table 4, where it can be seen that the performance of the summary-first method is similar to [Karpathy and Fei-Fei, 2015], which also attempts to align visual regions with words; more importantly, the transformation-first arrangement predominantly surpasses the summary-first arrangement on both the metrics of C and B4 in a sequential manner.

5 Conclusion

In this work, image captioning is formulated as a composition of modality transformation and text summarization by the proposed swell-shrink method, which firstly divides the whole image into multiple RoIs by Faster R-CNN, then each of the visual regions is transformed to a textual descriptive phrase. After that, a hierarchical LSTM is applied to summarize the main content among these multiple phrases, and finally an abstractive image caption is generated. The proposed method shows better interpretability on analyzing the process of language forming with regional visual information. The experimental results demonstrate the effectiveness of the proposed swell-shrink method.

References

[Anderson *et al.*, 2018] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, Jun. 2018.

[Aneja *et al.*, 2018] J. Aneja, A. Deshpande, and A. G Schwing. Convolutional image captioning. In *CVPR*, pages 5561–5570, Jun. 2018.

[Bahdanau *et al.*, 2015] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, May 2015.

[Cheng and Lapata, 2016] J. Cheng and M. Lapata. Neural summarization by extracting sentences and words. In *ACL*, pages 484–494, Aug. 2016.

[Donahue *et al.*, 2015] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634, Jun. 2015.

[Jin *et al.*, 2017] J. Jin, K. Fu, R. Cui, F. Sha, and C. Zhang. Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. *IEEE Trans. PAMI*, 39(12):2321–2334, Dec. 2017.

[Johnson *et al.*, 2016] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, pages 4565–4574, Jun. 2016.

[Karpathy and Fei-Fei, 2015] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, Jun. 2015.

[Krause *et al.*, 2017] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *CVPR*, pages 3337–3345, Jul. 2017.

[Krishna *et al.*, 2017] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.

[Lavie and Denkowski, 2009] A. Lavie and M. J. Denkowski. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23:105–115, Sep. 2009.

[Li *et al.*, 2017a] L. Li, S. Tang, L. Deng, Y. Zhang, and Q. Tian. Image caption with global-local attention. In *AAAI*, pages 4133–4139, Feb. 2017.

[Li *et al.*, 2017b] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang. Scene graph generation from objects, phrases and region captions. In *ICCV*, pages 1261–1270, Oct. 2017.

[Liang *et al.*, 2017] X. Liang, Z. Hu, H. Zhang, C. Gan, and E. P. Xing. Recurrent topic-transition GAN for visual paragraph generation. In *ICCV*, pages 3362–3371, Oct. 2017.

[Lin and Och, 2004] C. Lin and F. J. Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *ACL*, pages 605–612, Jul. 2004.

[Lin *et al.*, 2014] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, Sep. 2014.

[Miao and Blunsom, 2016] Y. Miao and P. Blunsom. Language as a latent variable: Discrete generative models for sentence compression. In *EMNLP*, pages 319–328, Nov. 2016.

[Nallapati *et al.*, 2016] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *SIGNLL CCNLL*, pages 280–290, Aug. 2016.

[Nallapati *et al.*, 2017] R. Nallapati, F. Zhai, and B. Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI*, pages 3075–3081, Feb. 2017.

[Papineni *et al.*, 2002] K. Papineni, S. Roukos, T. Ward, and W. Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, Jul. 2002.

[Rush *et al.*, 2015] A. M. Rush, S. Chopra, and J. Weston. A neural attention model for abstractive sentence summarization. In *EMNLP*, pages 379–389, Sep. 2015.

[Vedantam *et al.*, 2015] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, Jun. 2015.

[Vinyals *et al.*, 2015] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, Jun. 2015.

[Wu and Hu, 2018] Y. Wu and B. Hu. Learning to extract coherent summary via deep reinforcement learning. In *AAAI*, pages 5602–5609, Feb. 2018.

[Wu *et al.*, 2016] Q. Wu, C. Shen, A. Liu, L. Dick, and A. van den Hengel. What value do explicit high level concepts have in vision to language problems? In *CVPR*, pages 203–212, Jun. 2016.

[Xu *et al.*, 2015] K. Xu, J. Lei Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, Jul. 2015.

[Yang *et al.*, 2017] L. Yang, K. D. Tang, J. Yang, and L. Li. Dense captioning with joint inference and visual context. In *CVPR*, pages 1978–1987, Jul. 2017.

[Yin and Pei, 2015] W. Yin and Y. Pei. Optimizing sentence modeling and selection for document summarization. In *IJCAI*, pages 1383–1389, Aug. 2015.

[Zeiler and Fergus, 2014] M.D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833, Sep. 2014.