Journal of Southeast University (English Edition) Vol. 32, No. 2, pp. 170 - 176 June 2016 ISSN 1003—7985

Product image sentence annotation based on kernel descriptors and tag-rank

Zhang Hongbin[1,2] Ji Donghong[1] Yin Lan[1] Ren Yafeng[1] Yin Yi[2]

([1]Computer School, Wuhan University, Wuhan 430072, China)

([2]School of Software, East China Jiaotong University, Nanchang 330013, China)

Abstract: Dealing with issues such as too simple image features and word noise inference in product image sentence amnotation, a product image sentence annotation model focusing on image feature learning and key words summarization is described. Three kernel descriptors such as gradient, shape, and color are extracted, respectively. Feature late-fusion is executed in turn by the multiple kernel learning model to obtain more discriminant image features. Absolute rank and relative rank of the tag-rank model are used to boost the key words' weights. A new word integration algorithm named word sequence blocks building (WSBB) is designed to create *N*-gram word sequences. Sentences are generated according to the *N*-gram word sequences and predefined templates. Experimental results show that both the BLEU-1 scores and BLEU-2 scores of the sentences are superior to those of the state-of-art baselines.

Key words: product image; sentence annotation; kernel descriptors; tag-rank; word sequence blocks building (WSBB); *N*-gram word sequences

doi: 10.3969/j.issn.1003-7985.2016.02.007

Isolated words are usually annotated on images, which narrows the "semantic gap" between people's high-level cognitions and low-level image features. However, words remains independent so that the important combined semantic information between words is lost. More seriously, noisy words which affect people's high-level cognition about the images' content occur frequently. However, a sentence possesses both concise and unambiguous combined semantic information so that it can describe the image's content more accurately than those isolated words. Therefore, image sentence annotation has very promising applications such as a image retrieval system based on semantic information, a visual perception aided system for the blind, and the pilotless automobile based on real-time road monitoring etc. These applications will certainly bring people much more convenience than before.

Received 2015-09-12.

Biography: Zhang Hongbin (1979—), male, doctor, associate professor, zhanghongbin@whu.edu.cn.

Foundation items: The National Natural Science Foundation of China (No. 61133012), the Humanity and Social Science Foundation of the Ministry of Education (No. 12YJCZH274), the Humanity and Social Science Foundation of Jiangxi Province (No. XW1502, TQ1503), the Science and Technology Project of Jiangxi Science and Technology Department (No. 20121BBG70050, 20142BBG70011).

Citation: Zhang Hongbin, Ji Donghong, Yin Lan, et al. Product image sentence annotation based on kernel descriptors and tag-rank[J]. Journal of Southeast University (English Edition), 2016, 32(2): 170 - 176. doi: 10.3969/j.issn.1003-7985.2016.02.007.

1 Related Work

Image sentence annotation has three kinds of methods such as retrieval[1-2], generation[3-4], and summarization[5-10]. Farhadi et al.[1-2] analyzed the semantic correlations between images and texts to retrieve the best sentence for annotation. However, useful semantic information better at describing the image often crosses several sentences. It means that only one sentence cannot depict an image accurately. Yang et al.[3-4] generated a sentence according to the outputs of their object recognition system. Nevertheless, an object recognition system usually produces many noisy outputs, which affect annotation performance heavily. Ushiku et al.[5-7] summarized the correlated texts from the training images' captions to create a sentence. However, traditional features are too simple for capturing the correlated texts. Currently, product image sentence annotation attracts more and more attentions from different research fields including the computer vision and the natural language generation. Berg et al.[8] annotated a product image using several visual attributes. Kiapour et al.[9] tagged a product image using many fashion elements. However, text fragments[8-9] cannot describe a product image's content completely. Mason[10] retrieved the key correlated texts from training images' captions by gist. Gist focuses on the texture feature but it is too simple to describe the product image's content. The cross-media correlations between image features generated by the convolutional neural network (CNN) and word vectors are analyzed by the modality-biased log-bilinear language (MLBL) model[11]. A sentence is generated according to the correlations. CNN is so complex that the model easily sinks into overfitting. Therefore, a robust and effective annotation model must be proposed. The model must better describe the key visual characteristics of the product image as well as better summarize the key correlated words from training images' captions. The research innovations of this paper include:

1) Four new image features are achieved by fusing dif-

ferent kernel descriptors (KDES)[12]. MK-KDES-1 achieves the best annotation performance among all the features.

2) Two new word features are extracted by transplanting the tag-rank model[13] into the word relevance computing model. The two features improve annotation performance significantly.

3) A new word integration algorithm named word sequence blocks building (WSBB) is proposed to create valuable *N*-gram word sequences for sentence generation.

2 Presented Annotation Model

2.1 Image feature learning

To describe product images more accurately, KDES[12] features are extracted from three visual aspects such as gradient, shape, and color. New image features named MK-KDES-J ($J = 1, 2, \ldots, 4$) are created by fusing these KDES features in the multiple kernel learning (MKL) model. The product's class label is obtained in turn for sentence generation.

2.2 Word relevance computing

Training images are retrieved based on the MK-KDES-J feature. K key words like $\{\mathrm{wrd}_1, \mathrm{wrd}_2, \ldots, \mathrm{wrd}_K\}$ are summarized by the following word relevance computing model:

$$p(\mathrm{wrd}_j \mid I_q) = \log_2\Big(\sum_i p(\mathrm{wrd}_j \mid I_i)\, p(I_i \mid I_q)\Big) = \log_2\Bigg(\sum_i \frac{\mathrm{word\_fea}(\mathrm{wrd}_j)}{\sum\limits_{\mathrm{wrd}_j \in W'} \mathrm{word\_fea}(\mathrm{wrd}'_j)} \cdot \frac{\exp(-\mathrm{dist\_fun}(I_i \mid I_q))}{\sum\limits_{I_{ii} \in I'} \exp(-\mathrm{dist\_fun}(I_{ii} \mid I_q))}\Bigg) \tag{1}$$

where I_i denotes the training image; I_q denotes the testing image; wrd_j is the key word summarized from the training images' captions; W' is the word set; I' is the training image set; $p(\mathrm{wrd}_j \mid I_i)$ computes the semantic relevance score between wrd_j and I_i, and a higher score means higher importance of the word; $p(I_i \mid I_q)$ computes the visual similarity between I_i and I_q; word_fea represents the text feature of wrd_j; dist_fun computes the visual distance between two images. Finally, K key words with the highest relevance scores are summarized for sentence generation.

A tag-rank model[13] is absorbed into the word relevance computing model to better summarize the key correlated words. The model contains two metrics. One is absolute rank (AR) which evaluates a word's importance according to its absolute position in a sentence. The other is relative rank (RR) which evaluates a word's importance according to its relative position in a sentence. They are defined as

$$\mathrm{AR}(j) = \begin{cases} 0 & \mathrm{wrd}_j \notin s \\ \dfrac{1}{\log_2(1 + \mathrm{pos}_j)} & \mathrm{wrd}_j \in s \end{cases} \tag{2}$$

$$\mathrm{RR}(j) = \begin{cases} 0 & \mathrm{wrd}_j \notin s \\ 1 - \dfrac{\sum\limits_{k=1}^{\mathrm{pos}_j} n_{jk}}{n_j} & \mathrm{wrd}_j \in s \end{cases} \tag{3}$$

where $\mathrm{AR}(j)$ indicates that the weight of each word is based on its absolute position; pos_j is the average value calculated by all wrd_j's absolute positions in a sentence s; $\mathrm{RR}(j)$ indicates what percent of wrd_j's occurrences appear after pos_j; n_{jk} is the number of times that wrd_j appears in position k; and $n_j = \sum_k n_{jk}$ is the total occurrence frequency of wrd_j.

2.3 Word integration

Product image is often described by word sequences composed of several adjectives or nouns. A new word integration algorithm named WSBB is designed to create new word sequences. As is known, in children's block-building games, buildings are constructed by different kinds of blocks. These blocks are piled up based on their shapes overlapping between them. Therefore, the key correlated words summarized by the word relevance computing model are regarded as blocks. These words are piled up together based on the semantic overlapping between them. Final word sequences are regarded as buildings. Therefore, two definitions are proposed to better interpret the new word integration algorithm.

Definition 1 Semantic overlapping is defined as semantic relevance such as co-occurrence frequency or content overlapping between two word sequences.

Definition 2 N-gram word sequence is defined as a word sequence that contains N modified words such as adjectives or nouns ($N = 1, 2, \ldots, 4$).

Fig. 1 illustrates the new word integration algorithm. The 3-gram ($N = 3$) word sequence like "luxurious croco embossing" is generated easily. "Luxurious croco embossing" is created from bottom to up recursively. For example, if the co-occurrence frequency which is evaluated by the TF function between two 2-gram word sequences exceeds a threshold or there is content overlapping which is evaluated by the overlap function between two 2-gram word sequences, it means that the semantic overlapping between these 2-gram word sequences is prominent. A new 3-gram word sequence must be generated. On the contrary, if the semantic overlapping between two 2-gram word sequences is less prominent, the 2-gram word sequence with the highest semantic relevance score is output by the WSBB. The semantic relevance score between a new word sequence and a testing image is computed as

$$p(\text{seq} \mid I_q) = \log_2\left(\prod_{j=1}^{N} p(\text{wrd}_j \mid I_q)\right) \quad (4)$$

Word sequences are ranked in descend order according to the semantic relevance scores. Top M word sequences are chosen for sentence generation ($M = 1, 2, \ldots$).

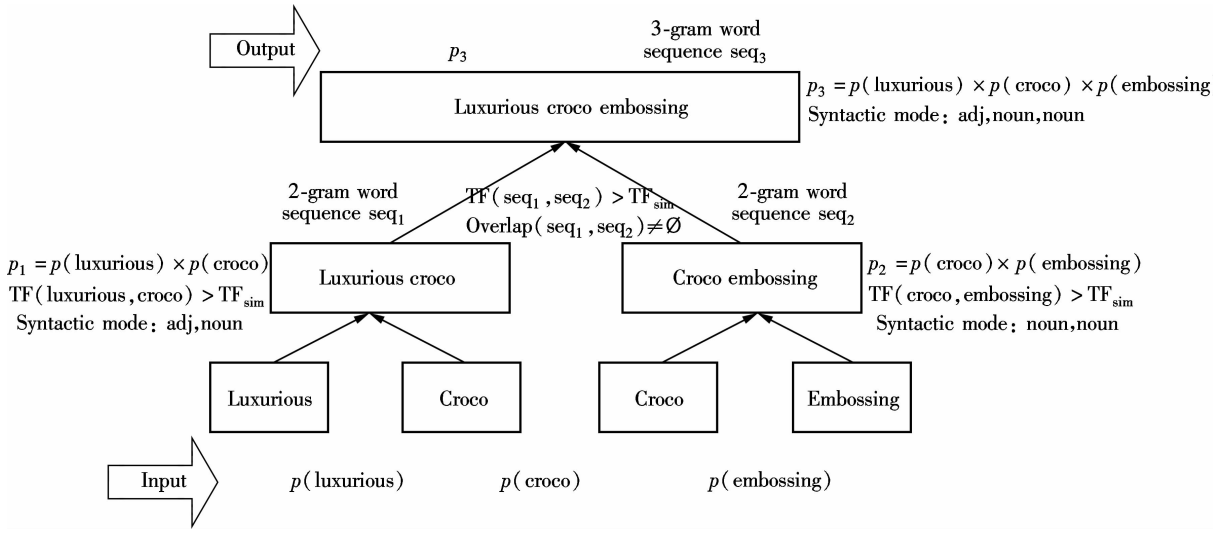

Fig. 1 A new 3-gram word sequence generated by WSBB

2.4 Sentence generation

The top M word sequences are output by WSBB. They are pushed into predefined templates to generate sentences. The virtues of templates include readability and conciseness. The general template is defined as "prefix phrase + core phrase + suffix phrase". The "prefix phrase" is defined as "This is a picture of a". The "core phrase" is created by arranging top M word sequences sequentially. The "suffix phrase" is defined as the product's class label obtained from image classification. In addition, a boosting strategy is designed to obtain a more accurate sentence. The sentence with the highest BLEU-3 score is chosen as the final annotation because both the semantic relevance and syntactic coherence of a sentence are evaluated by the BLEU-3 metric.

3 Experiments and Discussion

3.1 Dataset and baselines

Bags are the representative online products on E-commerce websites. Therefore, bags in attribute dataset[8] are selected to evaluate the presented annotation model. The dataset contains five kinds of products: clutch, evening, shoulder, hobo, and totes. The sample number is 8 127. 70% samples are chosen randomly as the training set and the rest of the samples are chosen as the testing set. The state-of-art baselines include beam search (based on the 3-gram language model)[6], gist-based[10], and MLBL[11].

3.2 Image feature learning

Gradient-KDES, shape-KDES, and color-KDES are extracted, respectively. These features are fused together by MKL to obtain MK-KDES-J ($J = 1, 2, \ldots, 4$) feature. Fusion results are shown in Tab. 1.

Tab. 1 Fusion results of different KDES features

New features	Gradient-KDES weight	Shape-KDES weight	Color-KDES weight	Classification accuracy/%
MK-KDES-1	0.68	0.32		87.65
MK-KDES-2	0.08		0.92	92.70
MK-KDES-3		0.02	0.98	92.86
MK-KDES-4	0.01	0.01	0.98	93.20

3.3 Quantitative evaluations of sentence annotation

A sentence is generated by randomly assembling K key words summarized by the word relevance computing model. The first experiment only tries to find the best image feature for sentence generation rather than to evaluate the new word feature AR (or RR) and the WSBB algorithm. Therefore, the traditional word feature named TF-IDF is used to compute the semantic relevance scores between words and images. The mean value of BLEU scores for each K is calculated and two BLEU score histograms are acquired as shown in Fig. 2.

MK-KDES-1 feature obtains the best annotation performance. It means that shape and texture (its nature is gradient variation) are the key visual characteristics of product image. For example, different bags have different shapes and different textures. MK-KDES-1 feature focuses on the visual characteristics of shape and gradient interprets the product image's content better than other features. Fig. 2 also informs us that people prefer to describe the product images' content from two aspects such as shape and texture.

Meanwhile, the word relevance computing model is designed to summarize the key words for better describing

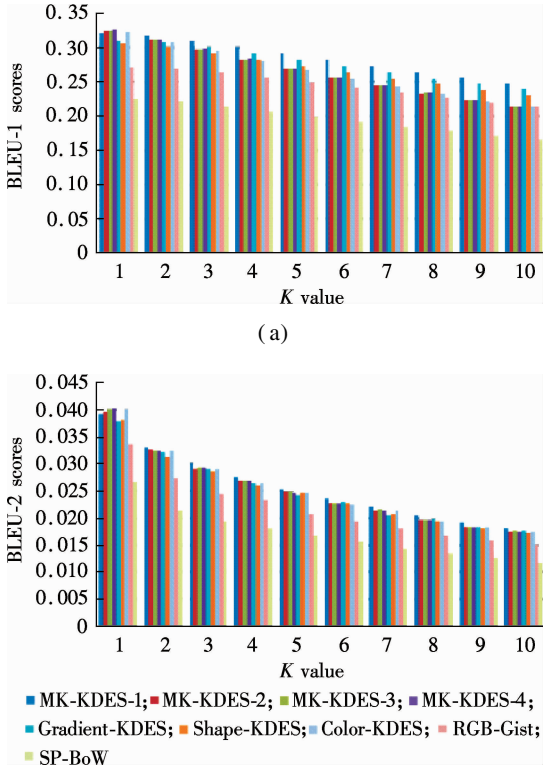

Fig. 2 Image feature selection. (a) BLEU-1 scores of each image feature; (b) BLEU-2 scores of each image feature

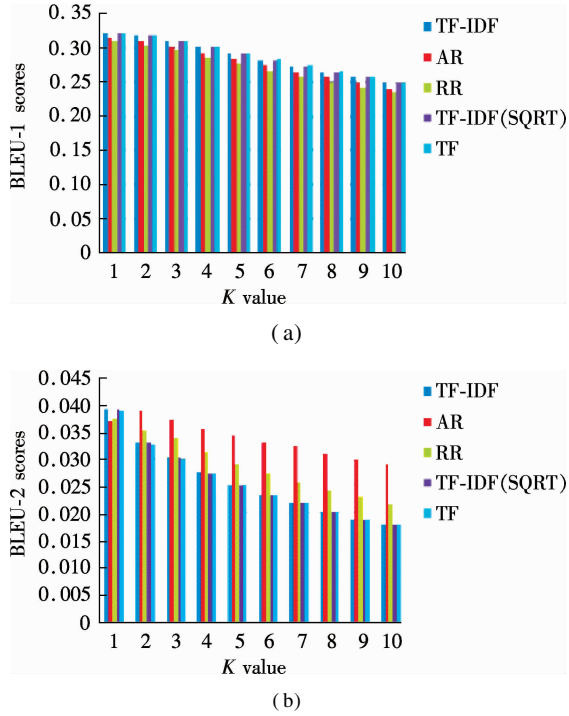

Fig. 3 Word feature selection. (a) BLEU-1 scores of each word feature; (b) BLEU-2 scores of each word feature

the image's content. Five word features including TF-IDF, TF, TF-IDF (SQRT), AR, and RR are evaluated fairly in the second experiment. The experimental procedure is similar to that of image feature selection. The mean value of BLEU scores for each K is calculated and two BLEU score histograms are acquired as shown in Fig. 3.

TF-IDF achieves the best BLEU-1 scores. This indicates that TF-IDF is good at content selection. To our surprise, the BLEU-1 scores of AR approximate those of TF-IDF. More importantly, AR achieves the best BLEU-2 scores. Especially, the BLEU-2 scores of AR are significantly superior to other features when $K > 5$. For example, the BLEU-2 score of AR is nearly 1.41 times that of TF-IDF when $K = 6$. The superiority keeps increasing with the growth of K value. Similar phenomena also occur in BLEU-3 evaluations. The results indicate that AR can greatly improve annotation performance. AR considers the absolute positions of words, and the key adjectives or nouns are often written at the beginning of a sentence by annotators, so AR assigns higher weights to these words. However, the RR weights are less important than the AR weights. Therefore, AR is the first choice for the key words summarization.

The word integration algorithm is taken into account in the third experiment. First, different word features (WF) such as TF-IDF, AR, and RR are assembled with WSBB to create different annotation models named WSBB-J (J is the model index). When TF-IDF is chosen as WF, WSBB-1 ($N = 1$), WSBB-2 ($N = 2$), WSBB-3 ($N = 3$), and WSBB-4 ($N = 4$) are achieved. When RR is chosen as WF, WSBB-5 ($N = 1$), WSBB-6 ($N = 2$), WSBB-7 ($N = 3$), and WSBB-8 ($N = 4$) are achieved. When AR is chosen as WF, WSBB-9 ($N = 1$), WSBB-10 ($N = 2$), WSBB-11 ($N = 3$), and WSBB-12 ($N = 4$) are achieved. Secondly, four boosting models are created by incorporating different WSBB models together. Boost-1 model is the late-fusion of TF-IDF and RR. Boost-2 model is the late-fusion of TF-IDF and AR. Boost-3 model is the late-fusion of RR and AR. Boost-4 model is the late-fusion of TF-IDF, AR, and RR. In the late-fusion procedure, only the corresponding 3-gram models and 4-gram models are chosen and fused together. For example, the Boost-1 model fuses WSBB-3, WSBB-4, WSBB-7, and WSBB-8. The BLEU scores of different models are shown in Tab. 2 ($M = 1$ for fair comparisons).

With the increase of N value, the BLEU-1 scores decrease rapidly when TF-IDF feature is chosen. This is due to the noisy words in texts. However, the BLEU-2 scores are improved rapidly after using the 3-gram (or 4-gram) word sequences generated by the presented WSBB algorithm. Similar phenomena also occur in BLEU-3 evaluations. AR beats any other feature in both BLEU-2 and BLEU-3 evaluations. As expected, the results are consistent with Fig. 3. More importantly, the BLEU-2 scores in Tab. 2 are all superior to those in Fig. 3 (b) due to the introduction of the WSBB algorithm. For example, the BLEU-2 score of WSBB-8 model is nearly 1.35 times the best result in Fig. 3(b). It means that both the AR feature and the WSBB algorithm help to improve annotation per-

Tab. 2 The BLEU scores of each model

Models	BLEU-1	BLEU-2	BLEU-3	Models	BLEU-1	BLEU-2	BLEU-3
WSBB-1	0.321 9	0.038 4	0.003 7	WSBB-9	0.312 2	0.039 1	0.003 7
WSBB-2	0.316 5	0.050 0	0.005 5	WSBB-10	0.313 0	0.052 1	0.006 3
WSBB-3	0.312 2	0.050 8	0.007 1	WSBB-11	0.312 0	0.052 9	0.008 0
WSBB-4	0.309 4	0.052 3	0.008 8	WSBB-12	0.311 6	0.054 4	0.009 5
WSBB-5	0.306 8	0.035 5	0.003 3	Boost-1	0.311 7	0.056 6	0.011 3
WSBB-6	0.310 1	0.049 3	0.005 0	Boost-2	0.313 2	0.057 0	0.011 6
WSBB-7	0.309 3	0.051 9	0.007 5	Boost-3	0.312 8	0.056 9	0.011 8
WSBB-8	0.309 4	0.052 8	0.007 9	Boost-4	0.309 4	0.058 3	0.013 1

formance. Meanwhile, image feature learning is another key factor for optimizing the final performance. In addition, the boosting model is superior to the corresponding individual model in both BLEU-2 and BLEU-3 evaluations. For example, the BLEU-2 scores of the Boost-1 model and its corresponding individual models are ranked as Boost-1, WSBB-8, WSBB-4, WSBB-7, WSBB-3. This indicates that our boosting strategy is helpful. As expected, the Boost-4 model achieves the best BLEU-2 score as well as the best BLEU-3 score among all boosting models, which means that different word features such as TF-IDF, RR, and AR are complementary mutual. Another interesting result is that AR accounts for 96.84% sentences, RR for 2.01% sentences, and TF-IDF only for 1.15% sentences in the generated results of the Boost-4 model. This also means that AR plays the most important role in the boosting procedure. Similar phenomena also occur in other boosting models.

Finally, the presented model is compared with the state-of-art baselines including beam search (based on 3-gram language model)[6], gist-based[10], and MLBL[11] in Fig. 4. The presented annotation model performs well in both BLEU-1 and BLEU-2 evaluations. Compared with MLBL and beam search, the presented annotation model is robust due to the fact that it is a nearly non-parametric model which cannot sink into overfitting. Though the better BLEU-1 score and BLEU-2 score are achieved, the BLEU-3 score of the presented annotation model still has a large lifting space. All in all, the presented model is both robust and effective to some extent.

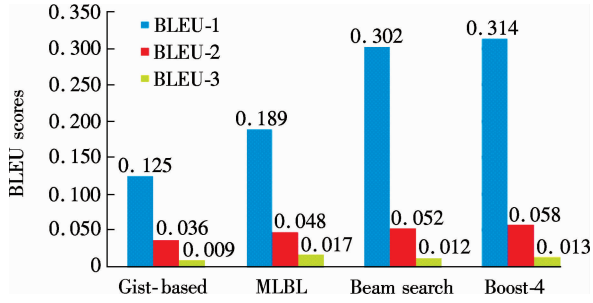

Fig. 4 The comparisons of BLEU scores with baselines

3.4 Qualitative measurement

Some good sentences generated by the presented annotation model are listed in Tab. 3.

Tab. 3 Some good sentences

Images	Boost-4's annotations	Original annotations
	1) This is a picture of a pvc trim hobo bag.	1) Metal mesh hobo with PVC trim.
	2) This is a picture of a classic metal mesh clutch bag.	2) Three different types of metal mesh classic framed clutch.
	3) This is a picture of a classic black print leather shoulder bag.	3) Deep color, slouchy supple leather, and functional details give tailored luxury a modern feel.
	4) This is a picture of a luxurious croco embossing totes bag.	4) The perfect handbag for everyday, this croco-embossed leather tote features: 12.5×12×5.5.
	5) This is a picture of a satin bow clutch bag.	5) A satiny evening bag in the shape of a bow, accented with a square rhinestone medallion.

First, our sentences are accurate. For example, the word sequences like "classic metal mesh" and "classic black print leather" describe the products' texture correctly. Secondly, our sentences are unambiguous. Different from isolated words, the above sentences composed of the correlated key combined semantic information such as "pvc trim" and "classic metal mesh" may depict products' characteristics unambiguously. Thirdly, our sentences are more concise. *N*-gram word sequences constructed by the key correlated words are adequate to describe product images' content. Finally, our sentences are more interesting. For example, "satin bow", "classic metal mesh",

and "luxurious croco embossing" all sound very interesting. They appear to have been written by people.

Certainly, many errors also remain in some sentences. Some poor sentences generated by the presented annotation model are listed in Tab. 4.

Tab. 4 Some poor sentences

Images	Boost-4's annotations	Original annotations
	1) This is a picture of a silk satin clutch bag.	1) Madison Brook handheld clutch purple-manmade handbags.
	2) This is a picture of a Jessica Simpson hobo bag.	2) The Gin Lane hobo handbag by Jessica Simpson.
	3) This is a picture of a black nylon hobo bag.	3) Yellow leather hobo with fringe and metallic star detail.
	4) This is a picture of a bold color combinations totes bag.	4) Vivid polka dots in bold color combinations help these totes make a statement.
	5) This is a picture of a nero black leather shoulder bag.	5) This patchwork leather shoulder bag in brown multi features: 16.5 ×4 ×14.

Many poor annotations occur in Tab. 4 due to three reasons. The first is the noisy words in text. For example, although "Jessica Simpson" and "bold color combinations" are generated correctly by the WSBB algorithm, these word sequences all focus on the non-visual characteristics of product images while the correlated visual descriptions are missed due to lower text weights. The second is the visual words ambiguities[14]. For example, the texture word of "black nylon" and that of "metallic star detail" fall into ambiguity. This finally confuses the annotation model. The last is the natural characteristic of the MK-KDES-1 feature. As shown in Tab. 1, the feature mainly describes the key texture and shape characteristics of product images while it ignores the color characteristics. Therefore, the word sequence like "nero black leather" contains a wrong word "black". Objectively speaking, although better annotation performances are achieved by both quantitative and qualitative evaluations, the presented annotation model still needs to be modified to resolve the above problems.

4 Conclusion

A robust and effective annotation model is introduced. First, the MK-KDES-1 feature is created by the KDES model and the MKL model. The feature obtains the best annotation performance. Secondly, the annotation model focuses on summarizing the key correlated words that have the tightest semantic relations with the product images. K key words are achieved by transplanting the tag-rank model including AR metric and RR metric into the word relevance computing model. The AR feature can greatly improve annotation performance. Thirdly, top M valuable N-gram word sequences are created by the presented WSBB algorithm. Fourthly, sentences are generated according to the N-gram word sequences and predefined templates. Finally, a useful boosting strategy is designed to boost annotation performance. Experimental results show how the presented model beats the state-of-art baselines, particularly, in the BLEU-1 and BLEU-2 evaluations. Most of our sentences are correct, unambiguous, concise, and interesting. More importantly, the presented annotation model is almost a non-parametric model which cannot sink into overfitting. In future, we will focus on overcoming the visual words ambiguities by semantic contexts[14] and introducing a syntactic generation tree to create more coherent sentences.

References

[1] Farhadi A, Hejrati M, Sadeghi M A, et al. Every picture tells a story: Generating sentences from images[C]//*European Conference on Computer Vision*. Berlin: Springer-Verlag, 2010: 15 - 29.

[2] Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: Data, models and evaluation metrics [J]. *Journal of Artificial Intelligence Resource*, 2013, **47**(1): 853 - 899.

[3] Yang Y, Teo C L, Daume H, et al. Corpus-guided sentence generation of natural images[C]//*Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK, 2011: 444 - 454.

[4] Kulkarni G, Premraj V, Dhar S, et al. Baby talk: Understanding and generating simple image descriptions [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**(12): 2891 - 2903. DOI: 10.1109/TPAMI.2012.162.

[5] Ushiku Y, Harada T, Kuniyoshi Y. Automatic sentence generation from images [C]//*Proceedings of the* 19*th ACM International Conference on Multimedia*. New York: ACM, 2011: 1533 - 1536.

[6] Feng F, Lapata M. Automatic caption generation for news images [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**(4): 797 - 812. DOI: 10.1109/TPAMI.2012.118.

[7] Gupta A, Verma Y, Jawahar C V, et al. Choosing lin-

guistics over vision to describe images[C]//*American Association for Artificial Intelligence*. Palo Alto, CA, USA: Association for the Advancement of Artificial Intelligence, 2012: 606 - 611.

[8] Berg T L, Berg A C, Shih J. Automatic attribute discovery and characterization from noisy web data[C]//*European Conference on Computer Vision*. Berlin: Springer, 2010: 663 - 676.

[9] Kiapour H, Yamaguchi K, Berg A C, et al. Hipster Wars: Discovering elements of fashion styles[C]//*European Conference on Computer Vision*. Zurich, Switzerland, 2014: 472 - 488.

[10] Mason R. Domain-independent captioning of domain-specific images[C]//*North American Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics Publication, 2013: 69 - 76.

[11] Kiros R, Salakhutdinov R, Zemel R. Multimodal neural language models[C]//*International Conference on Machine Learning*. Beijing, China, 2014: 595 - 603.

[12] Bo L, Ren X, Fox D. Kernel descriptors for visual recognition[C]//*Advances in Neural Information Processing Systems*. Vancouver, Canada, 2010: 1734 - 1742.

[13] Hwang S, Grauman K. Learning the relative importance of objects from tagged images for retrieval and cross-modal search[J]. *International Journal of Computer Vision*, 2012, **100**(2): 134 - 153. DOI: 10.1007/s11263-011-0494-3.

[14] Su Y, Jurie F. Visual word disambiguation by semantic contexts[C]//*IEEE International Conference on Computer Vision*. Barcelona, Spain, 2011: 311 - 318.

基于核特征和 tag-rank 的商品图像句子标注

张红斌[1,2] 姬东鸿[1] 尹 兰[1] 任亚峰[1] 殷 依[2]

([1] 武汉大学计算机学院,武汉 430072)

([2] 华东交通大学软件学院,南昌 330013)

摘要:针对商品图像句子标注中图像特征单一、关键词受噪声干扰等问题,提出一种聚焦图像特征学习和关键词摘取的商品图像句子标注模型. 从梯度、形状和颜色 3 个角度抽取图像核特征,并在多核学习模型内进行后融合. 利用 tag-rank 模型中的绝对排序和相对排序特征提升关键词权重,设计词序列拼积木算法把关键词拼装成 *N* 元词序列. 基于 *N* 元词序列和模板生成句子. 实验表明:句子的 BLEU-1 和 BLEU-2 评分优于对比模型.

关键词:商品图像;句子标注;核特征;tag-rank; 词序列拼积木;*N* 元词序列

中图分类号:TP391