Rich Image Captioning in the Wild

Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun
Cornelia Carapcea, Chris Thrasher, Chris Buehler, Chris Sienkiewicz

Microsoft Research

{ktran,xiaohe}@microsoft.com*

Abstract

We present an image caption system that addresses new challenges of automatically describing images in the wild. The challenges include generating high quality caption with respect to human judgments, out-of-domain data handling, and low latency required in many applications. Built on top of a state-of-the-art framework, we developed a deep vision model that detects a broad range of visual concepts, an entity recognition model that identifies celebrities and landmarks, and a confidence model for the caption output. Experimental results show that our caption engine outperforms previous state-of-the-art systems significantly on both in-domain dataset (i.e. MS COCO) and out-of-domain datasets. We also make the system publicly accessible as a part of the Microsoft Cognitive Services.

1. Introduction

Image captioning is a fundamental task in Artificial Intelligence which describes objects, attributes, and relationship in an image, in a natural language form. It has many applications such as semantic image search, bringing visual intelligence to chatbots, or helping visually-impaired people to see the world around them. Recently, image captioning has received much interest from the research community (see [24, 25, 26, 6, 7, 13, 11]).

The leading approaches can be categorized into two streams. One stream takes an end-to-end, *encoder-decoder* framework adopted from machine translation. For instance, [24] used a CNN to extract high level image features and then fed them into a LSTM to generate caption. [25] went one step further by introducing the attention mechanism. The other stream applies a *compositional* framework. For example, [7] divided the caption generation into several parts: word detector by a CNN, caption candidates generation by a maximum entropy model, and sentence re-ranking by a deep multimodal similarity model.

"Sasha Obama, Malia Obama, Michelle Obama, Peng Liyuan et al. posing for a picture with Forbidden City in the background."

"A small boat in Ha-Long Bay."

Figure 1: Rich captions enabled by entity recognition

However, while significant progress have been reported [26, 24, 6, 7], most of the systems in literature are evaluated on academic benchmarks, where the experiments are based on test images collected under a controlled environment which have similar distribution to the training examples. It is unclear how these systems perform on open-domain images.

Furthermore, most of the image captioning systems only describe generic visual content without identifying key entities. The entities, such as celebrities and landmarks, are important pieces in our common sense and knowledge. In

*Corresponding authors

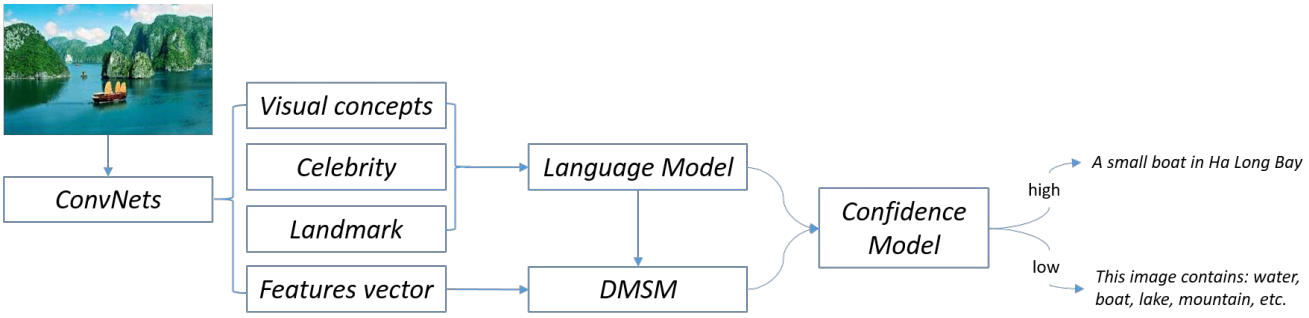

Figure 2: Illustration of our image caption pipeline.

many situations (e.g., Figure 1), the entities are the key information in an image.

In addition, most of the literature report results in automatic metrics such as BLEU [19], METEOR [1], and CIDEr [23]. Although these metrics are handy for fast development and tuning, there exists a substantial discrepancy between these metrics and human's judgment [5, 15, 4]. Their correlation to humans judgment could be even weaker when evaluating captions with entity information integrated.

In this paper, we present a captioning system for open domain images. We take a compositional approach by starting from one of the state-of-the-art image captioning framework [7]. To address the challenges when describing images in the wild, we enriched the visual model by detecting a boarder range of visual concepts and recognizing celebrities and landmarks for caption generation (see examples in Figure 1). Further, in order to provide graceful handling for images that are difficult to describe, we built a confidence model to estimate a confidence score for the caption output based on the vision and text features, and provide a back-off caption for these difficult cases. We also developed an efficient engine that integrates these components and generates the caption within one second end-to-end on a 4-core CPU.

In order to measure the quality of the caption from the humans perspective, we carried out a series of human evaluations through crowd souring, and report results based on human's judgments. Our experimental results show that the proposed system outperforms a previous state-of-the-art system [7] significantly on both in-domain dataset (MS COCO [16]), and out-of-domain datasets (Adobe-MIT FiveK [3] and a dataset consisting randomly sampled images from Instagram [1].) Notably, we improved the human satisfaction rate by 94.9% relatively on the most challenging Instagram dataset.

[1] Instagram data: https://gist.github.com/zer0n/061d6c5e0cb80b56d0a3

2. Model architecture

Following Fang et al. [7], we decomposed the image caption system into independent components, which are trained separately and integrated in the main pipeline. The main components include

- a deep residual network-based vision model that detects a broad range of visual concepts,
- a language model for candidates generation and a deep multimodal semantic model for caption ranking,
- an entity recognition model that identifies celebrities and landmarks,
- and a classifier for estimating the confidence score for each output caption.

Figure 2 gives an overview of our image captioning system.

2.1. Vision model using deep residual network

Deep residual networks (ResNets) [12] consist of many stacked "Residual Units". Each residual unit (Fig. 3) can be expressed in a general form:

$$\mathbf{y}_l = h(\mathbf{x}_l) + \mathcal{F}(\mathbf{x}_l, \mathcal{W}_l),$$
$$\mathbf{x}_{l+1} = f(\mathbf{y}_l),$$

where $\mathbf{x}_l$ and $\mathbf{x}_{l+1}$ are input and output of the l-th unit, and $\mathcal{F}$ is a residual function. In [12], $h(\mathbf{x}_l) = \mathbf{x}_l$ is an identify mapping and f is a ReLU [18] function. ResNets that are over 100-layer deep have shown state-of-the-art accuracy for several challenging recognition tasks on ImageNet [20] and MS COCO [17] competitions. The central idea of ResNets is to learn the additive residual function $\mathcal{F}$ with respect to $h(\mathbf{x}_l)$, with a key choice of using an identity mapping $h(\mathbf{x}_l) = \mathbf{x}_l$. This is realized by attaching an identity skip connection ("shortcut").

Training. In order to address the open domain challenge, we trained two classifiers. The first classifier was trained on MS COCO training data, for 700 visual concepts. And the second one was trained on an image set crawled from commercial image search engines, for 1.5K visual objects.

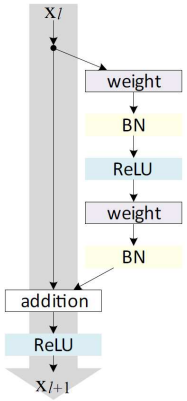

Figure 3: A residual unit. Here $\mathbf{x}_l/\mathbf{x}_{l+1}$ is the input/output feature to the l-th Residual Unit. Weight, BN, ReLU are linear convolution, batch normalization [10], and Rectified Linear Unit [18] layers.

The training started from a 50-layer ResNet, pre-trained on ImageNet 1K benchmark. To handle multiple-label classification, we use sigmoid output layer without softmax normalization.

Testing. To make the testing efficient, we apply all convolution layers on the input image once to get a feature map (typically non-square) and perform average pooling and sigmoid output layers. Not only our network provides more accurate predictions than VGG [22], which is used in many caption systems [7, 25, 13], it is also order of magnitude faster. The typical runtime of our ResNet is 200ms on a desktop CPU (single core only).

2.2. Language and semantic ranking model

Unlike many recent works [24, 25, 13] that use LSTM/GRU (so called gated recurrent neural network or GRNN) for caption generation, we follow [7] to use a maximum entropy language model (MELM) together with a deep multimodal similarity model (DMSM) in our caption pipeline. While MELM does not perform as well as GRNN in terms of perplexity, this disadvantage is remedied by DMSM. Devlin et al. [5] shows that while MELM+DMSM gives the same BLEU score as GRNN, it performs significantly better than GRNN in terms of human judgment. The results from the MS COCO 2015 captioning challenge[2] also show that the MELM+DMSM based entry [7] gives top performance in the official human judgment, tying with another entry using LSTM.

In the MELM+DMSM based framework, the MELM is used together with beam search as a candidate caption generator. Similar to the text-only deep structured semantic model (DSSM) [9, 21], The DMSM is illustrated in Figure 4, which consists of a pair of neural networks, one for mapping each input modality to a common semantic space. These two neural networks are trained jointly [7]. In training, the data consists of a set of image/caption pairs. The loss function minimized during training represents the negative log posterior probability of the caption given the corresponding image. The image model reuses the last pooling layer extracted in the word detection model, as described in section 2.1, as feature vector and stacks one more fully-connected layer with Tanh non-linearity on top of this representation to obtain a final representation of the same size as the last layer of the text model. We learn the parameters in this additional layer during DMSM training. The text model is based on a one-dimensional convolutional neural network similar to [21]. The DMSM similarity score is used as the main signal for ranking the captions, together with other signals including language model score, caption length, number of detected words covered in the caption, etc.

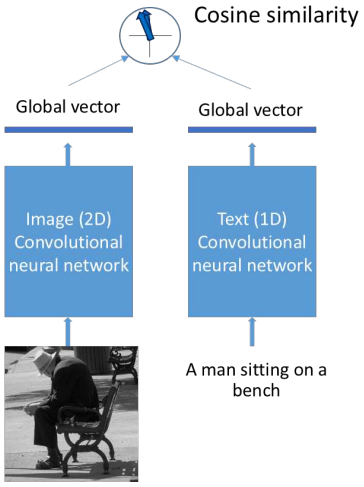

Figure 4: Illustration of deep multimodal similarity model

In our system, the dimension is set to be 1000 for the global vision vector and the global text vector, respectively. The MELM and the DMSM are both trained on the MS COCO dataset [16]. Similar to [9], character-level word hashing is used to reduce the dimension of the vocabulary.

2.3. Celebrity and landmark recognition

The breakthrough in deep learning makes it possible to recognize visual entities such as celebrities and landmarks and link the recognition result to a knowledge base such as Freebase [2]. We believe providing entity-level recognition results in image captions will bring valuable information to end users.

[2] http://mscoco.org/dataset/#captions-leaderboard

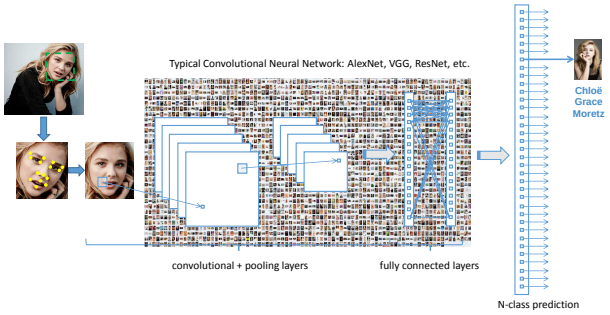

Figure 5: Illustration of deep neural network-based large-scale celebrity recognition. For each detected face in an input image, we crop and align the face region to a canonical view based on the facial landmarks, and then input this aligned face region to the deep neural work for celebrity prediction.

The key challenge to develop a good entity recognition model with wide coverage is collecting high quality training data. To address this problem, we followed and generalized the idea presented in [27] which leverages duplicate image detection and name list matching to collect celebrity images. In particular, we ground the entity recognition problem on a knowledge base, which brings in several advantages. First, each entity in a knowledge base is unique and clearly defined without ambiguity, making it possible to develop a large scale entity recognition system. Second, each entity normally has multiple properties (e.g. gender, occupation for people, and location, longitude/latitude for landmark), providing rich and valuable information for data collecting, cleaning, multi-task learning, and image description.

We started with a text-based approach similar to [27] but using entities that are catalogued in the knowledge base rather than celebrity names for high precision image and entity matching. To further enlarge the coverage, we also scrape commercial image search engines [8] for more entities and check the consistency of faces in the search result to remove outliers or discard those entities with too many outlier faces. After these two stages, we ended up with a large-scale face image dataset for a large set of celebrities.

To recognize a large set of celebrities, we resorted to deep convolutional neural network (CNN) to learn an extreme classification model, as shown in Figure 5. Training a network for a large set of classes is not a trivial task. It is hard to see the model converge even after a long run due to the large number of categories. To address this problem, we started from training a small model using AlexNet [14] for 500 celebrities, each of which has a sufficient number ($\geq$ 500) of face images. Then we used this pre-trained model to initialize the full model of a large set of celebrities. The whole training process follows the standard setting as described in [14]. After the training is finished, we use the final model to predict celebrities in images by setting a high threshold for the final softmax layer output to ensure a high precision celebrity recognition rate.

We applied a similar process for landmark recognition. One key difference is that it is not straightforward to identify a list of landmarks that are visually recognizable although it is easy to get a list of landmarks or attractions from a knowledge base. This implies that data collection and visual model learning are two closely coupled problems. To address this challenge, we took an iterative approach. That is, we first collected a training set for about 10K landmarks selected from a knowledge base to train a CNN model for 10K landmarks. Then we leveraged a validation dataset to evaluate whether an landmark is visually recognizable, and remove from the training set those landmarks which have very low prediction accuracy. After several iterations of data cleaning and visual model learning, we ended up with a model for about 5K landmarks.

2.4. Confidence estimation

We developed a logistic regression model to estimate a confidence score for the caption output. The input features include the DMSM's vision and caption vectors, each of size 1000, coupled with the language model score, the length of the caption, the length-normalized language model score, the logarithm of the number of tags covered in the caption, and the DMSM score.

The confidence model is trained on 2.5K image-caption pairs, with human labels on the quality (*excellent, good, bad, embarrassing*). The images used in the training data is a mix of 750 COCO, 750 MIT, and 950 Instagram images in a held-out set.

3. Evaluation

We conducted a series of human evaluation experiments through CrowdFlower, a crowd sourcing platform with good quality control[3]. The human evaluation experiments are set up such as for each pair of image and generated caption, the caption is rated on a 4-point scale: *Excellent*, *Good*, *Bad*, or *Embarrassing* by three different judges. In the evaluation, we specify for the Judges that *Excellent* means that the caption contains all of the important details presented in the picture; *Good* means that the caption contains some instead of all the important details presented in the picture and no errors; *Bad* means the caption may be misleading (e.g., contains errors, or miss the gist of the image); and *Embarrassing* means that the caption is totally wrong, or may upset the owner or subject of the image.

[3]http://www.crowdflower.com/

In order to evaluate the captioning performance for images in the wild, we created a dataset from Instagram. Specifically, we collected 100 popular Instagram accounts on the web, and for each account we constructed a query with the account name plus "instagram", e.g. "iamdiddy instagram", to scrape the top 100 images from Bing image search. And finally we obtained a dataset of about 10K images from Instagram, with a wide range of coverage on personal photos. About 12.5% of images in this Instagram set contain entities that are recognizable by our entity recognition model (mostly are celebrities). Meanwhile, we also reported results on 1000 random samples of the COCO validation set and 1000 random samples of the MIT test set [4]. Since the MELM and the DMSM are both trained on the COCO training set, the results on the COCO test set and the MIT test set represent the performance on in-domain images and out-of-domain images, respectively.

We communicated with the authors of Fang et al. (2015) [7], one of the two winners of the MS COCO 2015 Captioning Challenge, to obtain the caption output of our test images from their system. For our system, we evaluated three different settings: *Basic* with no confidence thresholding and no entity recognition, *Basic+Confi.* with confidence thresholding but no entity recognition, and *Full* with both confidence thresholding and entity recognition on. For *Basic+Confi.* and *Full*, we use templates such as *"this image is about ${top visual concept}"*, or *"a picture of ${entity}"* if entity recognizer fires, instead of the caption generated by the language model, whenever the confidence score is below 0.25. The results are presented in Tables 1, 2, and 3. Since the COCO and MIT images were collected in such a way that does not surface entites, we do not report *Full* in Tables 1 and 2.

As shown in the results, we have significantly improved the performance over a previous state-of-the-art system in terms of human evaluation. Specifically, the in-domain evaluation results as reported in Table 1 show that, compared to the baseline by Fang et al., our *Basic* system reduces the *Bad* and *Embarrassing* rates combined by 6.0%. Moreover, our system significantly improves the portion of captions that are rated as *Excellent* by more than 10%, mainly thanks to the deep residual network based vision model, plus refinement of the parameters of the engine and other components. Integrating confidence classifier to the system helps reduce the *Bad* and *Embarrassing* rates further.

The results on the out-of-domain MIT test set are presented in Table 2. We observed similar degree of improvements by using the new vision model. More interestingly, the confidence classifier helps significantly on this dataset. For instance, the rate of *Satisfaction*, a combination of *Excellent* and *Good*, is further improved by more than 10%.

[4] we randomly sampled a subset of the test images due to constraints on the need for human raters

System	Excel	Good	Bad	Emb
Fang et al. 2015	40.6%	26.8%	28.8%	3.8%
Ours (Basic)	51.4%	22.0%	23.6%	3.0%
Ours (Basic+Confi.)	51.8%	23.4%	22.5%	2.3%

Table 1: Human evaluation on 1K random samples of the COCO val-test set

System	Excel	Good	Bad	Emb
Fang et al. 2015	17.8%	18.5%	55.8%	7.9%
Ours (Basic)	23.9%	21.0%	49.0%	6.1%
Ours (Basic+Confi.)	28.2%	27.5%	39.3%	5.0%

Table 2: Human evaluation on 1K random samples of the MIT test set

System	Excel	Good	Bad	Emb
Fang et al. 2015	12.0%	13.4%	63.0%	11.6%
Ours (Basic)	15.1%	16.4%	60.0%	8.4%
Ours (Basic+Confi.)	23.3%	24.6%	47.0%	5.1%
Ours (Full)	25.4%	24.1%	45.3%	5.2%

Table 3: Human evaluation on Instagram test set, which contains 1380 random images from the 10K Instagram images that we scraped.

	Excel	Good	Bad	Emb
mean	0.59	0.51	0.26	0.20
stdev	0.21	0.23	0.21	0.19

Table 4: mean and standard deviation of confidence scores in each category, measured on the Instagram test set under the *Basic* setting.

Instagram data set contains many images that are filtered images or handcrafted abstract pictures, which are difficult for the current caption system to process (see examples in Figure 6). In the Instagram domain, the results in Table 3 shows that both the baseline and our *Basic* system perform quite poorly, scoring a *Satisfaction* rate of 25.4% and 31.5%, respectively. However, by integrating confidence classifier in the system, we improve the *Satisfaction* rate to 47.9%. The *Satisfaction* rate is further improved to 49.5% after integrating the entity recognition model, representing a 94.9% relative improvement over the baseline. In Figure 6, we show a bunch of images randomly sampled from the Instagram test set. For each image, we also show the captions generated by the baseline system (above, in green) and our *Full* system (below, in blue), respectively.

We also want to point out that, integrating the entity in the caption greatly improves the user experience, which might not be fully reflected in the 4-point rating. For ex-

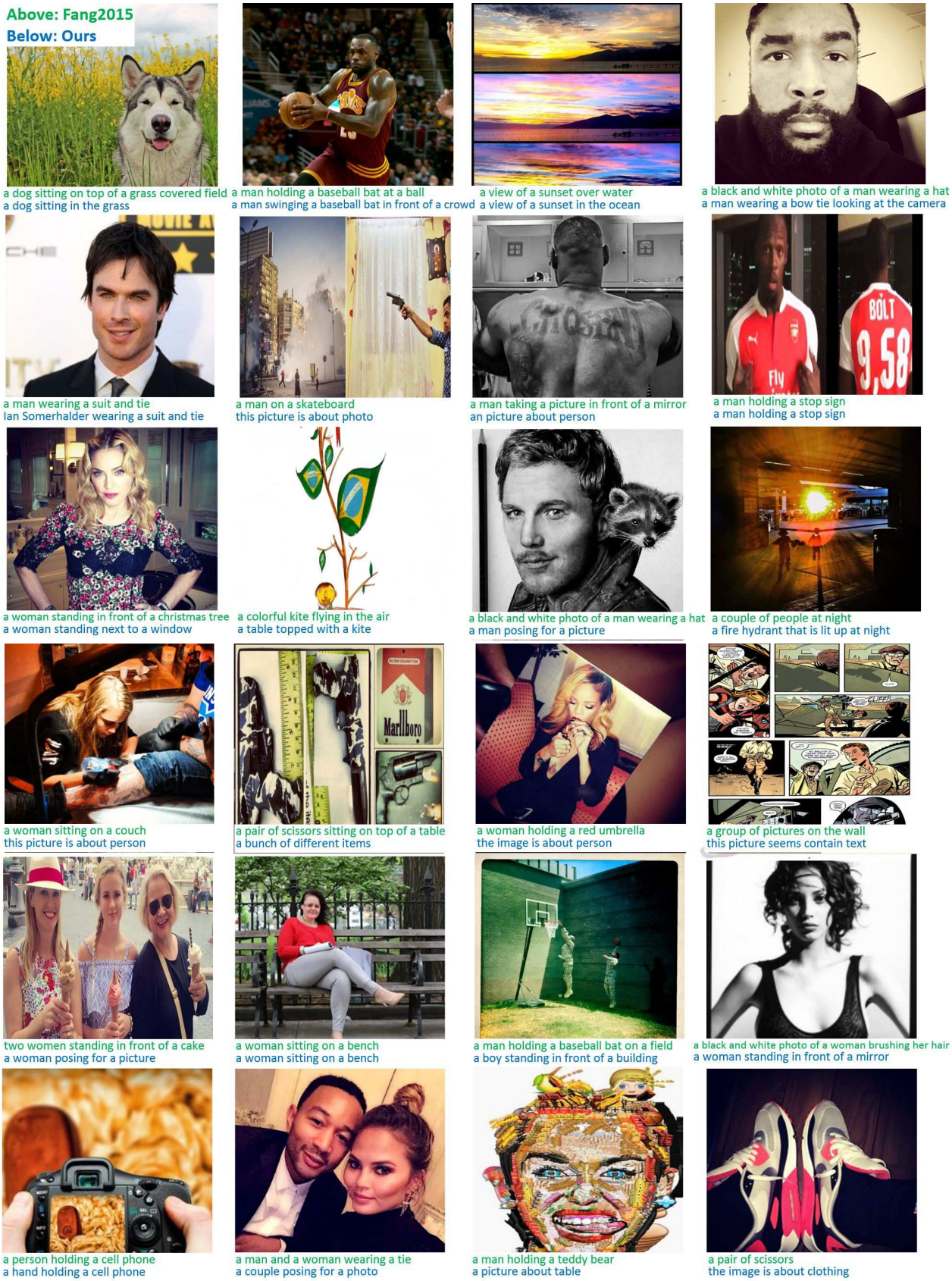

Figure 6: Qualitative results of images randomly sampled from the Instagram test set, with Fang2015 caption in green (above) and our system's caption in blue (below) for each image.

ample, for the first image in the second row of Figure 6, the baseline gives a caption "*a man wearing a suit and tie*", while our system produces "*Ian Somerhalder wearing a suit and tie*" thanks to the entity recognition model. Although both caption outputs are rated as *Excellent*, the latter provides much richer information than the baseline.

We further investigated the distribution of confidence scores in each of the *Excellent*, *Good*, *Bad*, and *Embarrassing* category on the Instagram test set using the *Basic* setting. The means and the standard deviations are reported in Table 4. We observed that in general the confidence scores align with the human judgements well. Therefore, based on the confidence score, more sophisticated solutions could be developed to handle difficult images and achieve a better user experience.

4. Conclusion

This paper presents a new state-of-the-art image caption system with respect to human evaluation. To encourage reproducibility and facilitate further research, we have deployed our system and made it publicly accessible [5], as part of Microsoft Cognitive Services.

5. Acknowledgments

The authors are grateful to Li Deng, Jacob Devlin, Delong Fu, Ryan Galgon, Jianfeng Gao, Yandong Guo, Ted Hart, Yuxiao Hu, Ece Kamar, Anirudh Koul, Allison Light, Margaret Mitchell, Yelong Shen, Lucy Vanderwende, and Geoffrey Zweig for valuable discussions.

References

[1] A. Agarwal and A. Lavie. Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 115–118. Association for Computational Linguistics, 2008.

[2] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM, 2008.

[3] V. Bychkovsky, S. Paris, E. Chan, and F. Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[4] C. Callison-Burch and M. Osborne. Re-evaluating the role of bleu in machine translation research. In *In EACL*. Citeseer, 2006.

[5] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell. Language models for image captioning: The quirks and what works. *arXiv preprint arXiv:1505.01809*, 2015.

[5] The landmark recognition feature will be available soon.

[6] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634, 2015.

[7] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, L. Zitnick, and G. Zweig. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1473–1482, 2015.

[8] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: Challenge of recognizing one million celebrities in the real world. In *IST International Symposium on Electronic Imaging*, 2016.

[9] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning deep structured semantic models for web search using clickthrough data. In *ACM International Conference on Information and Knowledge Management*, 2013.

[10] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[11] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. *arXiv preprint arXiv:1511.07571*, 2015.

[12] S. R. J. S. Kaiming He, Xiangyu Zhang. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[13] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[15] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*. Citeseer, 2011.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. Springer, 2014.

[17] T.-Y. 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*. 2014.

[18] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, pages 807–814, 2010.

[19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein,

A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[21] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil. A latent semantic model with convolutional-pooling structure for information retrieval. In *ACM International Conference on Information and Knowledge Management*, 2014.

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[23] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015.

[24] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.

[25] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015.

[26] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. *arXiv preprint arXiv:1603.03925*, 2016.

[27] X. Zhang, L. Zhang, X.-J. Wang, and H.-Y. Shum. Finding celebrities in billions of web images. *Multimedia, IEEE Transactions on*, 14(4):995–1007, 2012.