Multimodal Image Captioning for Marketing Analysis

Philipp Harzig, Stephan Brehm, Rainer Lienhart
University of Augsburg
Multimedia Computing and Computer Vision Lab
86159 Augsburg, Germany
{*philipp.harzig, stephan.brehm,*
rainer.lienhart}*@informatik.uni-augsburg.de*

Carolin Kaiser, René Schallner
GfK Verein
90419 Nuremberg, Germany
{*carolin.kaiser, rene.schallner* }
@gfk-verein.org

Abstract—**Automatically captioning images with natural language sentences is an important research topic. State of the art models are able to produce human-like sentences. These models typically describe the depicted scene as a whole and do not target specific objects of interest or emotional relationships between these objects in the image. However, marketing companies require to describe these important attributes of a given scene. In our case, objects of interest are consumer goods, which are usually identifiable by a product logo and are associated with certain brands. From a marketing point of view, it is desirable to also evaluate the emotional context of a trademarked product, i.e., whether it appears in a positive or a negative connotation. We address the problem of finding brands in images and deriving corresponding captions by introducing a modified image captioning network. We also add a third output modality, which simultaneously produces real-valued image ratings. Our network is trained using a classification-aware loss function in order to stimulate the generation of sentences with an emphasis on words identifying the brand of a product. We evaluate our model on a dataset of images depicting interactions between humans and branded products. The introduced network improves mean class accuracy by 24.5 percent. Thanks to adding the third output modality, it also considerably improves the quality of generated captions for images depicting branded products.**

Keywords-**image captioning; multimodal learning; marketing analysis; lstm; classification-aware loss**

I. INTRODUCTION

In recent years marketing companies have gained lots of experience with text analysis methods by collecting large text corpora from the Internet to identify new trends. One important task is to analyze the interaction with branded products in an emotional context, i.e., marketing companies analyze how people react to and interact with those products in order to track popularity and perception of brands over time. Not only text fragments play an important role in this analysis, but also social media pictures. These images could allow marketing companies to better understand how people interact with certain brands and perceive them emotionally. Consequently, we focus on transferring contents depicted by images into human readable text.

Our contributions are as follows: First, we provide a metric for measuring if an image is correctly classified with respect to objects of interest in the generated caption. Second, we implement a loss function which directly penalizes if the word for the company or brand does not appear in a generated caption. Third, we present a multimodal model which simultaneously generates captions and predicts three different image rating attributes.

II. RELATED WORK

Different approaches have appeared over time which try to convert the content of images into text fragments. Li et al. [1] use image features to compose sentences from scratch by using a text corpus comprised of n-grams gathered from the web. State of the art approaches use modern image classification DCNNs like [2] as feature extractors. One such architecture is the Show and Tell model by Vinyals et al. [3], [4], which makes use of recurrent neural networks (RNNs) for sequence modeling in the form of Long Short-Term Memory (LSTM) networks [5] to convert image features into human-like sentences. Karpathy et al. [6] also utilize neural networks for sequence modeling in form of a Bidirectional RNN to generate captions for images. Different approaches and their performances are reported on the website for the COCO [7] captioning challenge. The encoder/decoder structure used by current captioning approaches is heavily influenced by machine translation networks that transform a sentence from one language into another. E.g., Google [8] uses LSTM-driven neural networks for their translation system. Facebook [9] goes in another direction and eliminates the LSTM layers by replacing them through ordinary convolutional layers resulting in better performance both in speed and accuracy.

III. MULTIMODAL CAPTIONING AND RATING MODEL

Our model aims to produce captions relevant for marketing and to predict image ratings of three different categories from a single image. We therefore enhanced a popular image captioning model with two extensions to allow the combination of the three modalities of image, captions and image ratings in one single neural network architecture.

A. Show and Tell model

We use the Show and Tell model [3], [4] as a basis for our work. This model implements an encoder/decoder network architecture which is inspired by machine translation neural networks and is trained to maximize the probability $p(S|I)$ of generating the target sequence of words $S = (S_1, S_2, \dots)$ for an input image I in an end-to-end fashion. S consists of a variable number of words and the model is optimized for each (image, caption) pair. This model uses an LSTM network as the decoder RNN network and the loss $(L(I,S))$ that defines the error signal is given by the sum over the log likelihoods of the correct word at each position. During training, this loss is minimized w.r.t. (1) the parameters in the LSTM network, (2) the top layer of the CNN, which reduces the feature dimensions to a fixed length vector, and (3) the word embedding matrix $(\mathbf{W}_e)$, which reduces the feature dimensionality of one-hot vectors representing words in the vocabulary.

B. Dataset

We use a real-world dataset, which was created by a German market research nonprofit association (see Acknowledgment). The dataset consists of images that depict scenarios of persons interacting with branded products, e.g., a man holds a can with a Coca-Cola logo on it. Altogether, this dataset covers 26 different logo classes and consists of 10,529 images. Because of the small number of images in our dataset, we split it into a training and test dataset with a ratio of 9 to 1. We also use 8-fold cross-validation to select the best performing model due to the small number of training images. Two problems arise with our dataset. First, images are distributed unequally among the different classes. Second, we have extremely few training images compared to other datasets like MSCOCO [7]. Our dataset covers three different modalities. There are 10,529 images contained in our dataset with a total number of 52,645 captions associated with the images. Five annotators created a caption for each image. For a subset of all images (2,718) five annotators created image ratings with five possible values $(0-4)$. There are three kinds of image ratings, the first (r_1) describing whether the person interacts with the branded product in a positive (0) or negative (4) way, the second (r_2) describing if the person in the image is involved (0) with the branded product or uninvolved (4) and the third (r_3) describing if there is an emotional (0) or a functional (4) interaction with the branded product.

C. Classification aware loss for LSTM networks

As we want our model to describe the brand of an object correctly, we need to tell our model, whether or not a generated sentence contains the correct brand. Therefore, we introduce a special loss that punishes if the correct brand name is not contained within a generated sentence. We visualize the classification-aware loss in figure 1 with the

Show and Tell model on the bottom. We define a classword to be a word in the vocabulary V that clearly identifies a logo class (e.g., the word 'cocacola' is a classword.). We construct a one-hot mask vector $\mathbf{k}$ of size $|V|$, which has a one at every index of a classword.

$\mathbf{k}_t = \mathbf{n}_t \odot \mathbf{k}$ then provides the scores per word in the vocabulary by only focusing on the classwords, where $\mathbf{n}_t$ is the output of the LSTM at time step t. Since a classword can occur at any time step t within the sentence, we average over all words in the sentence ($\bar{\mathbf{k}}$). Note, that $\bar{\mathbf{k}}$ is a vector of size $|V|$ and $pred = \text{Softmax}(\bar{\mathbf{k}})$ is the probability distribution over classwords predicted by the generated sentence. Then, we define a cross-entropy loss that produces an error signal at the word for the correct class C with its corresponding one-hot vector $\mathbf{c}$ as $L_{\text{cls}}(I,C) = -[\log pred] \cdot \mathbf{c}$. The total loss of our modified model is now given by $L_{\text{total}} = L(I,S) + L_{\text{cls}}(I,C)$.

D. Regression of image ratings

Our third modality are three integer-valued image ratings in the range of $[0,4]$. We implement this prediction task with a linear regression model. To do so, we append a fully-connected layer with one output neuron ($\rho_r = \mathbf{W}_r \cdot \mathbf{\Theta}$) at the last layer ($\mathbf{\Theta}$) of the *Inception-v3* DCNN for each of the image ratings, where $\mathbf{W}_r$ is a 1×2048 weight matrix for image rating r. For the loss function we use the mean squared error function

$$L_r(I) = (\rho_r - g_r)^2 \qquad (1)$$

with g_r being the ground-truth for image rating r. The total loss of our model now changes to $L_{\text{total}} = L(I,S) + L_{\text{cls}}(I,C) + L_{r_1}(I) + L_{r_2}(I) + L_{r_3}(I)$. $L_{r_1}(I), L_{r_2}(I)$ and $L_{r_3}(I)$ are the image rating losses for ratings r_1, r_2 and r_3, respectively.

IV. EXPERIMENTS

In the following, we evaluate our model and compare it with a baseline model, which is trained with no modifications made to a vanilla Show and Tell model.

A. Sentence classification accuracy

The images of our dataset always show an interaction between a person and an object belonging to a certain company logo (e.g., a person is holding a Coca-Cola bottle). The sentence classification accuracy (SCA) metric measures the accuracy of correctly classified company logos within the generated captions.

For each class, we have a classword in the vocabulary that identifies the class (C). We generate three captions for each image using beam search with a beam size of $b = 3$. This heuristic considers the b best sentences up to the time step t as candidate sentences for the generation of sentences of length $t+1$. If the classification of an image is *cocacola*, we consider the set of three generated sentences
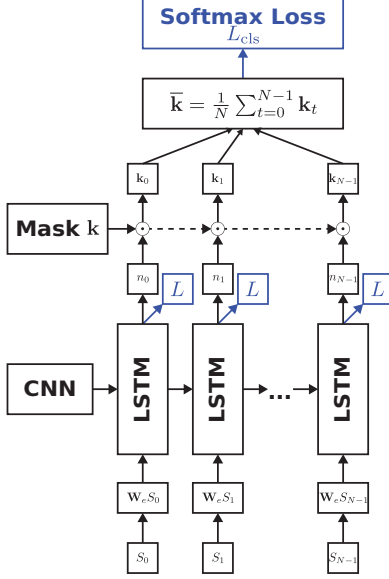
Figure 1. The classification aware loss function on top of the LSTM network.

(Φ) a match if the classword is contained in any of them, i.e., $\text{match}(I, \Phi, C) = 1$ if the classword for class C is contained in at least one sentence in Φ and equals to 0 otherwise. We then calculate the accuracy for every individual class (accuracy_C) by

$$\text{accuracy}_C = \frac{\sum_{I \in \mathbf{I}_C} \text{match}(I, \Phi, C)}{|\mathbf{I}_C|}. \qquad (2)$$

$\mathbf{I}$ and $\mathbf{I}_C$ are the sets of all test images and the test images belonging to class C, respectively. We use the mean accuracy (MA) metric because our dataset is distributed unequally among classes (see section III-B).

$$\text{MA} = \frac{\sum_{C \in \mathbf{C}} \text{accuracy}_C}{|\mathbf{C}|}, \qquad (3)$$

where $\mathbf{C}$ is the set of all class labels. We also measure the overall accuracy (OA) by

$$\text{OA} = \frac{\sum_{C \in \mathbf{C}} \sum_{I \in \mathbf{I}_C} \text{match}(I, \Phi, C)}{|\mathbf{I}|}. \qquad (4)$$

B. Training configuration

Inception-v3 network initialization: Instead of using an *Inception-v3* network pretrained on the 1000 classes of the ImageNet [10] classification task, we use an *Inception-v3* network fine-tuned to classify our 26 logo classes.

Show and Tell model: We train the Show and Tell model with the same parameters as Vinyals et al. [4]. For our extended model with the classification-aware loss, we just optimize all losses simultaneously according to L_{total}.

Image ratings: For the image ratings regression, we reduce the high learning rate of $\eta = 2.0$ by a factor of 2000 to $\eta = 0.001$. Because we have image ratings from five different annotators, we train our regression model with the mean of those five ratings as the ground-truth.

C. Results

In the following, we examine (a) sentence classification accuracy, (b) commonly used metrics for image captioning and (c) predicted image ratings for various models. The baseline model is the Show and Tell model trained on our dataset (*base*). Furthermore, we use the Show and Tell model with the *Inception-v3* model fine-tuned to our 26 logo classes (*base-ft*). Our third model implements the classification-aware loss discussed in section III-C and uses the pretrained *Inception-v3* model (*cls-aware*). We also conduct experiments, in which we merge our dataset with the MSCOCO dataset. To compensate for the fewer images in our dataset we use each image of our dataset eight times to balance both datasets. The resulting ratio is about $0.65 : 1$ for our dataset vs. the MSCOCO dataset. *fuse-1* is the model where both the LSTM part and the regression part are trained on the merged dataset for $1,700,000$ iterations. *fuse-2* continues training *fuse-1* with the parameters of the *Inception-v3* networks unfreezed and a learning rate of $\eta = 0.0005$ untill iteration $3,000,000$ We do not optimize by the classification-aware loss $L_{\text{cls}}(I, C)$ for both *fuse-1* and *fuse-2*. We initialize *fuse-3* with the parameters from *fuse-2*, i.e., *Inception-v3* parameters, LSTM parameters and image rating parameters. *fuse-3* is now trained on our dataset (see section III-B) with the classification-aware loss function enabled (in contrast to *fuse-1* and *fuse-2*).

Sentence classification accuracy: In table I we report SCA scores for our three models. We see that our problem mainly profits from fine-tuning the base network to the specific classification labels of our dataset. At a first sight, the classification-aware loss does only marginally improve the SCA scores (*cls-aware*). When using the fused datasets *fuse-1* and *fuse-2*, we notice a small degradation of the SCA scores, which is due to the fact, that the dataset now not only consists out of branded product images, but includes MSCOCO image-caption pairs. *fuse-3* uses all three modalities and allows changes to the *Inception-v3* base network, which as a consequence now allows to improve the SCA scores. The classification-aware loss is enabled in *fuse-3* and improves the mean accuracy in contrast to *cls-aware* by 1.3%. The multimodal architecture improves the mean accuracy by 24.5% in total in comparison with *base*.

Bleu-4, Meteor and Cider scores: In table I we also list how our models perform according to the well-known BLEU [11], Meteor [12] and CIDEr [13] scores. Note, that the scores for *gt* were computed by comparing each annotator against the other four and then averaging over all scores per annotator. For *base*, *base-ft* and *cls-aware* we compared the generated sentence against the five reference sentences. When using the classification-aware loss, all scores decrease compared to the baseline. In contrast to the SCA scores, the fused models profit heavily from the extended dataset, which helps the model to generate more diverse sentences.

model	OA	MA	Cider	Bleu-4	Meteor
base	0.757	0.587	1.771	0.548	0.341
base-ft	0.909	0.813	1.925	0.563	0.351
cls-aware	**0.910**	0.819	1.464	0.444	0.286
fuse-1	0.889	0.796	1.858	0.542	0.344
fuse-2	0.880	0.794	2.012	0.585	0.364
fuse-3	0.908	**0.832**	**2.107**	**0.612**	**0.371**
gt	-	-	1.943	0.520	0.382

TABLE II
DEVIATIONS FROM THE IMAGE RATINGS ON THE TEST SET FOR DIFFERENT MODELS.

model	d_{r_1}	d_{r_2}	d_{r_3}
base	0.5703	0.8141	0.8504
base-ft	0.5663	0.8656	0.8078
cls-aware	0.5644	0.8613	0.8054
fuse-1	0.6071	0.9379	0.8576
fuse-2	**0.5218**	**0.7827**	0.6974
fuse-3	0.5236	0.7876	**0.6968**
gt	0.2541	0.1894	0.3138

When finetuning the *Inception-v3* base network (*fuse-2*), the parallel training of the three modalities improve the Bleu-4, Meteor and Cider scores in comparison to *fuse-1*.

Image ratings: Since we modeled the image ratings as a linear regression model and possible values are in the range $[0, 4]$, we use the mean deviation from the ground-truth as a performance measure. We calculate the deviation for a single example ϵ with $\text{deviation}_r(\epsilon) = \sqrt{(\rho_r(\epsilon) - g_r(\epsilon))^2}$. Note, that this is the square root of the mean squared error as defined in eq. 1. We calculate the mean deviation (d_r) for an image rating r by averaging over all examples in the test set. In table II we present the calculated deviations on our test set. All image rating deviations increase, when using the fused dataset and start to decrease below the values from *base, base-ft* and cls-aware when evaluating the *fuse-2* and *fuse-3* models. Note, that training the three modalities in parallel produces best results for all three experiments. The scores for *gt* were computed by comparing each annotator against the other four and then averaging over all deviations per annotator.

V. CONCLUSION

We proposed an architecture specifically aimed to analyze images relevant for market research. Our model combines the three modalities of image, caption and image ratings into one unified model. We extended an image captioning model to prioritize on describing the brands of products depicted in a scene by using a special loss function, which penalizes if a brand name is not contained within the caption.

Furthermore, we enabled the model to simultaneously infer ratings from the input images. We evaluated six different models and merged our dataset with the famous MSCOCO dataset. As a result, we found that extending our specific dataset improves overall sentence quality but decreases the detection rate of brand names. Our approach of combining all three modalities in one model improves caption quality, brand name detection and image ratings and shows that multiple modalities help to get better performance at all tasks of such a model.

REFERENCES

[1] S. Li *et al.*, "Composing simple image descriptions using web-scale n-grams," in *SIGNLL CoNLL*. Association for Computational Linguistics, 2011, pp. 220–228.

[2] C. Szegedy *et al.*, "Rethinking the inception architecture for computer vision," in *IEEE CVPR*, 2016, pp. 2818–2826.

[3] O. Vinyals *et al.*, "Show and tell: A neural image caption generator," in *IEEE CVPR*, 2015, pp. 3156–3164.

[4] ——, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE TPAMI*, 2016.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *IEEE CVPR*, 2015, pp. 3128–3137.

[7] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

[8] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[9] J. Gehring *et al.*, "Convolutional sequence to sequence learning," *arXiv preprint arXiv:1705.03122*, 2017.

[10] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[11] K. Papineni *et al.*, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[12] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *ACL*, vol. 29, 2005, pp. 65–72.

[13] R. Vedantam *et al.*, "Cider: Consensus-based image description evaluation," in *IEEE CVPR*, 2015, pp. 4566–4575.