Creative Image-Captions Recommendation and Sentiment Extraction for personalization in marketing

MS Artificial Intelligence (Lentzas) SP2020

Author: Isha Gupta
Isha.gupta@columbia.edu
Columbia University in the City of New York

Abstract: *This paper lies at the intersection of Natural Language Processing, Computer Vision, and Sentiment Analysis. It proposes a novel solution to one of the most challenging problems of marketing – understanding the dynamic nature of consumer psychology -through an added feature on social media. Based on the picture posted or shared, the user is given a set of creative caption choices as recommendations, and based on the selection made – user sentiment can be extracted to market a product in a more user-relevant, interesting, and personalized manner. Due to information overload and intensified digital activity among consumers, a deeper level of personalization is needed that can target customers, not as part of presumed segments rather on individual levels. Providing engaging captions recommendation and then deriving the consumer sentiment behind a picture has strong potential in optimizing sequential marketing touchpoints leading to stronger consumer attraction and conversion.*

I. Introduction

'As the progression of pandemic moves through the globe, consumer sentiment is starting to waver.' says a study conducted by the consumer research team at Mckinsey and Company. The human element is the most critical aspect of successful inbound marketing. And inbound marketing is a cost-effective way of pulling in good-fit customers, generating leads, and pushing them through the sales funnel. Lead generation, which implies reaching new prospects and attracting new customers, is one of the most challenging yet most critical tasks for business growth. Moreover, due to the pandemic, consumer sentiment and behavior is shifting drastically. As per the survey, customers are reducing discretionary spending, developing intensified digital behaviors, and changing the level of optimism. Hence, understanding buyers' personas and perspectives is essential not just for stronger lead generation but also to stand out in the clutter and attract customers amidst the information overload. Personalization of content to reach the mind and heart of consumers with a unique and fresh approach using more interesting and relevant content can make a huge difference in the marketing of a product. This paper proposes a solution by creating an image caption generator on pre-trained InceptionV3 architecture and then creating creative captions of certain categories to recommend the user to choose from. Engaging a customer by providing a social media service can be a smart way to extract the emotion, sentiment, and type of a customer based on the caption selected. This gives the potential to go into a level deeper into personalization. This paper, hence, seeks to find an answer to the question: *'Can the intensified digital engagement of consumers, especially on social media, be leveraged to gain a more granular level of qualitative data on consumers' persona and behavior to enhance consumer attraction and retention through marketing?'*

II. Organization of paper

This paper has two parts – Image caption generation and then creative captions generation. Image caption generation is a famous artificial intelligence problem where a descriptive sentence is generated for an image. It is a combination of computer vision that understands the content of the image and Natural Language Processing which generates words in a sequential manner to describe the content. Furthermore, a Creative caption generator is created on top of it to generate captions including quotes based on various sentiments derived from the keywords from the description. Based on the type of caption selected, users can be segmented based on similarity in psychology to target in a more efficient manner.

III. Dataset

Flickr8k dataset [1] was used for training, validating, and testing the model. It is a standard benchmark for sentence-based image description. It has 8000 images, each with 5 captions, divided into the train, validation, and test sets. Next to it, rather than fitting the word vector while training the model, pre-trained word vectors were used which improved the performance significantly. GloVe (Global Vectors for Word Representation) 6B word vectors with 300-dimensional embeddings were used to create embeddings matrix which is trained on aggregated global word-word co-occurrence statistics from a corpus of Wikipedia and Gigaword 5.

IV. Methodology

1. Main Model

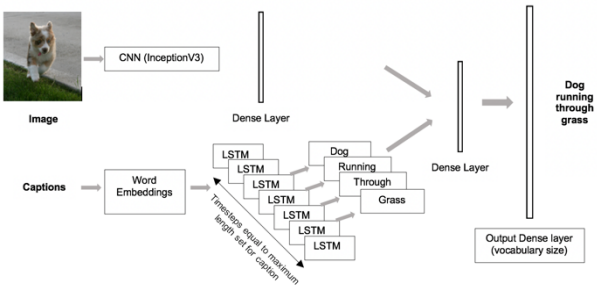

Figure 1. Model Architecture

1.1 Image caption generator

The model has three kinds of neural networks – Convolution Neural Network (CNN) for image feature extraction, Long Short Term Memory (LSTM) for word by word sequential generation of the caption of the image and standard encoder-decoder Recurrent Neural Network (RNN) architecture for combining the image feature generation model and the language generation model. The encoder encodes the content of the image into a fixed-length vector using an internal representation. The decoder reads the encoded image vector and generates a description in text format as the output.

A. Image Feature Extractor

CNN's work well with image classification because layers learn local patterns in the input space (e.g. edges and other shapes) – unlike a fully connected network that tries to learn global patterns. This has two implications: First, convnets learn patterns that are *translation invariant* which means that if a convnet learns a pattern in the top left corner of an image (e.g. an edge), it can recognize it elsewhere in the image. A fully connected neural net rather would have to learn the same pattern from scratch if it appeared in a new location. Thus, convnets are highly data efficient for data that possesses the translation invariance property, such as images.

Second, convnets can learn spatial hierarchies of patterns. First convolution can learn simple, local (in the input space) patterns (eg. Edges), the second convolution can learn more complex, larger patterns made from features of the first layer, and so on. Hence, convnets work well for data which is aptially hierarchical, such as images. Once the pre-trained model was loaded, the output of the model was removed and internal representation of the image is used as the encoding or internel representation of the input image.

B. Sequence Processor

LSTMs work well when learning and predicting sequential pattern of data is required. LSTMs are modified Recurrent Neural Networks (RNNs) which only transfer short term memory but are capable of dealing with long term dependencies as well. They generalize the idea of the leaky unit to allow for weights that change over time. LSTM setup allows the network to accumulate information over a long period of time but also to forget the old information once no longer relevant. And most importantly, model learns when to do this. An LSTM layer processes batches of sequences like other Keras layers by taking inputs of shape (batch size, timesteps, input features). Length of timesteps in caption generation would be the maximum length of caption set while model building and each timestep will represent adding predicted word to the new input sequence.

C. Decoder

Extracted image feature information can be combined with sequential generation of text through two types of architectures [2]:

1) Inject Model: It combines the image vector with every new word generated by model

2) Merge Model: It combines both the image vector with vector embedding of text which is decoded to generate next word in the sequence.

Merge model has been proven to outperform the inject model, and hence, merge model was used.

Therefore, once fixed-length vectors were generated after encoding image features and text word embeddings, both were merged together to be processed by dense layer to make final prediction in sequential manner.

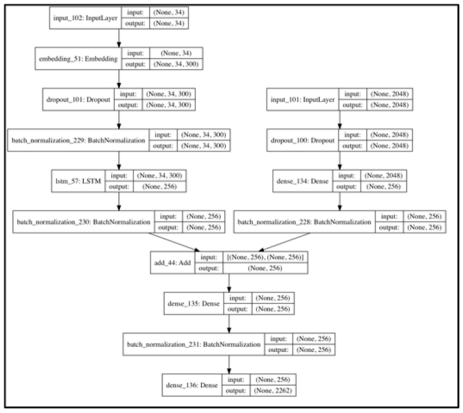

Figure 2. Image captioning Model

1.2 Creative Captions Generator

Caption generated by the previous model is more of a robotic description about the content of the image. The general structure of the predicted caption is of the format subject-verb-object: 'subject' in '<description of color or clothes> 'verb' 'object'. Further preprocessing this text by removal of special characters, single characters, additional spaces, and lemmatiation extracts major key words from the generated caption. To extract psychological insight of the consumer, quotes were scraped from this quotes website using selenium's webdriver which were then saved as a dictionary with keys as tags of quotes and values as the lists of quotes with quote and corresponding author. Using the embeddings dictionary saved from the GloVe word vector embeddings, vectors for each of the key word is extracted. Cosine similarity can also be used as a similarity metric although the concept of closeness in the space is more

relevant here rather than the same direction of vector. And hence, Euclidean distance is chosen as the similarity metric to find the best tage and hence the best quote for the given keyword in the originally generated sentence. Using Euclidean distance, closest tag is extracted for each of the key words of sentence. And quote for that closest tag is shown as an option. User is finally shown all the quotes corresponding to each keyword as well as the main descriptive sentence that was generated originally. Further, sentiment analysis is performed on chosen caption using NRC data [3], to derive consumer's psychology behind a post. NRC Emotion Lexicon has 2 sentiments (positive, negative) and 8 emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) for 14182 unigrams (words). NRC sentiment Analysis generates corresponding emotions and sentiments of user from the chosen caption.

This gives a good caption to the user for her/his image in the post and provides a strong consumer's behavioural insight to the marketing team.

1.3 Model Metrics

$$Log\ Loss\ =\ \sum -y.log\ log\ (p)\ -(1-y).log\ (1-p)$$

Figure 3. Cross Entropy Loss

Categorical cross entropy loss is used as one of the best metrics to evaluate a classifier. Hence, Categorical cross entropy loss was selected as the model evaluation metrics as output was a categorical variable with every word in the reduced vocabulary acting as a category. Also, BLEU scores were checked as well. BLEU (Bilingual Evaluation Understudy) is an algorithm used to evaluate quality of a sentence which has been translated by machine from one natural language to another. Quality implies the correspondence between a machine's output and that of a human's by combining both the precision and recall.

V. Model Optimization

1.1. Image Features Extraction by Transfer Learning

CNN models pretrained on Imagenet dataset were used to extract features of images in Flickr dataset. At first, VGG16 model was used which has a size of 528 MB with 138,357,544 parameters topological depth of 23 and Top-1 and Top-5 accuracies as 0.71 and 0.90 respectively. Final model performed better with transfer learning from InceptionV3 which is a deeper model with topological depth of 159 and size of 92 MB and 23,851,784 parameters.

1.2. Reduction and optimization of Vocabulary

Around 8000-words dictionary was extracted by fitting tokenizer on the descriptions. Initially complete vocabulary was used which was optimized further and reduced to improve the model performance. Words which were misspelled, occurred with extremeley low frequencies or had small lengths were removed to refine the vocabulary and reduce its size in order to improve model performance.

1.3. Using Pre-trained Word Vector on larger corpus

Initially word embeddings layer of the model was trained as part of the model fitting. To achieve better performance and optimize model training, GloVe word vectors were used to create embeddings matrix which are pre-trained on larger corpus of text from Wikipedia and Gigaword. This layer was then freezed for training while model fitting.

1.4 Data Generator Function

Initially sequences of input and ouput words were created for every description and fed into the model training. For memory optimization and creating one batch of inputs outputs at a time data generator function was created. 5 captions for each of the 6000 images, total of 30000 image-caption pairs were required. Further assuming each caption to be 7-words-long then it totaled to *30000* 7=210000* datapoints. Next to it, for every datapoint in *210000* datapoints, each had total length equal to the length of image vector i.e. *2048* (because of inceptionV3) + *34*300* (maximum length of caption * length of one word in embedding dimensions) making total length of one = *2048 + 10200 = 12248*. This leads to a total size of matrix as *210000*12248 = 2572080000* blocks. Hence, to optimally utilize the available memory and optimize model training, data generator function was created.

VI. Model Tuning

Layer (type)	Output Shape	Param #	Connected to
input_116 (InputLayer)	(None, 34)	0	
input_115 (InputLayer)	(None, 2048)	0	
embedding_58 (Embedding)	(None, 34, 300)	678600	input_116[0][0]
dropout_114 (Dropout)	(None, 2048)	0	input_115[0][0]
dropout_115 (Dropout)	(None, 34, 300)	0	embedding_58[0][0]
dense_155 (Dense)	(None, 256)	524544	dropout_114[0][0]
lstm_64 (LSTM)	(None, 256)	570368	dropout_115[0][0]
batch_normalization_254 (BatchN	(None, 256)	1024	dense_155[0][0]
batch_normalization_255 (BatchN	(None, 256)	1024	lstm_64[0][0]
add_51 (Add)	(None, 256)	0	batch_normalization_254[0][0] batch_normalization_255[0][0]
dense_156 (Dense)	(None, 256)	65792	add_51[0][0]
batch_normalization_256 (BatchN	(None, 256)	1024	dense_156[0][0]
dense_157 (Dense)	(None, 2262)	581334	batch_normalization_256[0][0]

Total params: 2,423,710
Trainable params: 2,422,174
Non-trainable params: 1,536

Figure 4. Model Summary

Model Tuning and architecture Advancement:

1) Keras Functional API: This offers a more flexible way to create models than the conventional Sequential API which assumes that the neural network has exactly one input and one output and that it consists of a linear stack of layers. Keras functional API allows to specify a model as a function of tensors and hence, is able to handle models with non-linear topology, models with shared layers, and models with mutiple inputs and outputs. In our case, model needs to have two inputs. It is essential for multimodal input of both the image and text where relation has to be predicted.

2) Batch Normalization: Normalization is a prominent term in Machine Learning which seeks to make datapoints in the dataset similar to each other, which helps the models to learn and generalize better. However, ML's simple normalization effect disappears in the neural network due to the successive passes through complex layers. It, in fact, vanishes after single layer transformation. Hence, using keras *BatchNormalization()* layer, data was continuously normalized as it passed through the network.
   It significantly helps with gradient propagation, thus allowing for the training of much deeper networks. It standardizes the inputs to every layer for each mini-batch. It stabalizes the learning process reducing the number of training epochs required for the entire training. This addition to the architecture improved the model performance to significant levels, in fact the most. Initially, only 2 layers of *BatchNormalization()* were added to the architecture for simplicity and checking the improvement. As significant reduction in loss was seen, adding the layer into architecture at every pass, generated the best results among all tuning techniques.
3) Regularization: It's one of the most effective and prominent generalization techniques in Machine learning. It regularizes the capacity of models by adding parameter norm penalty to the loss function or objective function to be minimized. Since, pretrained image classifier InceptionV3 was used, regularization could be added to only the text encoding. Although, model performance reduced on adding *kernel_regularization* of 0.001 as well as 0.0001 to the model. It was observed that addition of regularization parameters was leading to over-generalization of model that there was no reduction in the validation loss. And hence, image caption generator model performs better without the use of regularization terms.
4) Optimizer and learning rate: As the model training is time taking, during training of every architecture of model, the model was saved after every epoch if there was an improvement in the validation loss. To optimize the overall training, if an already trained architecture was further hypertuned then best model weights were loaded and initialized to train the model further and improve overall model performance on top of it. It saved a lot of time and optimized the model training significantly. Adam optimizer was used with varying learning rates. Model trained better with 0.0001 learning rate, which is 10 times slower than the default learning rat eof 0.001 of Adam optimizer.

VII. Results

Minimum validation loss achieved was 3.09885. When this model was trained on entire train as well as validation dataset, the loss on the test set was this. Further the BLEU [5] scores were also calculated for the prediction.

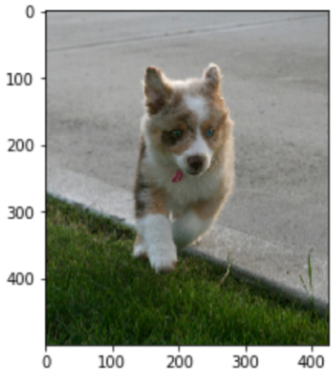

Figure 5. Input Image 1

Predicted:
dog running through the grass

Actual:
- *Beige white and black little dog is bounding towards the camera from the street onto the green grass*
- *Dog is running along the grass*
- *Multicolor puppy is running on the lawn beside road*
- *Puppy running along street*
- *Small dog is running on the grass beside the road*

Figure 6. Input Image 2

Predicted:
Young girl swings on swing

Actual:
- *blonde child swinging on swing*
- *smiling child wearing white tshirt with stripes is swinging*
- *young child in swing wearing skull and crossbones shirt*
- *the blonde haired child played on the swing*
- *the young girl in cartoon shirt is enjoying ride on swing*

Creative Captions generated:

```
"If you can make a woman laugh, you c
an make her do anything."          by
Marilyn Monroe

 "There is nothing to writing. All yo
u do is sit down at a typewriter and
bleed."                by Ernest Heming
way

 "The world as we have created it is
a process of our thinking. It cannot
```

be changed without changing our think ing." *by Albert Einstein*

VIII. Conclusion and Application in Marketing

In nutshell, this paper was successful in coming up with proposing a solution to personalize marketing on a deeper level. Varying global consumer sentiment and their intensified digital interaction can be leveraged to stand out in the clutter by providing a caption recommender along with consumer sentiment extractor to understand a user's perspective, sentiment and psychology behind a particular kind of image. Targeted marketing is picking up through excessively used social media platforms such as Instagram, Facebook, Twitter etc. For larger lead generation to bring more traffic to the landing page and push it through the sales funnel, consumers are shown advertisements as per the 'segment' they belong to – depending on various age, gender, locations, professions they belong to. The 'ideal' consumer segmentation is still a curiosity for marketing teams across the globe. The ideal consumer segmentation is the one where within each segment, consumers are fascinated by same kind of content, can be pulled to the landing page by showing same advertisement and lead generation could be improved significantly by segment-respective targeted marketing. Consumers in same segment thus need to have same psychology and personality at the time of advertisement which can be traced by offering creative image-captions recommendations to extract real time consumer sentiment for better real-time personalized marketing. Therefore, along with the conventional applications of image caption generator such as Assistant for visually impaired, this paper discovered and developed an application in the domain of marketing through social media monitoring. For example, during the pandemic of Coronavirus, what kind of pictures is our target segment is posting. Further, if offered caption recommendations – what aspects of an image do they 'most' care about. Which topic, sentiment or emotion are they most sensitive towards? What an individual craving or missing the most? By extracting sentiments on individual level, marketing can be made stronger through deeper personalization.

IX. Future Scope

1. A pre-trained Keras model InceptionV3 was used to extract image feature vectors although there is always a scope to improve upon these pre-trained models especially using newer techniques such as depthwise seperable layers, image auto augmentation and cutmix augmentation (which cuts and merges separate images and gives the label corresponding fraction of rating) etc. Also, different pre-trained models such as ResNet can be tried.
2. However while building this model various number of layers, batch sizes, learning rates, stacked layers, regularization, Normalization etc., the architecture of a neural network always has a deeper scope of improvement which might give better results.
3. The expert scores given to each of the descriptions in the Flicker Dataset can be used to provide weight to the descriptions while training. Giving higher weight to better and close to perfect descriptions is further expected to improve model performance.
3. For creative captions recommendations, context can be predicted by creating bidirectional LSTM and recommend hastags. It can help in deriving and constructing more creative captions to attract users to make the choice in order to capture their sentiments.
4. As creativity has no limits, more ways of caption creation can be built. For instance, extracting sentences from twitter and google based on the key words and hashtags generated. The more interesting caption choices are created, the better will be the consumer sentiment extraction.

X. Code

Code for the entire research project is presnet at *Github*.

XI. References

[1]. M. Hodosh, P. Young and J. Hockenmaier (2013) "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics", Journal of Artifical Intellegence Research, Volume 47, pages 853-899 http://www.jair.org/papers/paper3994.html

[2]. Tanti, M., Gatt, A., & Camilleri, K. P. (2017). What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator?. *arXiv preprint arXiv:1708.02043*.

[3]. Mohammad, Saif M., and Peter D. Turney. "Crowdsourcing a word–emotion association lexicon." *Computational Intelligence* 29.3 (2013) 436-465.

[4]. Brownlee, Jason. "How to Develop a Deep Learning Photo Caption Generator from Scratch." *Machine Learning Mastery*, 19 Apr. 2020, machinelearningmastery.com/develop-a-deep-learning-caption-generation-model-in-python/.

[5]. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, [Online] Available: https://papers.nips.cc/paper/4824-imagenetclassificationwith-deep-convolutionalneural-networks.pdf

[6]. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database

76]. BLEU: a Method for Automatic Evaluation of Machine Translation Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu IBM T. J. Watson Research Center Yorktown Heights, NY 10598, USA