#okboomer : A multi-modal approach towards generating hashtags on social media posts

Pradyumna Tambwekar
School of Interactive Computing
Georgia Institute of Technology
pradyumnat@gatech.edu

Sameer Dharur
College of Computing
Georgia Institute of Technology
sameerdharur@gatech.edu

Abstract

Filtering of data into appropriate categories is among the oldest problems in computer science. Since the advent of the internet and the subsequent surfeit of data collected by websites, this problem has become a ubiquitous and exponentially challenging one. Among the popular manifestations of this today is on social media platforms that host several thousands of gigabytes of data and turn to hashtags as a means of achieving this categorization. We propose a novel (to the best of our knowledge) multi-modal deep learning method for the generation of hashtags – conditioned on an input image and a corresponding caption to the image. Viewed another way, this is an extension of the auto-correct feature on most text applications today, with the addition being that the predicted values are dependent not just on textual inputs but also on images. As part of this project, we have also collected a new dataset comprising Instagram posts with their images, captions and hashtags. In our evaluations, our best results have a BLEU score of 0.69 on the validation split while trained over a subsample of 7160 images across 20 epochs which we believe could get better when trained without computational and time constraints.

1. Introduction

The advent of social media has spawned a continuous stream of large-scale data generation on a daily basis. Instagram alone has over 95 million images uploaded each day [1], in the form of posts and stories. It becomes essential, therefore, to group these massive amounts of data into intuitive and relevant categories that filter information and help drive user engagement. Owing to the mutual interest of consumers as well as advertizers on social media platforms, accurate categorization of data is an essential priority that benefits every participant on these media to maximize their reach and reduce time spent sifting through vast troves of data to find relevant content. If successful, our solution could make social media a much more efficient space for end-users, advertizers and platform-owners alike.

The current paradigm for hashtagging [2] of posts is entirely manual and unhelpful to a new user on the platform. The choice of hashtags also tends to be subjective

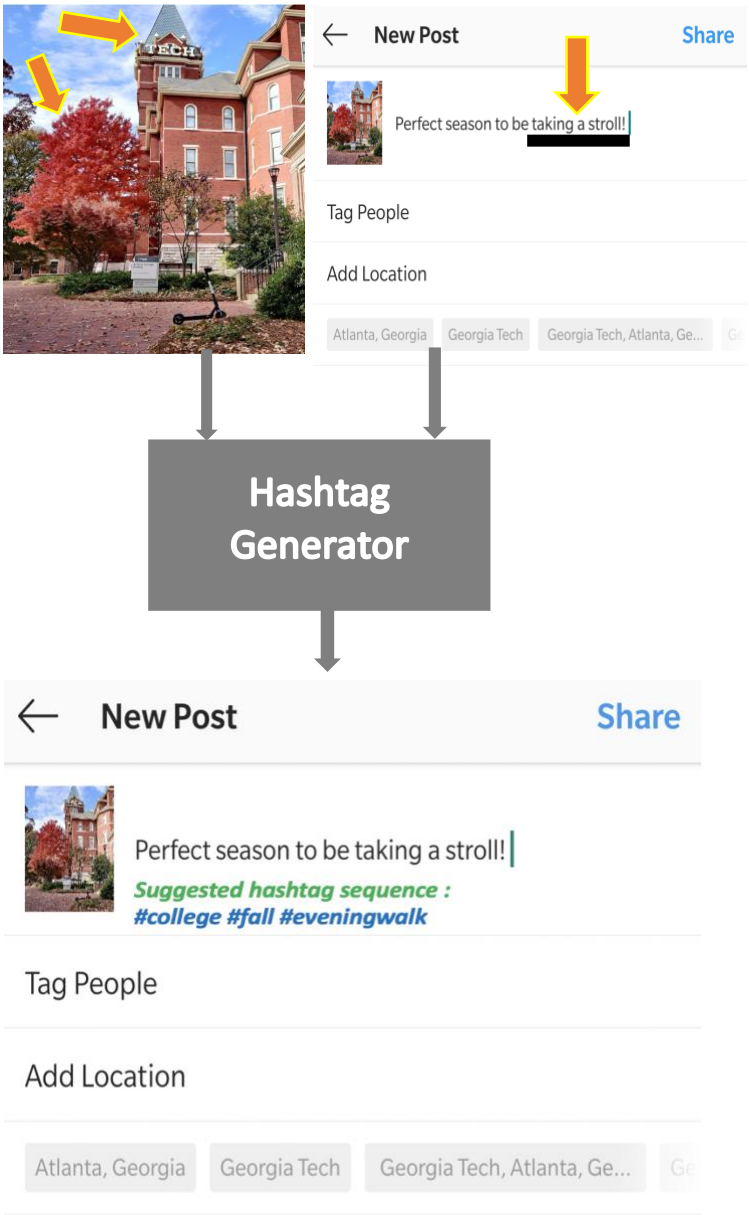

Fig 1. The workflow of our solution, which takes the image and the caption as input and generates a natural language sequence of relevant hashtags as the output.

among users who exhibit different levels of engagement with social media, typically diverging among people of different ages and cultures. While this is an unsurprising reflection of the diversity of users on the platform, it comes with the drawback of relevant data being grouped in a less robust manner, pointing to the need for recommending a unified way of grouping data that overcomes individual preferences. It is, therefore, useful to incorporate an automated framework to generate hashtags for a post, given its content in the form of an image and a caption (Fig 1). This could potentially be a recommendation feature while making a new post on the platform, similar to autocomplete suggestions [3] on most keypads today. Unlike in the case of autocomplete, which is a text-only problem, our solution would handle inputs that include an image and its corresponding caption, to give us the output of the desired hashtags. Therefore, we model this as a multi-modal problem spanning vision and language where the inputs – an image and a caption – are encoded together, and the output is a decoded natural language piece of text containing the hashtags relevant to the encoding.

2. Related Work

We view our task as being a combination of visual question answering (VQA) [4] and image captioning [5]. VQA is a task that requires models to associate image features extracted from a Convolutional Neural Network (CNN) [6] classifier with text features extracted from a Recurrent Neural Network (RNN) to eventually output the class of the highest probability over a finite set of outputs. Image captioning, instead, does not deal with multi-modal inputs – taking just the image features as the input and giving us a natural language caption (through an RNN) as the decoded output.

Our task combines the encoding of a VQA setup and the decoding of the image captioning setup, with the inputs being an image whose features are extracted from a CNN and a caption encoded through an RNN, to generate a sequence of hashtags through an RNN decoder. Although we saw potential in experimenting with the outputs as a finite set of words that could be modelled as a classification problem - in our empirical observations, it was necessary to design the pipeline this way since from an encoding perspective, the contents of the image as well as the caption are relevant to the output (hashtags), whereas the hashtags themselves tend to contain temporal dependencies and can often be long winding with unpredictable sizes.

The dominant learning paradigm in VQA of <question, image, answer> may not necessarily be universally applicable since the outputs are not always in a softmax-able range of values. It is to this end that our decoder was chosen to output a free form natural language chunk of text, similar to captioning, with each word being a relevant hashtag for the image and its corresponding caption.

The use of Instagram posts with their hashtags in deep learning has been attempted before, albeit for a different purpose. In [7], Mahajan et al explore the idea of augmenting the Imagenet dataset with data from Instagram, using the hashtags of images as their classification labels in a weakly supervised setting. Although our problem statement is entirely different, this was a useful trigger for us to explore the idea of working on social media data.

3. Setup and Dataset Analysis

For the purpose of this new task, we did not have a prior dataset to pick up and import off-the-shelf. All data for this project was collected manually from scratch via a Python tool called *instaloader*. We manually identified 130 most commonly used hashtags on Instagram and scraped publicly uploaded data on the social media platform through the tool, fetching 1000 images for each hashtag along with their corresponding captions. This was followed by the pre-processing and cleaning of the data to remove images containing no captions or hashtags, or those with captions and hashtags in languages other than English. The data also had to be organized in appropriate JSON files after the cleaning process, along with creating a 90:5:5 split of training, validation and testing data for the images and their corresponding captions and hashtags. The hashtags were also widely representative of various topics that commonly gain traction on social media platforms across sports, entertainment, business, politics and personal experiences.

While the entire dataset has a size of about 120 GB, we were unable to train on all of it owing to our constraints on the availability of compute power. In our experience, the process of downloading and uploading such enormous data to a cloud VM and running even a few epochs is prohibitively time and cost expensive, greatly limiting the scope for hyperparameter tuning. We've reported the results on this paper by training on a subset of $1/10^{th}$ the original images on 2 GPU cores of the NVIDIA P100.

4. Hashtag generation

Each post on most social media platforms, such as Instagram and Twitter, broadly consists of three components – an image, an optional caption and one or more hashtags associated with the content of the post which too are optional. For the purposes of this project, we've cleaned up our data to remove entries whose captions or hashtags contain emoji or non-English content, although this could be modelled too in future upon collecting sufficiently represented data. As mentioned earlier, our input is the image and a caption, while the output is the series of hashtags.

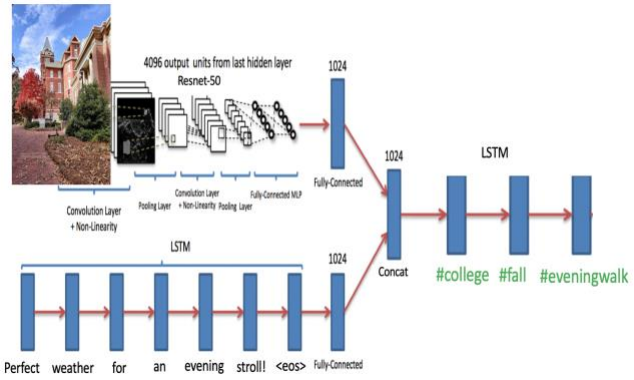

Fig 2. A representation of our multi-modal encoder-decoder architecture.

The input encoder consists of a Convolutional Neural Network (CNN) to process the image's features and a Recurrent Neural Network (RNN) to process the caption (Fig 2). Our baseline comes from the work by Kazemi *et al* [8] in a VQA setting, using the pretrained weights from a ResNet-50 model to extract the features of the input image.

We used data augmentation techniques of random horizontal flips and random resized crops to improve the generalizability of the model's predictions, ran the extracted features through adaptive average 2-d pooling with 1x1 kernels, followed by batch normalization with a momentum value of 0.01. To encode the caption input, we used a single layer LSTM [9] with 300-dimensional word embeddings taken from GLoVe [10].

The decoder RNN used is a single layer LSTM as well, with 300-dimensional GLoVe embeddings and a default sequence length of 20, which is a reasonable cap on the number of hashtags on most Instagram posts. We use the teacher-forcing method with a ratio of 0.5 to train this sequential decoder.

Other relevant hyperparameters of interest are our use of Xavier initialization, a common choice while training multi-modal networks, the choice of Cosine Annealing Learning Rate initialized at 2e-3 as recommended by Loschilov *et al* [11] for stochastic gradient descent with warm restarts, the preference for the Adam optimizer and a batch size of 16. To quantitatively evaluate our results, we use the Bilingual Evaluation Understudy (BLEU) score [12] – a common yardstick in natural language problems such as machine translation and image captioning – averaging them over each batch of outputs.

As is often the case while working with very large datasets, hyperparameter tuning in our problem was slow, made especially tricky by the unavailability of limitless compute time. We eventually worked with a much smaller sample set (of ~700 images) to explore and experiment with the subtle interplay of hyperparameters on our dataset. Specifically, we tried different values of weight decay before ultimately settling down on 0.0001 as a suitable trade-off between overfitting and underfitting, different batch sizes to bypass frequent memory outage issues on the NVIDIA P100 GPU, and different learning rates based on the trends of the losses observed.

5. Results

As alluded to in the previous sections, there were a bunch of different data splits we were working with. The first one contained the entire set of images downloaded from the *instaloader* API which amounted to about 130,000 images in entirety. For the purpose of our experiments, we picked 20,000 images which were further split into train, val and test datasets in a 90:5:5 ratio.

For this problem, we've chosen to quantitatively measure performance with the help of two common metrics : 1) The loss, which was picked to be Cross Entropy and 2) The BLEU score, which is appropriate for the task owing to the natural language output from our LSTM decoder.

Our losses (Fig 3) showed a steady decline in both the training and the validation stages. It is noteworthy that our validation losses were consistently lower in magnitude compared to the training losses at every epoch. This suggests that our model was able to generalize better than expected on the validation set, under the combination of hyperparameters we finally zeroed in on.

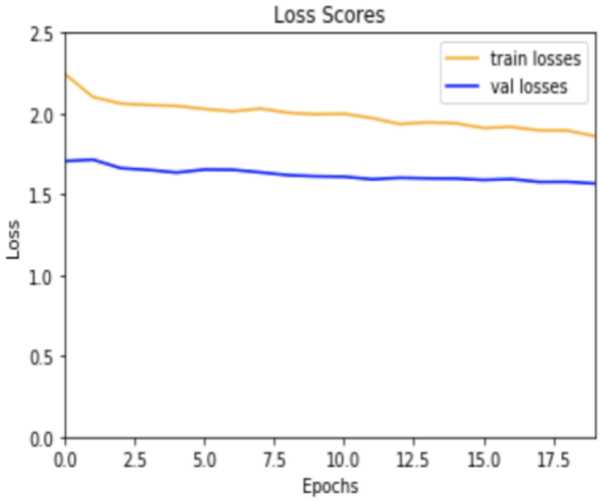

Fig 3. Our loss scores on the training and the validation datasets, showing a clear declining trend, with the validation losses being lower in magnitude.

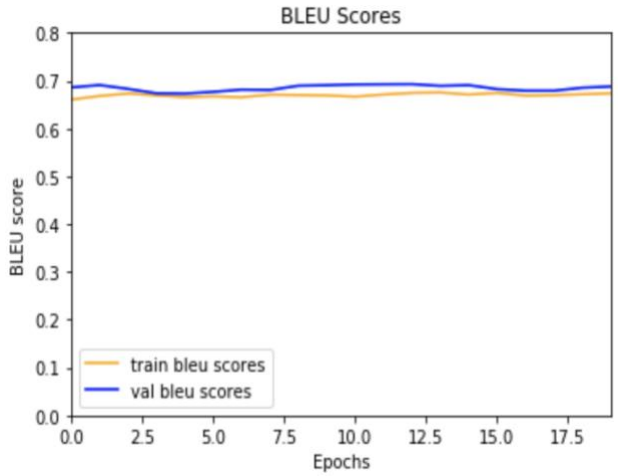

Fig 4. Our BLEU scores on the training and the validation datasets, oscillating in the range of 0.66 to 0.69, with the validation scores being slightly better at almost every epoch.

Our BLEU scores on the training and validation datasets too confirmed this estimate of the model generalizing its results well – owing to the consistently (albeit marginally) better score on the validation dataset compared to the training dataset at every epoch. The scores oscillated between 0.66 and 0.69 across the duration of training (Table 1).

	Loss Scores (Train)	Loss Scores (Val)	BLEU Scores (Train)	BLEU Scores (Val)
Highest	2.24	1.71	0.68	0.69
Lowest	1.86	1.57	0.66	0.67
Mean	1.99	1.62	0.67	0.68

Table 1. A summary of our loss and BLEU scores across the training and validation datasets.

Fig 5.
Caption : ***soja***.
Predicted Hashtags : ***#love #instagood #food #delicious***

Seen here in Fig 5 is an example of a qualitative output of our model. The caption for the image is 'soja' which is an alternative spelling for 'soya', a commonly consumed plant-based food product. We see that the model correctly identifies the presence of food in the image and assigns it relevant domain-specific hashtags used on Instagram such as '#love' (generic enough across a range of contexts on the platform, especially on posts related to food), '#food', '#instagood' (yet another generic caption across a range of happy situations) and '#delicious' which is clearly a hashtag specific to the content of the image.

Likewise, Fig 6 demonstrates a funny post with a sarcastic caption whose essence from both the image and the caption is accurately identified in the predicted hashtags.

Fig 6.
Caption : ***Summer vibes.***
Predicted Hashtags : ***#memes #funnymemes #lol #follow #fun***

6. Discussion

Our intent behind the project was to get a baseline model working on this new task and add as much complexity as time would permit. A task of this kind is, by nature, open-ended – especially with the surfeit of architectures for vision and language modelling being released in the research community on a regular basis.

Alternative implementations of this solution could explore different sequence modelling techniques such as attention and transformers [14] on the encoding and decoding stages to achieve better results on the captions

and the hashtags respectively. One of the other additions to this could also be explicitly encoding the sentiment of the caption and passing it along to the decoder, with the image and the original caption, as an explicit signal for generating the hashtags – since they could also vary based on the sentiment of the caption.

It may also be useful to model the hashtag generation as a 'vanilla' classification problem over a finite set of outputs. Although it would rob the problem of its temporal dependency on the decoder, the single most likely hashtag predicted for an image and a caption might still be appropriate and easier to train on.

7. Conclusion

We've defined a new task for hashtag generation on social media posts conditioned on an image and a caption, along with putting together a dataset for this purpose. We've demonstrated a baseline architecture based on the ResNet-50 convolutional feature extractor along with LSTM encoders and decoders and report a BLEU score of 0.7 on the validation split. We believe this work could be built upon extensively with different combinations of hyperparameters and model architectures to achieve even better results that handle the hugely complex perceptive and reasoning characteristics exhibited by posts made on social media platforms.

8. References

[1] Salman Aslam, Instagram By The Numbers, https://www.omnicoreagency.com/instagram-statistics/. Accessed 3 Dec 2019.

[2] Rebecca Hiscott, The Beginner's Guide To The Hashtag, https://mashable.com/2013/10/08/what-is-hashtag/. Accessed 3 Dec 2019.

[3] Danny Sullivan, How Google autocomplete works n Search, https://www.blog.google/products/search/how-google-autocomplete-works-search/. Accessed 3 Dec 2019.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. 2015

[5] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, Show and Tell : A Neural Image Caption Generator. 2015.

[6] Yann LeCun, Leon Bottou, Yoshua Bengio, Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. 1998.

[7] Dhruv Mahajan, Ross Girshik, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, Laurens van der Maaten. Exploring the Limits of Weakly Supervised Pretraining. 2018.

[8] Vahid Kazemi, Ali Elqursh, A Strong Baseline For Visual Question Answering. 2017.

[9] Sepp Hochreiter, Jurgen Schmidhuber, Long Short-Term Memory. 1997.

[10] Jeffrey Pennington, Richard Socher, Christopher D. Manning, Global Vectors for Word Representation. 2014.

[11] Ilya Loschilov, Frank Hutter, SGDR : Stochastic Gradient Descent with Warm Restarts. 2017.

[12] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, BLEU : A Method for Automatic Evaluation of Machine Translation. 2002.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, Illia Polosukhin. Attention Is All You Need. 2017.