InstaHash - Describing Images in a Human Fashion Using Instagram Hashtags

Samantha Ray
Texas A&M University
College Station, Texas, United States

Duc Hoang
Texas A&M University
College Station, Texas, United States

Josiah Coad
Texas A&M University
College Station, Texas, United States

Mahalakshmi Sridharan
Texas A&M University
College Station, Texas, United States

ABSTRACT

Machine image captioning currently focuses on correctly identifying the content of the image. We seek to improve machine description of images by improving their contextual understanding of the images akin to how a human would understand and interpret the image. To achieve this, we focus on developing a hashtag recommendation system that captions social media images based on both their content and their context. Our recommendation system combines the Word2Vec and AutoRec models to learn the associations between the an image's human-labeled hashtags and machine-generated descriptions of its contents. Our approach takes advantage of the strengths of word embeddings and collaborative filtering to recommend a list of hashtags for the image based on its description alone. Evaluation of the cosine similarity of the recommended hashtags to the ground truth hashtags shows that this method recommends relevant hashtags as approximately 95% have positive cosine similarity, 75% with similarity greater than 0.3.

KEYWORDS

image captioning, hashtags, collaborative filtering, word embeddings

1 INTRODUCTION

Image description tasks generally seek to caption the image by describing the objects in the image. Captioning images in this fashion can be useful for information extraction and computer vision tasks. However, algorithms for providing logical descriptors of an image only capture a relatively small amount of the information in the image. Beyond the objects in an image, an image caries more information in subtle cues that determines how a human would interpret and describe it. Social media images have rich meaning beyond the objects in the image. Users supplement their images with hashtags to emphasize the important elements in the image (e.g. #ring), give context to the image (e.g. #motivation or #goals), or express the user's feelings about the image (e.g. #love). Using a computer to describe images using hashtags serves as a unique variation of the image captioning problem. Our approach seeks to describe images in a human fashion by modeling the language dependencies between the hashtags assigned to the image and the objects in the image.

2 BACKGROUND

The social media platform Instagram consists of images with captions that other users can interact with via likes and comments. Instagram users can supplement their images with hashtags to make the pictures easier to find for others looking for that type of content. These hashtags can be interpreted as a succinct description of the image.

Park et al. constructed the HARRISON (HAshtag Recommendation for Real-world Images in SOcial Networks) dataset [5] for the purpose of creating hashtag recommendation systems. This dataset consists of 57,383 photos from Instagram with a hashtag vocabulary of 997 words. On average, each image has 4.5 hashtags and each hashtag appears 261 times. The major findings of this work include that recommending hashtags for an image requires contextual understanding of said image.

Attend to You [4] proposes a novel model for the task of personalized image captioning, i.e. captioning images in the style of a specific user. Cesc Park et al. collected 1.1 million Instagram posts from 6,300 users to train their model and made the dataset available as InstaPic-1.1M. This dataset has a hashtag vocabulary of 435,651 hashtags with each hashtag appearing 8 times on average. Notably, the vocabulary has a right skewed distribution where the most popular tag, love, appears 25,603 times, and over half of the hashtags only appear once, causing a median count of 1. If we ignore hashtags that appear only once, the mean becomes 20 and the median 4.

3 APPROACH

3.1 Overview

Our model takes an image and returns a list of hashtags that describe the image. The list of hashtags returned gives a socially-aware description of the image in contrast to an object-focused description. To accomplish this, our system combines two models that learn the relationships between the hashtags and the image descriptions: Word2Vec [3] [6] and AutoRec [7].

The Word2Vec model learns dense vector representations of the words in the vocabulary. These word vectors can be used to find the most similar words to a sequence of words using cosine similarity. The AutoRec model, a variation of collaborative filtering, learns associations between the words in the vocabulary so that it can rank all of the words in the vocabulary given an input set of words. As both models generate distinct recommendations for a given image description, InstaHash takes the intersection of the two results sets

to determine the final output of the recommender. If the two models produce disjoint recommendations, the recommender defaults to the top 5 ranked hashtags. This methodology results in the system only recommending hashtags with high confidence.

3.2 Data Preprocessing

To reduce the sparsity present in the InstaPic-1.1M hashtag vocabulary, we apply the preprocessing steps performed on the HARRISON dataset to the InstaPic-1.1M dataset. This process involves limiting the hashtag vocabulary to the 1000 most common hashtags, filtering the image hashtag sets to contain only the hashtags in this smaller vocabulary, and removing any image that had no hashtags remaining. The filtered InstaPic-1.1M dataset has 380 thousand images with a mean hashtag count of 1,217 and median hashtag count of 706.

Object Detection The Instagram datasets do not contain image descriptions beyond the captions or broad category of the image in the case of HARRISON, so we use object detection algorithms [2] [1] to describe the contents of the images. We use the Google Cloud Vision API to obtain the objects contained within an image. We reject all objects returned from the Google API with less than 50% confidence. On average 9.5 words describe an image in our dataset. Some of these object detections clearly read as computer generated, e.g. leg, but sometimes the Google API returns a more high-level description such as "atmosphere".

3.3 Word2Vec

Word2Vec models the dependencies of natural language through how related words appear in similar contexts, allowing words to be translated into a dense vector representation. However, most English Word2Vec models train on corpuses such as news documents and thereby only contain representations of formal English. Instagram hashtags on the other hand contain informal English, acronyms, and contemporary slang due to the casual, social nature of the media platform. Consequently, InstaHash uses Gensim's Word2Vec model to train its own word embeddings for the unique vocabulary of Instagram hashtags. Each sample in the training corpus contains the words present in the corresponding image's hashtags and the objects from its description. The order of the words in the sample does not matter due to the independent nature of hashtags, so these samples consist of true bags of words. However, Word2Vec takes sequences of words as input, so it would learn time dependencies that do not actually exist in the data. To account for this, we create permutations of each of sample to train the Word2Vec on every possible order of the words in the sample. Due to the factorial growth of the number of permutations consisting of n objects taken r at a time, we set r to be a small number such as 3 as samples have an average length of 13.5 for the HARRISON dataset and 12 for the InstaPic-1.1M dataset.

3.4 AutoRec

AutoRec, a collaborative filtering system with an autoencoder architecture, learns profiles of opinions on a set of items to make its recommendations about other items. Generally, these profiles consist of opinions in the form of ratings on items such as movies. With minor modification, this architecture can learn profiles consisting of image descriptions with items being words in a vocabulary. Representing the presence or absence of a word as a positive or negative rating, respectively, AutoRec can recommend a ranked list of words. Normally, AutoRec uses sparse training where items with no labels receive zeros to signify an unknown opinion. However, the image descriptions do not have negative samples inherently because they only contain words that describe the images. With no negative samples, the AutoRec model could not learn what words do not have associations as the weights could only increase during training. To fix this, we enforce negative ratings on unknown labels, interpreting the absence to mean that this word does not describe the image.

3.5 Recommender

To further describe the image, the Instagram-trained Word2Vec model outputs the n, e.g. n=20, most similar words in the vocabulary to the set of words from the ground truth hashtags and from the Google API image description to serve as an expanded image description.

The AutoRec model takes the set of the image's hashtags, the image description from the Google API, and the Word2Vec expanded image description as input. During training, words belonging to the image's hashtags or description set receive a positive signal to signify that these words describe the image. The Word2Vec expansion receives a decaying rating to account for the fact that the recommended words become less similar to the description as n increases. The rest of the vocabulary receives a negative signal to signify that the author and the description generation models did not choose these words to describe the image.

After the AutoRec model finishes training to 0.28 RMSE after 50 epochs, the AutoRec model takes just the Google API description as a sparse input. In other words, any word in the input set receives the positive signal while the rest of the vocabulary receives zeros. This allows nodes in the output layer to activate based upon the learned weights.

At the recommendation phase of the system, both the Word2Vec model and AutoRec model generate a ranked list of hashtags using the image description as input. The system takes the intersection of the two output sets as the initial recommendation as both models endorsed the selected words. If the intersection results in an empty set, the system defaults to recommending the top five recommendations of the AutoRec result set. To match the domain of the task, we filter the intersection to contain only words that appeared in the hashtag vocabulary. The system recommends this final set of hashtags as the context-aware description of the image.

4 RESULTS

4.1 Evaluation

The images have a ground truth description consisting of the author's hashtags. While these hashtag sets do not necessarily describe the content of the image, i.e. some images only have abstract hashtags such as #motivation, we take them as the best description of the image to compare our results against as the author knows the full context and inspiration for the image.

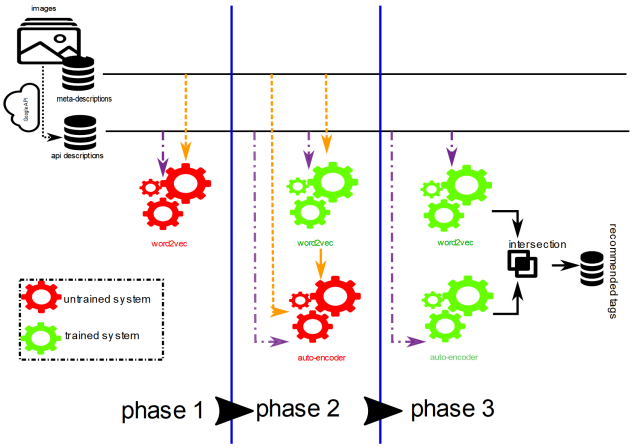

Figure 1: The InstaHash Model

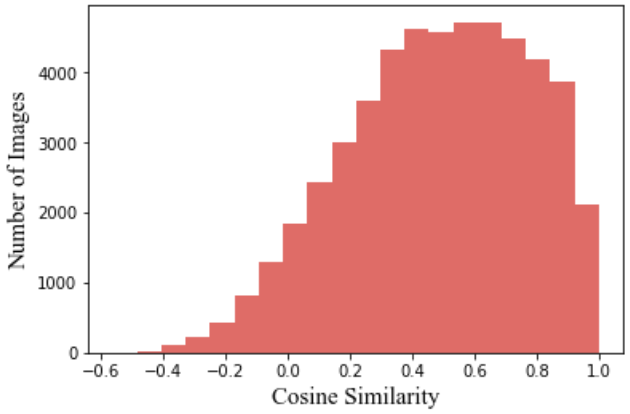

Figure 2: Instapic-1.1M - Cosine Similarity to Ground Truth

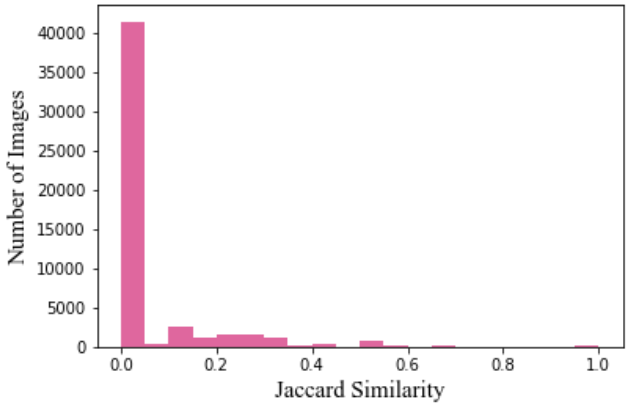

Figure 3: Instapic-1.1M - Jaccard Similarity to Ground Truth

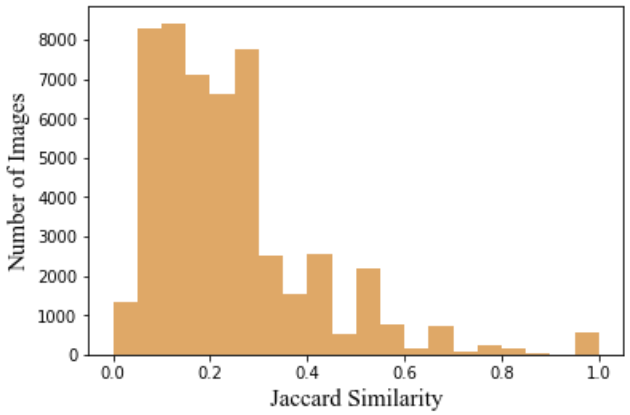

Figure 4: Instapic-1.1M - Jaccard Similarity to Description

Our model performs well with respect to generating relevant hashtags. To gauge the relevancy of the output set of hashtags to the ground truth from the original image, we take the cosine similarity of the image's hashtags and the generated hashtags using our Instagram-trained Word2Vec model. Negative cosine similarity strongly indicates that the generated hashtags do not match the original image. Positive cosine similarity, on the other hand, does not prove to be a strong signal of quality as words can be clustered around a topic without being a good hashtag, e.g. #dog and #canidae. Nonetheless, positive cosine similarity does indicate the relevance of the recommendation as the words exist near each other in vector space.

We take 0.3 as the lower bound for a good recommendation and 0.7 as the lower bound for a great recommendation. Highly-related words may not have the highest cosine similarity, e.g. #brother and #sister have a cosine similarity of 0.7432 in GloVe Word2Vec model despite being gendered variations of the same concept. Low to negative cosine similarity signifies a bad recommendation.

To assure that our model generated novel hashtags, i.e. the model did not simply repeat the image description or ground truth, we also assessed the output based on its Jaccard similarity to the ground truth. Lower Jaccard similarity paired with higher cosine similarity means that the generated hashtags matched the input image while demonstrating that the model had learned the relationship between words in the vocabulary.

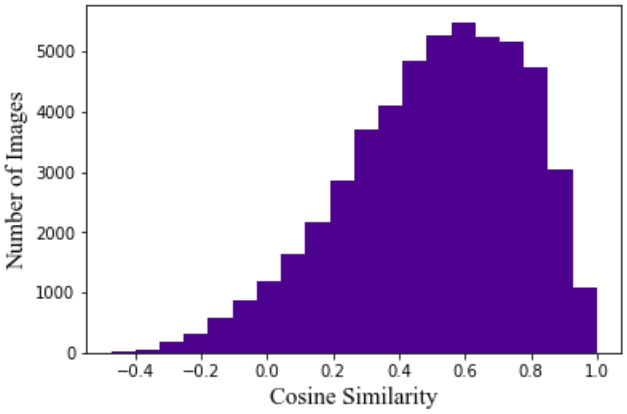

Figure 5: HARRISON - Cosine Similarity to Ground Truth

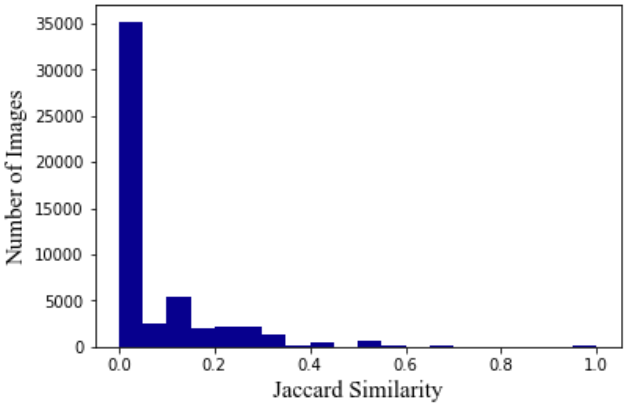

Figure 6: HARRISON - Jaccard Similarity to Ground Truth

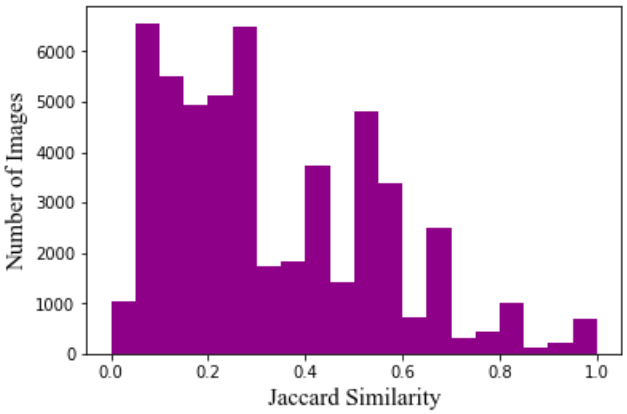

Figure 7: HARRISON - Jaccard Similarity to Description

Table 1: Cosine Similarity to Ground Truth

	Mean	> 0.7	> 0.3	< 0.3	< 0.0
HARRISON	0.5060	0.2735	0.7782	0.2218	0.0473
InstaPic-1.1M	0.4832	0.2710	0.7289	0.2711	0.0633

Table 2: Jaccard Similarity to Ground Truth

	Mean	> 0.7	> 0.3	< 0.3
HARRISON	0.0682	0.0038	0.0606	0.9416
InstaPic-1.1M	0.0497	0.0043	0.0552	0.9458

Table 3: Jaccard Similarity to Description

	Mean	> 0.7	> 0.3	< 0.3
HARRISON	0.3289	0.0888	0.4836	0.5647
InstaPic-1.1M	0.2347	0.0201	0.2747	0.7678

Shortcomings of this model include the dense clustering of hashtags for certain topics of hashtags and the lack of world knowledge in recommendations. Some hashtags appeared so often in the same context that the model functionally recommended them as a group. These dense categories include dogs, cats, food, Instagram-related words, and weddings. Lack of world knowledge refers to when the model recommends a synonym that does not work for the image given its context. For example, #brother and #boyfriend exist nearby in the vector space, but a human knows these words cannot be interchanged while picking hashtags. Other cases involve when a word exists nearby in vector space but has nothing to do with the image, e.g. #banana on a photo that contains non-banana fruit.

4.2 Examples

We provide detailed examples in Tables 4, 5, 6, and 7 to demonstrate how our system performs on different types of image and hashtags combinations. The ground truth hashtags can be content-related, context-related, and/or abstract. Naturally, content-related hashtags describe the content of the image. Context-related hashtags describe elements about the image such as the location or the author's opinion. Abstract hashtags consist of hashtags that have no strong connection to the content or context of the image.

5 CONCLUSION

5.1 Discussion

Our model only used a subset of the InstaPic-1.1M dataset images and restricted the vocabulary due to computational limitations. We want to train our system on all our available data to assess the effect on the quality of recommendations and better model the scope of the hashtag recommendation task. Additionally, our evaluation only considers the relevancy of the recommendations with respect to the given image's original hashtags in terms of the cosine similarity of their words. We want to perform a more thorough evaluation of

Table 4: HARRISON Sample Image - Image of Boat with Descriptive Hashtags

	Hashtags
GT	sky boat sea cloud
Desc	calm cloud sunset ocean horizon vehicle sky morning boat sea
Word2Vec	horizon sunrise sunset calm sea afterglow cloud resource coast sky ocean wave boat water lake view reflection daytime wind coastal
AutoRec	sky cloud sunset horizon sea sunrise morning ocean
Intersect	horizon sunrise sunset sea cloud sky ocean
Recommend	horizon sunrise sunset sea cloud sky ocean
Sim. GT	0.8465
Sim. Desc	0.9591
Jaccard GT	0.3750
Jaccard Desc	0.5455

Table 5: InstaPic-1.1M Sample Image - Image with Non-Descriptive Hashtags

	Hashtags
GT	colorado
Desc	fun sibling youth family people friendship happy child smile event
Word2Vec	people friendship youth smile fun happy team photobombing father child social daughter sisters community laugh mother bestfriends interaction sibling grandparent
AutoRec	smile family fun happy people friendship event child
Intersect	people friendship smile fun happy child
Recommend	friendship smile fun happy
Sim. GT	0.2495
Sim. Desc	0.9635
Jaccard GT	0.0000
Jaccard Desc	0.4000

the output of our recommender system by having humans judge the quality of our generated hashtags.

5.2 Future Work

Online and Related Alternatives: The model makes offline recommendations based off the image's description. A natural follow-up to this system involves making online recommendations as a user selects hashtags and learning from these selections. If continuously trained on online image-hashtag pairs (including data from other social media platforms such as Facebook or Twitter), the model will gain the ability to recommend trending hashtags. Data augmentation, an inexpensive alternative to making the model live online, will extend the offline data in hand to be more representative of the online data while avoiding issues like overloading the system.

Balancing accuracy and novelty: Models that recommend hashtags fairly close to the ground truth will help enhance the existing recommender in achieving a reasonable balance between recommending ground truth tags and recommending novel tags. Initial experiments conducted by training a feed-forward fully-connected neural network with the term frequency (TF) vector representations of the image descriptions as input and hashtags as output resulted in a simple yet robust model that recommended at least one of the ground truth hashtags for 77% of the test data. Notably, the neural network recommended specific ground truth tags like #vsco, #vscocam that AutoRec generally did not. Additionally, hashtags not present in the ground truth tended to be closely related in meaning, e.g. sky and clouds, shoes and boots, and landscape and nature.

Capturing emotions and Personalization: Combining AutoRec with VGG16 pretrained model will pave the way for more accurate emotion-based hashtag recommendations. Additionally, the extending the input features to include the demographics and past hashtags of the user will enable the personalization of the recommendations as supported by prior works in hashtag recommendation.

Table 6: InstaPic-1.1M Sample Image - Image with Abstract Hashtags

	Hashtags
GT	shoplocal vsco
Desc	picture frame branch furniture rectangle black tshirt
Word2Vec	frame picture rectangle photographic bed bedroom paper wall bedding collection shelf couch mattress closet sheet room design furniture square visual
AutoRec	black tshirt
Default	black room furniture vsco vscocam
Recommend	black furniture vsco vscocam
Sim. GT	0.5373
Sim. Desc	0.7362
Jaccard GT	0.2000
Jaccard Desc	0.2222

Table 7: InstaPic-1.1M - Image with Mixed Hashtags

	Hashtags
GT	dogsofinstagram sunday
Desc	companion rare dog puppy vertebrate canidae breed mammal carnivore
Word2Vec	breed companion canidae dog puppy carnivore sporting vertebrate terrier dogsofinstagram mammal retriever snout chihuahua labrador dogs instadog petsofinstagram rare cur
AutoRec	mammal carnivore vertebrate
	dog breed canidae companion puppy terrier dogsofinstagram
Intersect	breed companion canidae dog puppy carnivore vertebrate terrier dogsofinstagram mammal
Recommend	dog puppy dogsofinstagram
Sim. GT	0.8773
Sim. Desc	0.9510
Jaccard GT	0.2500
Jaccard Desc	0.1818

REFERENCES

[1] Google Developers. 2019. Vision AI. https://cloud.google.com/vision/

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. *CoRR* abs/1703.06870 (2017). arXiv:1703.06870 http://arxiv.org/abs/1703.06870

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[4] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to You: Personalized Image Captioning with Context Sequence Memory Networks. In *CVPR*.

[5] Minseok Park, Hanxiang Li, and Junmo Kim. 2016. HARRISON: A Benchmark on HAshtag Recommendation for Real-world Images in Social Networks. *CoRR* abs/1605.05054 (2016). arXiv:1605.05054 http://arxiv.org/abs/1605.05054

[6] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.

[7] Suvash Sedhain, Aditya Krishna Menon, Scott Sanner, and Lexing Xie. 2015. AutoRec. In *Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion*. ACM Press. https://doi.org/10.1145/2740908.2742726