Learning User Representations for Open Vocabulary Image Hashtag Prediction

Thibaut Durand
Borealis AI Simon Fraser University
thibaut.durand@borealisai.com

Abstract

In this paper, we introduce an open vocabulary model for image hashtag prediction – the task of mapping an image to its accompanying hashtags. Recent work shows that to build an accurate hashtag prediction model, it is necessary to model the user because of the self-expression problem, in which similar image content may be labeled with different tags. To take into account the user behaviour, we propose a new model that extracts a representation of a user based on his/her image history. Our model allows to improve a user representation with new images or add a new user without retraining the model. Because new hashtags appear all the time on social networks, we design an open vocabulary model which can deal with new hashtags without retraining the model. Our model learns a cross-modal embedding between user conditional visual representations and hashtag word representations. Experiments on a subset of the YFCC100M dataset demonstrate the efficacy of our user representation in user conditional hashtag prediction and user retrieval. We further validate the open vocabulary prediction ability of our model.

1. Introduction

Understanding the content of an image is a fundamental and challenging computer vision task because an image can contain a large variety of semantic concepts and a semantic concept can have diverse visual appearances. The vocabulary to describe visual concepts in an image is very large and it is necessary to go beyond the semantic categories used in standard classification datasets (*e.g.* COCO [31], ImageNet [36]) which focus on a small subset of categories that have precise physical description. Moreover, the annotations of these datasets are independent of the users who take the picture and ignore sentiment/subjective concepts like *fun* or *happy*.

The hashtag problem serves as a lens into the general problem of image understanding because the user's intent is not separable from the image content. Images annotated with hashtags are available in great abundance because of social networks, but hashtags are inherently subjective because they are provided by users as a form of self-expression [43]. As a consequence, hashtags may have synonyms (different hashtags referring to the same visual content) and may be ambiguous (the same hashtag referring to different visual content). This self-expression leads to user-specific variation in hashtag supervision that is independent of the image content, and therefore limits the effectiveness of standard image classification methods. To overcome this problem, [43] introduced a user-specific model that models the joint distribution of images, hashtags and users instead of the image-hashtag pairs as in standard image classification models. But this approach has two main limitations: it cannot deal with new users or new hashtags without retraining the model.

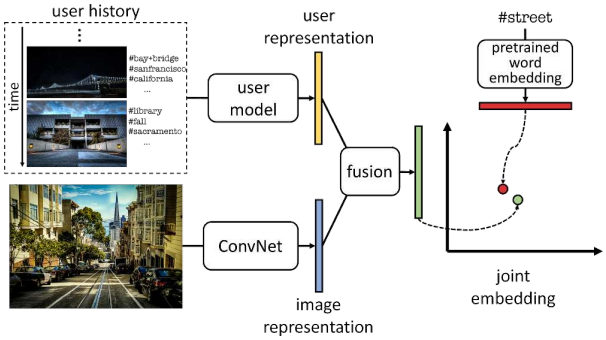

Figure 1. Overview of the proposed user conditional joint embedding model for hashtag prediction. First, a user representation (yellow) is extracted from the user history. Then, the model computes a user conditional image representation (green) which is projected into a joint embedding space with the hashtag representation of `#street` (red).

To deal with new users, we propose a new model that extracts a representation of a user from his/her image history (top-left of Figure 1). Given a user, our model only uses the images and their corresponding hashtags from the user's history to compute a user representation – our model works like a standard image classification model if a user does not have an image history. Our approach is inductive and can extract a representation of a new user without retraining the model. Another advantage of our model is that

it constructs a dynamic (time-varying) user representation so it can easily improve the representation of a user if new images from this user are available. Then, the user representation (yellow) is fused with the image representation (blue) to compute a user conditional image representation (green).

Unlike existing hashtag prediction models for images [12, 43] or text [47, 13], we propose an open vocabulary hashtag prediction: our model can generalize and map across new concepts that have not been seen at training time. Open vocabulary models are important because social networks are constantly evolving: user interests can change quickly and new hashtags appear frequently. Our model uses pretrained word embeddings to represent each hashtag in a continuous and semantic space (red in Figure 1), and then the hashtag representations are projected into a joint embedding space with the user conditional visual representation. A continuous semantic embedding space is more appropriate than using separate classifiers because it can share knowledge between synonymous hashtags. Similarly, it can deal with the long tail distribution problem (infrequent hashtags) and unseen hashtags because it can share information between hashtags. We show that our model is scalable and can deal with a vocabulary of 550k hashtags. Finally, our approach is symmetric and can be used for both image-to-hashtag and hashtag-to-image retrieval.

To summarize, our contribution is threefold. **(1)** We propose a new model to extract a user representation from his/her image history. This approach allows to deal with new users or to improve a user representation with new images without retraining the model. **(2)** We introduce an open vocabulary model based on pre-trained word embeddings that can deal with infrequent hashtags and hashtags unseen during training. **(3)** Our experiments show that the image history can be used to extract effective user representations. We investigate the efficacy of our user representation for both user-specific image tagging and user retrieval. We also evaluate the ability of our model to generalise to predict hashtags unseen during training.

2. Related Work

Image tagging with user representation. Recent works [12, 43] show that modelling the user is important to analyze images annotated with hashtags because of the self-expression problem. Denton et al. [12] introduced a user representation that exploits user metadata (age, gender, GPS coordinates and country). Even if this user representation can deal with geographical domain shift [38] (the same semantic category of objects can look quite different in images taken in different geographical locations), it cannot fully represent a user because these user metadata are not informative enough to fully capture user behaviour. Another limitation is that it is not always possible to have access to user metadata. To overcome this problem, Veit *et al*. [43] proposed to learn an embedding for each user based on the images and the corresponding hashtags. However, learning an embedding per user is limited to a transductive setting; it is not applicable to new users. In this paper, we propose a model that can capture user behaviour and deal with new users without retraining the model. Our model extracts a representation of a user by only exploiting the images with their corresponding hashtags from his/her image history. Our model can also improve a user representation with new images without retraining the model. Note that [47, 23] also address the problem of hashtag prediction but do not model the user.

Conditional models for visual recognition. Our work is related to conditional models for visual recognition. The most popular example is probably the Visual Question Answering (VQA) task [19, 25, 49, 14] where the input image is conditioned by a question. Recently, [39] proposed a model for the personality-captions task by conditioning the input image on the given style and personality traits. While [39] uses an addition to fuse the visual and the personality representation, we use a bilinear product as in most of the VQA models to fuse the visual and the user representation. Our model is also related to the Conditional Similarity Networks [42] that learn embeddings differentiated into semantically distinct subspaces to capture different notions of similarities. However this model can only deal with a fixed number of similarities.

Open vocabulary. Standard image classification models [28, 22] are not suitable for open vocabulary prediction because the classes are usually fixed before training and the models are designed to predict among those classes for a given image. [18] introduced a vocabulary-free image tagging model, that uses an image search engine to collect images for each tag in the vocabulary, but it cannot deal with new hashtags after training. A strategy to deal with new categories is to use a Zero-Shot Learning (ZSL) model [29, 30]. ZSL models are learned on some categories and tested on other categories based on the knowledge extracted during training [35]. A more realistic scenario is the Generalized Zero-Shot Learning (GZSL) [9, 48] where both seen and unseen classes are present at test time. A lot of ZSL/GZSL models [30, 17, 2, 3, 50, 1, 7, 48, 45] learn an embedding between a visual space and a semantic space (attributes, text description). Our model is similar to [17], but the main difference is that [17] is designed for single-label object classification whereas our model works for multi-label classification with a large and diverse hashtags set, which can represent abstract concepts like *fun*. Another important difference is that [17] preprocesses the labels based on the WordNet hierarchy to clean the vocabulary and avoid synonyms whereas our model works for hashtags in the wild without this preprocessing.

Multi-modal embeddings. Over the past few years, a lot of models using visual-text embeddings [27, 20, 46, 37, 16, 15, 8, 39, 44] have been proposed for several applications. Today, most of the methods that build cross-modal embeddings between text and images use a triplet loss [27]. While the original triplet loss averages over all triplets in the mini-batch, [16] introduced a hard negative sampling because the average strategy can lead to vanishing gradients since, as the optimization progresses, most of the triplets tend to contribute less to the error. [16, 21, 15] observe a significant improvement by using hard negatives in the loss. However, the hard negative triplet loss is sensitive to noise/outliers and needs a few epochs to "warm up" at the beginning of the learning process because a very limited amount of triplets are contributing to the gradient, when many are violating the constraints. Recently, [8] introduced an adaptive strategy that automatically adapts the number of triplets used in the loss. These triplet losses work well for tasks like caption retrieval [16] because the number of triplets is the size of the mini-batch, but they are not scalable for our task because the hashtag vocabulary is too large (> 400k). The complexity is exacerbated for the multi-label setting because each example can be a positive example for several hashtags. [21] show that randomly sampling some triplets is not helpful because most of the triplets incur no loss and therefore do not improve the model. Moreover it is difficult to define negative examples because hashtags have synonyms.

3. Model

Our goal is to learn a user-specific hashtag prediction model. Our model uses the user image history to compute the user representation and hence it can deal with new users. We first present our model to extract a user representation from an image history and then our user conditional joint embedding model for open vocabulary hashtag prediction.

Notations. We denote by $\mathcal{U} = \{u_1, \dots, u_U\}$ a set of U users and a vocabulary of K hashtags by $\mathcal{H} = \{h_1, \dots, h_K\}$. In the open vocabulary setting the vocabulary of hashtags for training is a subset of the vocabulary of hashtags for testing, *i.e.* $\mathcal{H}^{train} \subset \mathcal{H}^{test} = \mathcal{H}$, whereas in the fixed vocabulary setting (standard setting for image classification) the vocabulary of hashtags is the same for training and testing, *i.e.* $\mathcal{H}^{train} = \mathcal{H}^{test} = \mathcal{H}$. For each user u, we have access to an ordered list by time[1] of N_u images with their associated hashtags: $\bar{\mathcal{I}}^{(u)} = [(\mathcal{I}_1^{(u)}, \mathcal{H}_1^{(u)}), \dots, (\mathcal{I}_{N_u}^{(u)}, \mathcal{H}_{N_u}^{(u)})]$, where $\mathcal{I}_j^{(u)}$ is the image and $\mathcal{H}_j^{(u)} \subset \mathcal{H}$ is the nonempty hashtag set of the j-th image. Each image is associated with a unique user and we use disjoint sets of users for training and testing.

[1]This constraint can be satisfied by using the uploaded time on social networks.

Model overview. We define our problem as an automatic image labelling based on inferring hashtags, conditioned on an image $\mathcal{I}$, and a user u. During training, we aim at learning a model f that outputs the probability distribution over a tag y_i conditional on the image $\mathcal{I}$ and the user u:

$$p(y_i = 1 | \mathcal{I}, u; \Theta) = f(\mathcal{I}, u, y_i; \Theta) \tag{1}$$

where Θ are the whole set of parameters of the model. The general architecture of our approach is shown in Figure 1. Our model first extracts a representation of the user from his image history (yellow vector) and a visual representation of the image (blue vector). Then, these representations are fused to compute a user conditional image representation (green vector). Finally, the model learns a joint embedding between the user conditional image representations and the hashtag representations (red vector).

3.1. User representation

A key component of our model is the user representation because hashtags are inherently subjective and depend on the user. To extract a representation of a user, we propose to exploit his/her image history. Our approach allows to extract a user representation of a new user by exploiting only the image history and without retraining the model. Extracting a good user representation is a challenging problem because the user representation should encode some information about the user, for instance the hashtags used (each user only uses a small subset of hashtags based on topics of interest), language (English, Spanish, French, *etc.*), but also the correlations between images and the hashtags.

We now explain our method to extract a user representation which is shown in Figure 2. Given a user u, we assume that we know his/her image history (or a subset) $\bar{\mathcal{I}}^{(u)}$. Thereafter, we ignore the u notation for the sake of clarity because we only consider one user. To predict the hashtags of the T-th image, we use the $T-1$ past images and their corresponding hashtags to extract the user representation $\mathbf{u}_{1:T-1} \in \mathbb{R}^{d_u}$. The model first extracts a representation for each image-hashtags pair in the user history. Then, it aggregates these representations with a Gated Recurrent Unit (GRU [10]) to compute the user representation.

Image-hashtags representation. The goal is to compute a vectorial representation of each image-hashtags pair. We first extract a visual representation for each image in the user history with a ConvNet f^{im}:

$$\mathbf{x}_t^{im} = f^{im}(\mathcal{I}_t) \in \mathbb{R}^{d_i} \qquad \forall t < T \tag{2}$$

Similarly, we compute a representation of the hashtags associated with each image. We first extract a word representation for each hashtag (Section 3.2.2), then we sum

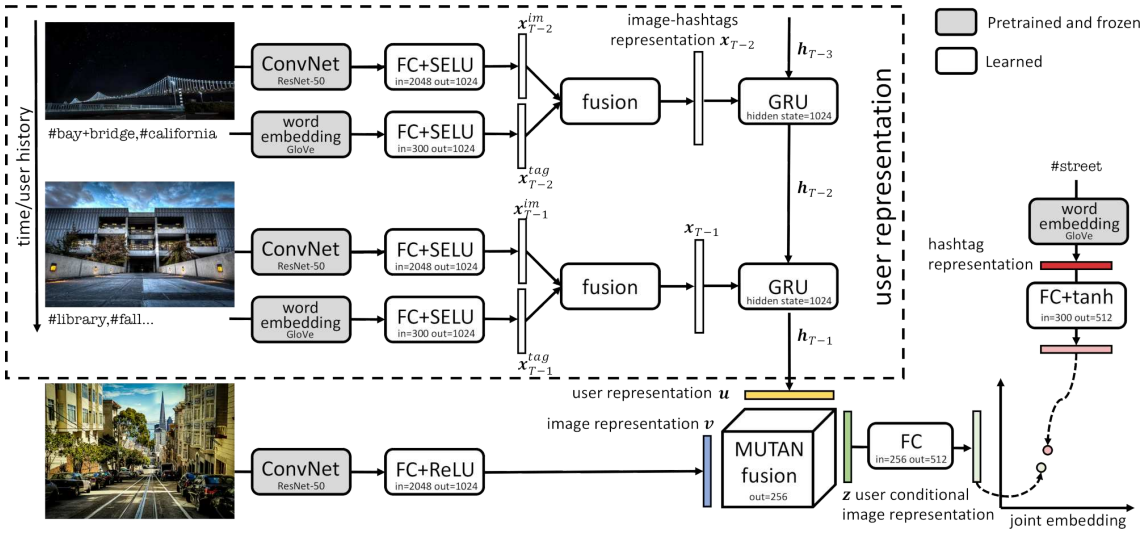

Figure 2. Our open vocabulary model for hashtag prediction where the user representation is extracted based on the user history.

each hashtag representation to have a fixed size representation $\mathbf{y}_t$, and finally we learn a non-linear mapping f^{tag}:

$$\mathbf{x}_t^{tag} = f^{tag}(\mathbf{y}_t) \in \mathbb{R}^{d_t}, \quad \mathbf{y}_t = \sum_{y \in \mathcal{H}_t} \psi(y) \quad \forall t < T \quad (3)$$

where $\psi(y) \in \mathbb{R}^{d_w}$ is a pretrained word embedding of hashtag y. If a hashtag is composed of several words, we sum the representation of each word *e.g.* $\psi(\texttt{black+white}) = \psi(\texttt{black}) + \psi(\texttt{white})$. If a hashtag does not have a word representation, it is possible to approximate it by using some algebraic properties. Pretrained word embeddings are used as auxiliary information to share information between hashtags so that the knowledge learned from seen hashtags can be transferred to unseen hashtags. It also allows to deal with the long-tail distribution problem because it can transfer knowledge from the data-rich head to the data-poor tail hashtags. In our experiments we use GloVe [34], but our model works with any word embeddings (*e.g.* [33, 47, 6]). Note that these word embeddings do not require additional supervision because they are learned in an unsupervised way from large text corpora.

Finally, we aggregate the image and hashtag representations to compute a representation of each image-hashtags pair:

$$\mathbf{x}_t = fusion(\mathbf{x}_t^{im}, \mathbf{x}_t^{tag}) \qquad \forall t < T \qquad (4)$$

We use an element-wise product to fuse the two modalities. In Section 4.2, we analyze several fusion operators and we observe that the choice of the fusion is important.

User representation. The goal is to compute a fixed size user representation $\mathbf{u}_{1:T-1}$ given the set of features $\{\mathbf{x}_t\}_{t=1,\dots,T-1}$ representing each image-hashtags pair of the user history. To take into account the temporal information of the images, we use a Gated Recurrent Unit [10]:

$$\mathbf{h}_t = f_{GRU}(\mathbf{x}_t, \mathbf{h}_{t-1}) \qquad \forall t < T \qquad (5)$$

where $\mathbf{h}_t$ is the hidden state of the GRU at step t and $\mathbf{h}_0 = \mathbf{0}$. GRUs turn variable length sequences into meaningful, fixed-sized representations. The last hidden state $\mathbf{h}_{T-1}$ is used as user representation $\mathbf{u}_{1:T-1}$. To aggregate the image-hashtags representations, it is possible to use any pooling function (*e.g.* max, average), but our experiments show that taking into account the temporal information improves performance. Thereafter, we use all the previous images as the user history and we denote the user representation by $\mathbf{u}$ for the sake of clarity. Note that it is possible to replace the GRU by other temporal models like TCN [4].

3.2. User conditional joint embedding model

We now introduce the user conditional joint embedding model. Given an image and user representations, our model first computes the user conditional image representation and then it learns a joint embedding between the user conditional image and hashtag representations.

3.2.1 User conditional image representation

The image $\mathcal{I}$ and user u are firstly embedded into vectors $\mathbf{v}$ and $\mathbf{u}$ respectively. We use a ConvNet to extract a fixed-size vectorial representation $\mathbf{v} \in \mathbb{R}^{d_v}$ of the visual content of an image. We use a different ConvNet from the ConvNet used in the user model because these two networks have different goals (representing an image vs representing a user). Experimentally, we observe that using separate networks improves performance. The image and user representations $\mathbf{v}$ and $\mathbf{u}$ are then fused using a bilinear operator to produce a user conditional image representation

$\mathbf{z} \in \mathbb{R}^{d_c}$. Bilinear models are powerful solutions used in particular in computer vision to capture multi-modal interactions [12, 19, 25, 43]. The bilinear model is more expressive than straightforward concatenation, element-wise product, or element-wise sum and is defined as follows:

$$z_j = \mathbf{v}^T \mathbf{W}_j \mathbf{u} + b_j \qquad j \in \{1, \dots, d_c\} \tag{6}$$

where $\mathbf{W}_j \in \mathbb{R}^{d_v \times d_u}$ is a weight matrix and $b_j \in \mathbb{R}$ is a bias of the j-th dimension. $\mathbf{z} = [z_j]_{j=1,\dots,d_c}$ is the output of the bilinear model and represents the image-user pair. We need to learn the tensor $\mathbf{W} = [\mathbf{W}_j]_{j=1,\dots,d_c} \in \mathbb{R}^{d_v \times d_u \times d_c}$ and the bias $\mathbf{b} = [b_j]_{j=1,\dots,d_c} \in \mathbb{R}^{d_c}$.

3.2.2 Joint embedding

In this section, we introduce our joint embedding model that can deal with hashtags unseen during training (Figure 2). Our aim is to learn functions that take the representation of an arbitrary hashtag and a user conditional image representation as inputs and embed them into a joint embedding. To learn the joint embedding space, we define a similarity function between the two modalities. We first project each modality in a joint embedding space by learning a mapping function $\phi^{iu} : \mathbb{R}^{d_c} \rightarrow \mathbb{R}^d$ (resp. $\phi^{tag} : \mathbb{R}^{d_w} \rightarrow \mathbb{R}^d$) from the user conditional image (resp. hashtag) space to the joint embedding space. Then, we define the similarity function in the joint embedding space to be the usual inner product. Given a user conditional image representation $g(\mathbf{v}, \mathbf{u})$ $(= \mathbf{z})$, we compute the compatibility score of any given hashtag y as follows:

$$f(\mathbf{v}, \mathbf{u}, y; \Theta) = \phi^{iu}(g(\mathbf{v}, \mathbf{u}))^T \phi^{tag}(\psi(y)) \tag{7}$$

The intuition is to maximize the similarity between the user conditional image representation and its associated hashtags in the joint embedding space. Unlike standard visual-semantic embeddings, our joint embedding also depends on the user, so an image can be mapped to different points in the joint embedding space with respect to the user profile. Note that unlike existing image hashtag prediction models [12, 43], our model is scalable because the number of learnable parameters of our model is independent of the hashtag vocabulary size.

3.3. Learning

Our training objective is to increase the similarity to the present hashtags, while decreasing the similarity to the other hashtags. Because the triplet loss commonly used to learn joint embedding is not scalable, we employ a classification loss for this task. Recent works [24, 40, 43, 32] suggest that softmax classification can be very effective even in multi-label settings with large numbers of classes such as ours. Given a user u and an image $\mathcal{I}_n$, the posterior hashtag probability is:

$$p(\hat{y}|\mathcal{I}, u; \Theta) = \frac{\exp(f(\mathcal{I}, u, \hat{y}; \Theta))}{\sum_{y \in \mathcal{H}^{train}} \exp(f(\mathcal{I}, u, y; \Theta))} \tag{8}$$

The probability distribution is computed only on the hashtags known during training ($\mathcal{H}^{train}$). Following [24, 43], we select a single hashtag $\hat{y}_n^{(u)}$ uniformly at random from hashtag set $\mathcal{H}_n^{(u)}$ as target class for each image. All the weights except the ones for the ResNets (due to the limitation of GPU memory) are optimized jointly in an end-to-end manner by minimizing the negative log-likelihood of the probability distribution:

$$\mathcal{L}(\Theta) = -\frac{1}{U} \sum_{u \in \mathcal{U}} \frac{1}{N_u} \sum_{n=1}^{N_u} \log p(\hat{y}_n^{(u)}|\mathcal{I}, u; \Theta) \tag{9}$$

Due to technical constraints, it is not possible to have several users in memory at the same time. A mini-batch contains the consecutive images of a single user.

4. Experiments

Implementation details. We use PyTorch in our experiments and each experiment runs on 1 GPU. We train our model using ADAM [26] during 20 epochs with a start learning rate 5e-5. We use ResNet-50 [22] as the ConvNet and GloVe embeddings [34] as pre-trained word embeddings. GloVe was trained on Common Crawl dataset with a vocabulary of 1.9M words[2]. Despite their appealing modelling power, bilinear models are intractable for our task, because the size of the full tensor is prohibitive. In our experiments, we use the MUTAN model [5] to approximate the bilinear product (Equation 6) but other models [19, 25, 49, 14] can be used.

Datasets. We perform experiments on a subset of the YFCC100M dataset [41]. YFCC100M consists of about 99 million images from the Flickr image sharing site. We collect the images from all the users having between 100 and 200 images with at least one hashtag. We use all the hashtags for which we can compute a GloVe representation. The training set has a vocabulary of 442k hashtags and the test set has a vocabulary of 568k hashtags (about 125k hashtags are unseen during training). We ignore all the images that do not have at least one valid hashtag. Finally, we keep all the users that have at least 50 images. We split the sets by user ID in order to ensure that images from the same user do not occur in both sets. We assign 70% (resp. 10% and 20%) of the images to the training (resp. validation and test) set. Thereafter, this dataset is named open vocabulary

[2] https://nlp.stanford.edu/projects/glove/

	Model	User rep.	User fusion	A@1	A@10	P@10	R@1	R@10
Fixed vocab (~ 18.5k hashtags)	[A] frequency		-	0.01	0.13	0.03	0.00	0.07
	[B] user agnostic		-	14.57	37.60	7.52	4.79	15.86
	[C] used hashtags	✓	max	61.62	80.43	37.37	26.02	55.88
	[D] hashtag occurences	✓	sum	62.09	80.56	37.58	26.26	56.13
	[43] Tensor (MCLL)	✓	-	14.75	37.66	7.53	4.86	15.94
	Ours (hashtag)	✓	GRU	71.90	85.21	47.60	31.51	62.83
	Ours (image+hashtag)	✓	GRU	**74.13**	**87.49**	**50.88**	**33.36**	**66.49**
Open vocab (~ 440k hashtags)	[A] frequency		-	0.00	0.01	0.01	0.00	0.00
	[B] user agnostic		-	13.47	34.71	6.64	4.26	13.49
	[E] hashtag sum	✓	sum	59.93	79.75	36.24	23.42	54.20
	[43] Tensor (MCLL)	✓	-	13.49	34.73	6.65	4.26	13.50
	Ours (hashtag)	✓	GRU	65.06	83.31	44.84	26.87	60.69
	Ours w/o Glove (image+hashtag)	✓	GRU	46.24	64.17	20.36	17.08	31.49
	Ours (image+hashtag)	✓	GRU	**67.46**	**86.32**	**46.68**	**27.90**	**62.99**

Table 1. Hashtag prediction results on both datasets (higher is better). We compare several strategies to extract a user representation based on user image history. The performance on the open vocabulary dataset is evaluated only with the hashtags seen during training. The performance with the unseen hashtags is shown in Table 2. Ours w/o Glove means that the pretrained GloVe embeddings are not used.

dataset. We also proposed a fixed vocabulary version of the open vocabulary dataset. We use a similar hashtag preprocessing as [43] except the dataset splits are by user ID. On this dataset, we propose a variant of our model without pretrained word embeddings to have a fair comparison with [43] (see subsection A.3 of supplementary). More information and analysis about these datasets can be found in subsection A.1 of supplementary.

Metrics. To evaluate the hashtag prediction performance of the models, we use three standard metrics [12, 43]: Accuracy@k (A@k), Precision@k (P@k) and Recall@k (R@k). More information about these metrics is available in subsection A.2 of supplementary. We use $k = 1$ and $k = 10$: for instance, A@1 measures how often the top-ranked hashtag is in the ground-truth hashtag set and A@10 how often at least one of the the ground-truth hashtags appears in the 10 highest-ranked predictions.

4.1. Hashtag prediction

In this section, we evaluate our model for the hashtag prediction task which attempts to rank an image's ground-truth hashtags higher than hashtags it does not contain. In these experiments, we use all the previous images in the user history to extract the user representation. Image retrieval results are shown in subsection A.5 of supplementary.

Baseline models. We compare our model with the following models:
[A] Frequency: this simple baseline ignores input image and user representation, always ranking hashtags by their frequency in the training data.
[B] User agnostic: this model is equivalent to a standard image classification: there is no user representation.
[C] Used hashtags: this user representation is a binary vector of the hashtags used in previous images by the user:

$$\mathbf{u} = [u_1, \ldots, u_K] \quad \text{where} \quad u_i \in \{0, 1\} \tag{10}$$

where $u_i = 1$ (resp. $u_i = 0$) means that the i-th hashtag has been used (resp. has never been used) by the user.
[D] Hashtag occurrences: this user representation is similar to [C] except that it indicates the number of occurences of each hashtag:

$$\mathbf{u} = [u_1, \ldots, u_K] \quad \text{where} \quad u_i \in \mathbb{N} \tag{11}$$

where u_i indicates the number of times that the i-th tag has been used by the user.
[E] Hashtag sum: this user representation is the sum of each hashtag word embedding used in previous images by the user.

The models [C] and [D] are not used on the open vocabulary dataset because they require a fixed hashtag vocabulary. Note that it is not possible to compare with the user representation proposed in [12] because it uses user metadata that are not available in the dataset. We also report the results of our model with only the hashtag branch in the user representation model (*i.e.* $\mathbf{x} = \mathbf{x}^{tag}$) and without the pretrained GloVe embeddings (they are randomly initialised). To compare our model with [43], we reimplement the user-specific Tensor (MCLL) model. Because this model was proposed for a transductive setting, we use a vector filled with the value $1/d_u$ to represent a user that is not present in the training set.

MODEL	UNSEEN HASHTAGS (∼ 120k hashtags)					ALL HASHTAGS (∼ 560k hashtags)				
	A@1	A@10	P@10	R@1	R@10	A@1	A@10	P@10	R@1	R@10
[B] user agnostic	0.06	0.40	0.08	0.03	0.25	12.89	33.21	6.07	3.78	12.05
[E] sum hashtags	36.41	55.40	32.51	26.60	48.12	58.91	79.47	34.08	21.35	51.42
Ours (hashtag)	44.07	60.15	39.35	33.97	53.05	65.75	83.90	43.99	26.09	59.14
Ours (image+hashtag)	**45.98**	**62.62**	**41.31**	**35.53**	**55.30**	**68.06**	**86.91**	**45.80**	**27.03**	**61.39**

Table 2. Hashtag prediction results on hashtags unseen during training and all the hashtags on the open vocabulary dataset.

Results. The performance of all the models are summarized in Table 1 and we make seven observations. First, the user agnostic models ([A, B]) perform poorly for all metrics with respect to the user-specific models as already shown in [12, 43]. It also demonstrates that the user history can be used to extract good user representations. Second, the Tensor (MCLL) model [43] has similar performance than the user agnostic model because it cannot deal with user unseen during training. It is necessary to retrain [43] to deal with new user. We also compare our model with [43] in a transductive setting in a next paragraph. Third, we observe that the hashtag occurrences user representation [D] is slightly better than the used hashtags user representation [C]. The reason is that the [D] is richer than [C] because it encodes user hashtag frequency. Fourth, modelling the temporal information of the hashtags with a recurrent network (our model with only hashtags) significantly improves performance with respect to hashtag pooling strategy ([C, D]). Fifth, using visual information improves the results because it can exploit the correlations between the hashtags and the visual content of the images. Sixth, we observe that the pretrained word embeddings are very important on imbalanced data because it allows to transfer knowledge between hashtags. Finally, we observe the same behaviour on the closed set and open set datasets, so our user representation model can be used in both settings. Visual results are shown in subsection A.4 of supplementary.

Results on unseen hashtags. We also evaluate the ability of our model to generalize to predict unseen hashtags. In the first experiment, named UNSEEN HASHTAGS, we only evaluate the results of unseen hashtags (equivalent to ZSL setting). In the second experiment, named ALL HASHTAGS, we evaluate the performance for all the hashtags (similar to GZSL setting). While the first experiment directly evaluates the performance on unseen hashtags, the second experiment is more realistic because the model has to predict hashtags among both seen and unseen hashtags. The results of these experiments are shown in Table 2 on the open vocabulary dataset. We observe that our model is able to predict unseen hashtags so our model is able to deal with new hashtags without retraining the model. We draw the same conclusions about the user representation that for seen hashtags in Table 1: modeling the user is important for unseen tags, and our user representation model has the best results because it models the temporal information and exploits the visual content.

MODEL	A@1	A@10	P@10	R@1	R@10
[43]	35.92	63.07	11.51	15.91	37.79
Ours-FH	48.20	69.59	33.03	20.50	46.41
Ours	**73.19**	**87.28**	**50.44**	**32.19**	**65.86**

Table 3. Comparison with [43] on a fixed set of users. Ours-FH means that our user representation is computed with on a fixed history (training images).

Comparison with state-of-the-art model [43] in a transductive setting. We compare our model with [43] on the fixed vocabulary dataset with a transductive setting *i.e.* the same set of users during both training and testing. The results are summarized in Table 3. We report the results of our model with the user representation computed only on the training images (fixed user history) and our model with the user representation computed on all the previous images. We observe that our model is better than [43] because [43] needs a lot of images to have good performance, and cannot exploit the temporal information because each image is processed independently during training. Another advantage of our dynamic approach is that it can improve the user representation by exploiting new images without retraining the model.

4.2. Model analysis

In this section, we analyse important parameters of our model: the dimension of the user representation and importance of the image-hashtags fusion. The impact of the image history size to extract the user representation is analyzed in the supplementary (subsection A.6).

User representation dimension. We first analyze the importance of the user representation dimension, which is the hidden state dimension of the GRU in our model. We show in Figure 3 the R@10 and the computation time for a large range of user dimension (32 to 8192). We observe that using a large user representation is better than small user representation for all metrics. However, using a large user represen-

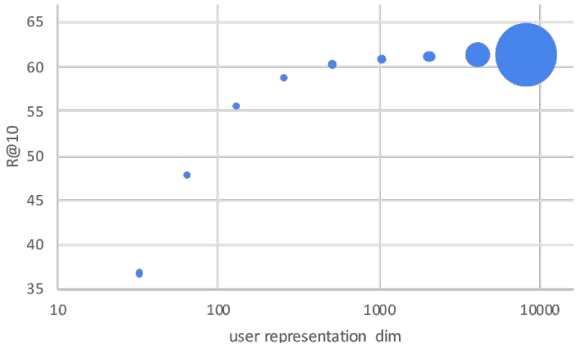

Figure 3. Analysis of the user representation dimension with respect to R@10. The width of the circle is proportional to the computation time of the user representation. (x-axis log scale)

FUSION	A@1	A@10	R@1	R@10
only hashtags	65.16	83.26	26.12	60.89
sum	65.29	83.21	26.19	60.75
concatenation	65.36	83.24	26.21	60.71
bilinear [5]	65.95	85.63	26.69	59.94
GLU [11]	66.02	85.77	26.73	60.28
TIRG [44]	63.97	81.94	25.10	59.35
eltwise product	**67.28**	**86.27**	**27.18**	**62.88**

Table 4. Analysis of the image-hashtags fusion operator.

tation is more time consuming and requires more memory to store (the GRU becomes the bottleneck of the model). We observe that 1024 is a good trade-off between accuracy and computation time and we use this dimension for ours experiments.

Image-hashtags fusion. Our second analysis is about the combination of the image and hashtags branches in the user representation model (Equation 4). In Table 4, we show the results for several standard multi-modal fusion operators, and our model with only the hashtags branch. We compare standard fusion operators (element-wise sum, concatenation, element-wise product) and more complex operators like bilinear (MUTAN) [5], GLU [11], and TIRG [44]. We use ReLU for each model except for the element-wise product model where we use SELU to avoid having a vector with too many zeros (using a ReLU with the element-wise product significantly drops the performances). We note that only the element-wise product fusion improves significantly the performance. We believe this is because the element-wise product fusion forces the model to exploit both image and hashtags representations. We observe that both bilinear and GLU operators improve the performance but are not able to efficiently exploit the visual representation. This experiment also shows that the hashtags branch is more informative than the image branch. We note that our conclusions are different from [44] which shows the best fusion depends on the task. Finally, we want to point out that it is probably possible to use/design better fusion strategies between the user embedding and image embedding but it is out of the scope of this work.

	USER REP.	A@1	A@10	MR	DIM
FIXED	[C] used	33.48	46.95	16	18,583
	[D] occurence	33.64	46.94	17	18,583
	Ours (tag)	42.95	58.47	3	1024
	Ours (im+tag)	**45.64**	**61.45**	**2**	1024
OPEN	[E] sum tags	35.19	44.81	29	300
	Ours (tag)	45.15	59.27	3	1024
	Ours (im+tag)	**47.90**	**61.56**	**2**	1024

Table 5. User retrieval results. MR is the median rank (lower is better) and dim is the user representation dimensionality.

4.3. User retrieval

In this section, we analyze the discriminative power of our user representation model. To achieve it, we consider the user retrieval task: given a user representation, the goal is to find a user representation of the same user computed with non-overlapping image histories *i.e.* each image is used only in one image history. We use users from the test set and an image history size of 20. For instance, given a user, we first use the first 20 images to compute a user representation, then we use the next 20 images to compute another user representation of the same user. For this experiment, we compute 33,648 user representations from 6,139 users. The user representations are ℓ_2 normalized and we use the cosine similarity to rank the users. To evaluate the performance, we use the Accuracy@k metric and the median rank metric. The results in Table 5 show that our model is able to extract accurate user representations from different image history sizes. Note that our user representation model is not trained for this task. About the user representation model, we observe the same conclusions as for hashtag prediction. Despite our user representation being 18 times smaller than [C] and [D] (which are sparse vectors), we note that our model improves the A@1 performance by 12 pt. On the contrary [E] has a smaller dimension than our model, but the representations are not discriminative enough.

5. Conclusion and Future Work

We introduced a new model for hashtag prediction that can deal with new users and new hashtags without retraining the model. This paper shows that the images and their corresponding hashtags in the user history can be efficiently used to extract a user representation. Our user representation can be successfully used for user specific hashtag prediction and user retrieval. Our user representation model could be extended to exploit user relationships or user metadata.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-Embedding for Image Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016. 2

[2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[3] Jimmy Ba, Kevin Swersky, Sanja Fidler, and Ruslan Salakhutdinov. Predicting Deep Zero-Shot Convolutional Neural Networks using Textual Descriptions. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[4] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. In *arXiv 1803.01271*, 2018. 4

[5] Hedi Ben-younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 5, 8

[6] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. In *Transactions of the Association for Computational Linguistics*, 2017. 4

[7] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Improving Semantic Embedding Consistency by Metric Learning for Zero-Shot Classification. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[8] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. Cross-Modal Retrieval in the Cooking Context: Learning Semantic Text-Image Embeddings. In *ACM Special Interest Group on Information Retrieval (SIGIR)*, 2018. 3

[9] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An Empirical Study and Analysis of Generalized Zero-Shot Learning for Object Recognition in the Wild. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[10] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. In *Advances in Neural Information Processing Systems Workshop (NeurIPS)*, 2014. 3, 4

[11] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International Conference on Machine Learning (ICML)*, 2017. 8

[12] Emily Denton, Jason Weston, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. User conditional hashtag prediction for images. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2015. 2, 5, 6, 7

[13] Bhuwan Dhingra, Zhong Zhou, Dylan Fitzpatrick, Michael Muehl, and William W. Cohen. Tweet2Vec: Character-Based Distributed Representations for Social Media. In *Association for Computational Linguistics (ACL)*, 2016. 2

[14] Brendan Duke and Graham W. Taylor. Generalized Hadamard-Product Fusion Operators for Visual Question Answering. In *arXiv 1803.09374*, 2018. 2, 5

[15] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Finding beans in burgers: Deep semantic-visual embedding with localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[16] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improved Visual-Semantic Embeddings. In *British Machine Vision Conference (BMVC)*, 2018. 3

[17] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. DeViSE: A Deep Visual-Semantic Embedding Model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013. 2

[18] Jianlong Fu, Yue Wu, Tao Mei, Jinqiao Wang, Hanqing Lu, and Yong Rui. Relaxing From Vocabulary: Robust Weakly-Supervised Deep Learning for Vocabulary-Free Image Tagging. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[19] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 2, 5

[20] Albert Gordo, Jon Almazan, Naila Murray, and Florent Perronnin. LEWIS: Latent Embeddings for Word Images and their Semantics. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 3

[21] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. End-to-end Learning of Deep Visual Representations for Image Retrieval. In *International Journal of Computer Vision (IJCV)*, 2017. 3

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 5

[23] Hamid Izadinia, Bryan C. Russell, Ali Farhadi, Matthew D. Hoffman, and Aaron Hertzmann. Deep Classifiers from Image Tags in the Wild. In *ACM Multimedia*, 2015. 2

[24] Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning Visual Features from Large Weakly Supervised Data. In *European Conference on Computer Vision (ECCV)*, 2016. 5

[25] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard Product for Low-rank Bilinear Pooling. In *International Conference on Learning Representations (ICLR)*, 2017. 2, 5

[26] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 5

[27] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. In *arXiv 1411.2539*, 2014. 3

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. ImageNet Classification with Deep Convolutional Neural Net-

works. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012. 2

[29] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2

[30] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2014. 2

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2014. 1

[32] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the Limits of Weakly Supervised Pretraining. In *European Conference on Computer Vision (ECCV)*, 2018. 5

[33] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013. 4

[34] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 4, 5

[35] Marcus Rohrbach, Michael Stark, and Bernt Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 2

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 2015. 1

[37] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning Cross-modal Embeddings for Cooking Recipes and Food Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[38] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. In *NIPS 2017 workshop: Machine Learning for the Developing World*, 2017. 2

[39] Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. Engaging Image Captioning Via Personality. In *arXiv 1810.10665*, 2018. 2, 3

[40] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 5

[41] Bart Thomee, Benjamin Elizalde, David A. Shamma, Karl Ni, Gerald Friedland, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The New Data in Multimedia Research. *Communications of the ACM*, 2016. 5

[42] Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Conditional Similarity Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[43] Andreas Veit, Maximilian Nickel, Serge Belongie, and Laurens van der Maaten. Separating Self-Expression and Visual Content in Hashtag Supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 5, 6, 7

[44] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing Text and Image for Image Retrieval - An Empirical Odyssey. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 8

[45] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning Two-Branch Neural Networks for Image-Text Matching Tasks. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018. 2

[46] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning Deep Structure-Preserving Image-Text Embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[47] Jason Weston, Sumit Chopra, and Keith Adams. # tagspace: Semantic embeddings from hashtags. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 2, 4

[48] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018. 2

[49] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 5

[50] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 2