Image2Emoji: Zero-shot Emoji Prediction for Visual Media

Spencer Cappallo[†] Thomas Mensink[†] Cees G. M. Snoek[†‡]

[†]University of Amsterdam [‡]Qualcomm Research Netherlands

{cappallo, tmensink, cgmsnoek}@uva.nl

ABSTRACT

We present *Image2Emoji*, a multi-modal approach for generating emoji labels for an image in a zero-shot manner. Different from existing zero-shot image-to-text approaches, we exploit both image and textual media to learn a semantic embedding for the new task of emoji prediction. We propose that the widespread adoption of emoji suggests a semantic universality which is well-suited for interaction with visual media. We quantify the efficacy of our proposed model on the MSCOCO dataset, and demonstrate the value of visual, textual and multi-modal prediction of emoji. We conclude the paper with three examples of the application potential of emoji in the context of multimedia retrieval.

1. INTRODUCTION

Visual classification and retrieval models traditionally rely on a limited set of pre-defined concepts, which are frequently trained on millions of photos and require a costly re-training process to adapt to new concepts, *e.g.* [1, 6]. Furthermore, at the end of this process, the model only outputs a list of concepts and their likelihood, which must be translated in some way to a more digestible form for an end user. To address these limitations, we suggest the use of ideograms as a final representation for concepts, and propose an approach to predict any arbitrary set of ideograms without re-training the visual classifier.

Ideograms maintain a visual grammar of interaction, limiting the semantic gap between a query and the returned media, and allow for language-independent, youth-friendly user interfaces which adapt seamlessly to a touchscreen saturated world. This work uses emoji as a candidate set of ideograms for visual search. Emoji are a set of over 700 ideograms, which is widely prevalent and supported natively by most smartphones, as well as many major websites such as Facebook and Twitter. Their widespread use suggests that, as a set of ideograms, emoji have sufficiently broad semantic coverage as to be interesting. See Figure 1 for examples of images with emoji predicted by our model.

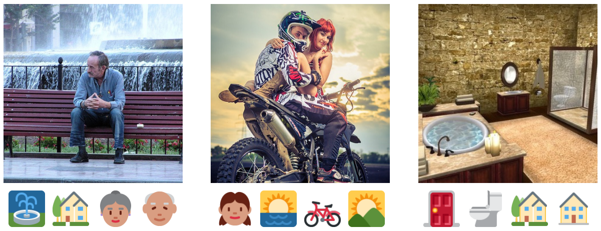


Figure 1: Emoji predictions for three input images.

To avoid costly data annotation and training steps, and to be applicable to any arbitrary set of ideograms, our proposed approach for emoji labeling is zero-shot, relying on a previously-learned semantic embedding space to translate visual concept detections and user text into the target labels. We see several opportunities for the use of ideograms and emoji within the multimedia community. As a pre-defined set of queries, they hold advantages over text lists (see Figure 2) for retrieval or exploration tasks. As clear iconography they present opportunities as a means for interaction on small screens such as smart watches. They also suggest possibilities for description and summarization tasks, for example when describing image collections and videos.

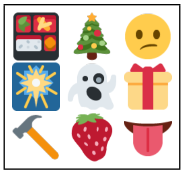
The task we have chosen resembles that of zero-shot concept detection. In particular, our model's treatment of the visual modality is similar to that of [5]. In [5], a semantic embedding predicts unseen ImageNet classes from a classifier's probability outputs for known classes by finding the distance between the high-probability known concepts and the unseen concepts in a vector space. In both [5] and our work, the semantic embedding is generated using word2vec [4]. We differentiate our work from [5] through the introduction of a second, textual modality to the model; the motivated use of a significantly different training corpus for learning our semantic embedding; and an overall focus on the prediction of emoji labels, which is a multi-label setting in contrast to their single-label scenario. We call our approach *Image2Emoji* and detail its technicalities next.

2. IMAGE2EMOJI

Image2Emoji combines visual concepts and user text to predict unseen emoji labels in a zero-shot manner.

The problem can be formalized thus: The objective is to predict a target set of ideogram labels, $\mathcal{Z}$, by relying on a set of input concepts, $\mathcal{Y}$, where we assume $\mathcal{Z} \cap \mathcal{Y} = \emptyset$. Our input concepts are comprised of visual concepts from a pre-trained classifier, $\mathcal{Y}_v$, and textual concepts extracted

MM'15, October 26–30, 2015, Brisbane, Australia.
© 2015 ACM. ISBN 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806335.



bento	christmas tree	confused
fireworks	ghost	gift
hammer	strawberry	tongue

Figure 2: Comparison of emoji concepts presented as ideograms (L) and as text (R).

from user text, $\mathcal{Y}_t$, s.t. $\mathcal{Y}_v, \mathcal{Y}_t \in \mathcal{Y}$. To accomplish this without training, we rely on an intermediary vector space where semantic relationships between known labels $\mathcal{Y}$ and target labels $\mathcal{Z}$ can be exploited to score the emoji concepts. See Figure 3.

2.1 Semantic Embedding

Image2Emoji uses a word2vec vector representation [4] as a semantic embedding space. The embedding space is designed to minimize the distance between the vectors of semantically similar words. The word2vec model consists of a neural network with a single hidden layer. When given as input a token t_i, the network tries to predict n tokens from surrounding context on each side, $\{t_{i-n}, ..., t_{i-1}, t_{i+1}, ..., t_{i+n}\}$, which is known as a skip-gram model. Once trained, the d node-values of the hidden layer are used as a projection into the learned d-dimensional semantic vector space. These vector representations are normalized when calculating semantic similarity.

Our semantic embedding is a 500-dimensional word2vec model, which is trained on the title, description, and tag text from the 100M Flickr photos in the YFCC100M dataset [7]. An 11% increase in mean average precision for our task was observed by using the text from Flickr, compared against training the model on the text of Wikipedia, as used in [5]. We expect this increase is due to the language on Flickr being more closely tied to visual discussion than the language used on Wikipedia.

The semantic embedding space is denoted as $\mathcal{S}$, and a function w is defined which returns the vector representation of a given token within the space $\mathcal{S}$, namely $w : \mathcal{Y}, \mathcal{Z} \to \mathcal{S}$. By placing an input label $y \in \mathcal{Y}$ and target label $z \in \mathcal{Z}$ in a single vector space, their semantic closeness can be found by calculating the Cosine Similarity between their normalized vector representations. The combined operation of embedding the labels in the semantic space and finding their similarity is denoted by $cos(z, y) = w(z) \cdot w(y)$.

2.2 Emoji Scoring

For a given input image x_v, the visual classifier produces probabilities $p(y_v|x_v)$ across the set of visual concepts $y_v \in \mathcal{Y}_v$. For a given target label, $z \in \mathcal{Z}$, the influence of a given visual concept, y_v, is found through the product of their similarity in the semantic embedding space and the probability output of the classifier. The contribution of the visual modality to the target label prediction is the sum of these products for every $y_v \in \mathcal{Y}_v^*$, where $\mathcal{Y}_v^* \subset \mathcal{Y}_v$ consisting of the N_v labels with the highest probabilities for the given input image x_v. The scoring function for emoji labels based on the visual modality is labeled as S_v:

$$S_v(z, x_v) = \sum_{y_v \in \mathcal{Y}_v^*} cos(z, y_v) \cdot p(y_v|x_v) \tag{1}$$

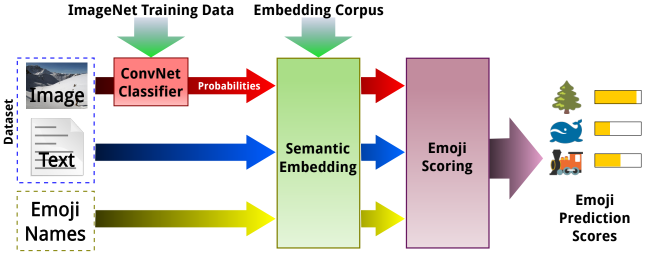


Figure 3: Data flow within Image2Emoji. Probability scores of visual classes along with the image's accompanying text are mapped to the semantic vector space, and their similarity to the emoji label names are used to score the emoji. Both modules are detailed in Section 2.

The predictions based on the text modality follow a similar pattern. However, due to the discrete nature of the text input and the fact that the probabilities $p(y_t|x_t)$ for a given text input x_t and text label $y_t \in \mathcal{Y}_t$ are restricted to being either 1 or 0 (present in the text or not), the maximum of the product is used instead of the sum. The scoring function based on the text modality is given as S_t:

$$S_t(z, x_t) = \max_{y_t \in \mathcal{Y}_t} cos(z, y_t) \cdot p(y_t|x_t) \tag{2}$$

To combine the predictions from the modalities, we employ a late-fusion strategy with a weighting factor α that is restricted to the range $[0, 1]$. The full scoring function is found through combining 1 and 2:

$$S(z, x_v, x_t) = \alpha S_v(z, x_v) + (1 - \alpha) S_t(z, x_t) \tag{3}$$

for a target label $z \in \mathcal{Z}$, an input image x_v, and accompanying user text x_t.

The value of α can be discovered through validation on known annotations, or simply left as 0.5 for an equal weighting of input modalities. Though our approach focuses on visual and text modalities, it is trivially adaptable to any number of input modalities, provided they can be mapped to the semantic embedding space.

Since the proposed method forgoes a costly training stage, it is computationally lightweight. Calculating the cosine similarity matrix needs only be done once per set of target labels and input modality, and has a complexity of $\mathcal{O}(N_{\mathcal{Y}} N_{\mathcal{Z}} D)$, where $N_{\mathcal{Y}}$ is the number of input labels, $N_{\mathcal{Z}}$ the number of target labels, and D is the dimensionality of the semantic embedding. Predicting an emoji representation for a given input has complexity $\mathcal{O}(N_T N_{\mathcal{Z}})$ per input image and modality, where N_T is the number of input labels selected ($N_T = N_v$ in the visual modality, and in the textual modality $N_T = \#$ of words.)

3. EXPERIMENTS

Dataset Although there is no existing dataset with emoji annotations, we wish to quantitatively evaluate the efficacy of our zero-shot approach before demonstrating its performance on emoji. To do this, we test our model on the train set of MSCOCO [3]. MSCOCO is selected for this task because 37 out of its 80 label categories correspond to concepts present within the set of emoji. The MSCOCO train data consists of 83k images with multi-label annotations in 80 classes and Mechanical Turk provided captions.

	im-mAP	mAP
Random lower bound	0.086	0.037
Image2Emoji-visual	0.460	0.355
Supervised upper bound	0.750	0.562

Table 1: Image2Emoji performance using only the visual modality. For context, upper and lower bounds are also reported in the form of a supervised SVM approach and random rankings. Our zero-shot approach is nearer the performance of the supervised upper bound than the lower bound.

Implementation details Our visual classifier is a deep convolutional network in the style of GoogLeNet [6], trained on ImageNet data [2] to classify 15,293 concepts – the number of concepts with at least 200 positive images in the dataset. To ensure that the evaluation remains zero-shot, 110 of these visual concepts are excluded due to their class names overlapping with any of the 80 MSCOCO classes, resulting in a total of 15,183 visual concepts.

Results are reported for two different sources of text. One is text source is the first listed caption per image from the high-quality Mechanical Turk annotations. Out of concern that these descriptions are not representative of user text under normal circumstances, we also collect the title, description, and tag information from Flickr. Results are also reported for this second, more realistic text source.

Evaluation criteria To demonstrate the relative performance of the text and visual modalities within our model, the two are evaluated individually without contribution from the other. Following this, we report the change in performance when both modalities are combined, with varying values of the fusion weighting parameter α. Two evaluation metrics are reported for each test. The first, corresponding to a retrieval task, is the mean average precision when ranking images for target labels, which we denote with 'mAP'. Also reported is the mean average precision *per image*, which corresponds to the success of ranking emoji labels for describing or summarizing a given input, and which we refer to as '*image*-mAP' or '*im*-mAP'.

4. RESULTS

Visual Prediction We first evaluate the performance of the model when relying solely on visual input. This corresponds to removing the blue arrows from the diagram in Figure 3. To provide context to our results, we provide upper and lower bound results.

To establish the upper bound, a supervised method is tested. A linear SVM is trained via a one-vs-rest paradigm to predict the 80 MSCOCO labels. The 15k concept probabilities are used as the feature representation, and the model is trained on the 40k images of MSCOCO's validation set. The performance of the supervised approach along with the performance of random predictions can be seen in Table 1. Encouragingly, our zero-shot approach is closer to the performance of the supervised upper bound than the random lower bound. These values also motivate the inclusion of a text modality in a zero-shot approach, in contrast to using only visual concepts as in [5].

	Captions		Flickr Text	
	im-mAP	mAP	*im*-mAP	mAP
Baseline	0.337	0.288	0.380	0.342
Image2Emoji-text	0.647	0.536	0.604	0.555

Table 2: Results of Image2Emoji using only the text modality. Two text sources are tested: Mechanical Turk-provided descriptions, and user text harvested from Flickr. Results are compared to a simple baseline that omits a semantic embedding in favor of a direct mapping. Our model outperforms the baseline, performing well even with the Flickr text.

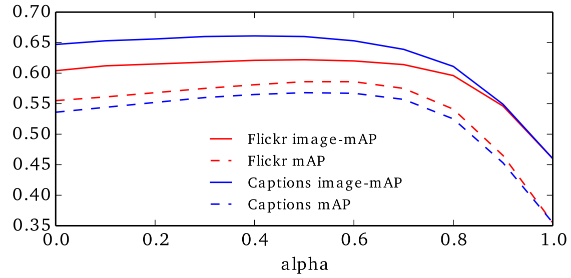


Figure 4: Effect of the modality-weighting parameter α. Peak performance occurs near $\alpha = 0.5$, where Image2Emoji has mAPs of 0.586 and 0.568, and *im*-mAPs of 0.622 and 0.660 for the Flickr text and the captions, respectively. In the mAP setting, our zero-shot, multi-modal model actually outperforms the visual-only supervised upper bound.

Textual Prediction Image2Emoji using solely the text modality is evaluated. This corresponds to removing the red arrows of the visual modality from the diagram in Figure 3.

We compare the text portion of our model against a simple zero-shot baseline. This baseline matches directly the terms in the input text with the target labels. In Table 2, results are presented for both the high-quality captions and the noisier Flickr text. Our model, utilizing an intermediary semantic embedding, outperforms the simple, direct baseline. This reflects the more complete relationships between words which are captured by the embedding. It is worth noting that our method yields roughly similar results for both the captions and the Flickr text. Despite the noise inherent to the Flickr text, it is usually longer than the single-sentence captions provided by Mechanical Turk users, which provides a greater number of semantic data points.

Fusion Prediction Lastly, we combine the predictions. This corresponds to the full diagram in Figure 3, containing both the text and visual modalities. The performance of Image2Emoji for varying values of α, and using either the caption text or Flickr text, are in Figure 4. Notably, when using the Flickr text, the mAP performance of the model for predicting labels actually exceeds that of the supervised visual baseline (0.586 to 0.562). Furthermore, the performance discrepancy with varying values of α is very small around the optimal selection. This suggests that selecting a suboptimal α should have only marginal effects on the overall performance.

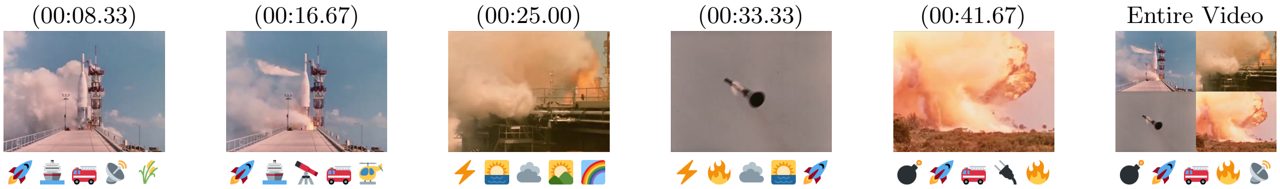


Figure 5: The five highest scoring emoji for frames at regular intervals across a video. Emoji predictions from frame-level visual detections were combined with the video's user-provided title. The emoji representation for the entire video is also shown.

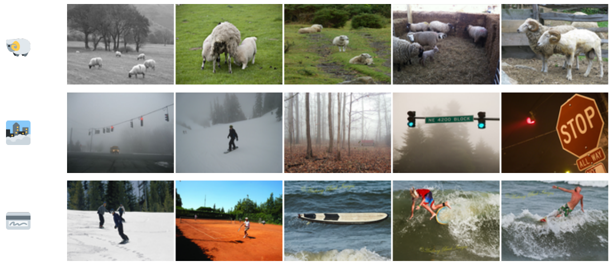


Figure 6: Top images for queries of the *sheep*, *foggy*, and *credit card* emoji. *Credit card* performs poorly, perhaps due to a lack of credit cards in the dataset.

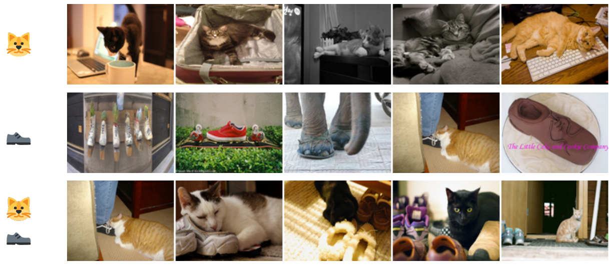


Figure 7: Example of composition with emoji concepts to create more nuanced queries. *Shoe* and *cat* are combined to retrieve images containing both.

5. APPLICATION POTENTIAL

Following the numerical evaluation of the proposed model, we showcase some possible applications of Image2Emoji. All examples presented in this section use a subset comprised of 385 emoji. These have been manually selected by removing emoji which were not pictographs and therefore not clearly tied to a visual concept (such as 🔡 *abcd* or ↔️ *left-right arrow*) and also removing those characters which repeat a concept already represented by another, very similar, character (e.g., 🐷 *pig face* and 🐖 *pig*).

Query-by-emoji The focus of the proposed model is to use emoji as a means for retrieval and exploration of visual data. Three examples of the top-ranked images for a given emoji query are shown in Figure 6. It is worth noting that 🌁 *foggy* returns very sensible results, despite the concept of 'foggy' being very distant from the collection of concrete nouns used as concepts for the visual modality. The text modality is likely crucial in capturing this meaning. The query 💳 *credit card* performs very poorly, which we suspect is due to a lack of credit cards in the MSCOCO dataset we tested on. Also shown is an example of query composition in Figure 7, where the top images for 🐱 *cat*, 👞 *shoe*, and 🐱 *cat* + 👞 *shoe* are listed. The combination of emoji can allow for more specific or nuanced queries to be constructed from a limited concept set.

Emoji ranking In Figure 1, we show results of our model for the task of ranking descriptive emoji for a given visual input. Worth noting is the presence of multiple concepts that share similar semantics in the embedding space, such as 👴 *old man* and 👵 *old woman*. The training process for the semantic embedding will assign similar vector directions to tokens which are used in a similar manner, resulting in similar scores.

Emoji summarization Representational or descriptive emoji are most useful for describing collections of images or video. For this reason, Image2Emoji applied to a video is shown in Figure 5. The top-scoring emoji for five frames evenly distributed throughout the video are displayed, along with an emoji representation for the entire video using the average across all frames. The video-level title text from YouTube was used for the text modality, with a large α used for the frame-level predictions to emphasize the more local, visual information. In this example, the emoji summary gives a compelling summary of the video contents.

In this paper, we have proposed and evaluated a multimodal, zero-shot approach to generating ideogram labels for visual media, and have investigated several uses of emoji for exploration and representation purposes.

Acknowledgments This research is supported by the STW STORY project and the Dutch national program COMMIT.

6. REFERENCES

[1] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *MM*, 2013.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[3] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[5] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[7] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *arXiv:1503.01817*, 2015.