EmojiGAN: learning emojis distributions with a generative model

Bogdan Mazoure*
Department of Mathematics & Statistics
McGill University

Thang Doan*
Desautels Faculty of Management
McGill University

Saibal Ray
Desautels Faculty of Management
McGill University

Abstract

Generative models have recently experienced a surge in popularity due to the development of more efficient training algorithms and increasing computational power. Models such as adversarial generative networks (GANs) have been successfully used in various areas such as computer vision, medical imaging, style transfer and natural language generation. Adversarial nets were recently shown to yield results in the image-to-text task, where given a set of images, one has to provide their corresponding text description. In this paper, we take a similar approach and propose a image-to-emoji architecture, which is trained on data from social networks and can be used to score a given picture using ideograms. We show empirical results of our algorithm on data obtained from the most influential Instagram accounts.

1 Introduction

The spike in the amount of user-generated visual and textual data shared on social platforms such as Facebook, Twitter, Instagram, Pinterest and many others luckily coincides with the development of efficient deep learning algorithms (Perozzi et al., 2014; Pennacchiotti and Popescu, 2011; Goyal et al., 2010). As humans, we can not only share our ideas and thoughts through any imaginable media, but also use social networks to analyze and understand complex interpersonal relations. Researchers have access to a rich set of metadata (Krizhevsky, 2012; Liu et al., 2015) on which various computer vision (CV) and natural language processing (NLP) algorithms can be trained.

For instance, recent work in the area of image captioning aims to provide a short description (i.e. caption) of a much larger document or image (Dai et al., 2017; You et al., 2016; Pu et al., 2016). Such methods excel at conveying the dominant idea of the input. On the other hand, we use ideograms, also popular under the names of emojis or pictographs as a natural amalgam between annotation and summarization tasks. Note that, in this work, we use the terms emoji, ideogram and pictograph interchangeably to represent the intersection of these three domains. Ideograms bridge together the textual and visual spaces by representing groups of words with a concise illustration. They can be seen as surrogate functions which convey, up to a degree of accuracy, reactions of social media users. Furthermore, because each emoji has a corresponding text description, there is a direct mapping from ideograms onto the word space.

In this paper, we model the distribution of emojis conditioned on an image with a deep generative model. We use generative adversarial networks (GANs) (Goodfellow et al., 2014), which are notoriously known to be harder to train than other distributional models such as variational auto-encoders (VAEs) (Kingma and Welling, 2013) but tend to produce sharper results on computer vision tasks.

2 Related Work and Motivation

Since the release of word2vec by Mikolov and colleagues in 2013 (Mikolov et al., 2013), vector representations of language entities have become more popular than traditional encodings such as bag-of-words (BOW) or n-grams (NG). Because word2vec operations preserve the original semantic meaning of words, concepts like word similarity and synonyms are well-defined in the new space and correspond to closest neighbors of a point according to some metric.

The aforementionned word representation was followed by doc2vec (Le and Mikolov, 2014). Orig-

*These authors contributed equally.

Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 273–279

https://doi.org/10.18653/v1/P17

inally, doc2vec was meant to efficiently encode collections of words as a whole. However, since empirical results suggest a similar performance for both algorithms, researchers tend to opt for the simpler and more interpretable word2vec model.

One of the most recent and the most interesting vector embeddings has been emoji2vec (Eisner et al., 2016). It consists of more than 1,600 symbol-vector pairs, each associating a Unicode character to a real $300-$dimensional vector. The abundance of pictographs such as emojis on social communication platforms suggests that word-only analyses are limited in their scope to capture the full scale of interactions between individuals. Emojis' biggest advantage is their universality: no information is lost due to faulty translations, mistyped characters or even slang words. In fact, emojis were designed to be more concise and expressive than words. They, however, have been shown to suffer from varying interpretations which depend of factors such as viewing the pictograph on an iPhone or a Google Pixel (Miller et al., 2016). This in turn implies that the subject of conversation highly impacts the choice of media (text or emoji) picked by the user (Kelly and Watts, 2015). Reducing a whole media such as a public post or an advertisement image to a single emoji would almost certainly mean loosing the richness of information, which is why we suggest to instead model visual media as a conditional distribution over emojis that users employ to score the image.

Deep neural models have previously been used to analyse pictographic data: (Cappallo et al., 2015) used them to assign the most likely emoji to a picture, (Felbo et al., 2017) predicted the prevalent emotion of a sentence and (Zhao and Zeng, 2017) used recurrent neural networks (RNNs) to predict the emoji which best describes a given sentence.

We build on top of this work to propose EmojiGAN $-$ a model meant to generate realistic emojis based on an image. Since we are interested in modeling a distribution over image-emoji tuples, it is reasonable to represent it using generative adversarial networks. They have been shown to successfully memorize distributions over both text and images. For example, a GAN can be coupled with RNNs in order to generate realistic images based on an input sentence (Reed et al., 2016).

We train our algorithm on emoji-picture pairs obtained from various advertisement posts on Instagram. A practical application of our method is to analyze the effects of product advertisement on Instagram users. Previous works attempted to predict the popularity of Instagram posts by using surrogate signals such as number of likes or followers (Almgren et al., 2016; De et al., 2017). Others used social media data in order to model the popularity of fashion industry icons (Park et al., 2016). A thorough inspection of clothing styles around the world has also been conducted (Matzen et al., 2017).

3 Proposed Approach

3.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have recently gained huge popularity as a blackbox unsupervised method of learning some target distribution. Taking roots in game theory, their training process is framed as a two player zero-sum game where a generator network G tries to fool a discriminator network D by producing samples closely mimicking the distribution of interest. In this work, we use Wasserstein-GAN (Arjovsky et al., 2017), a variant of the original GAN which uses the Wasserstein metric in order to avoid problems such as mode collapse. The generator and the discriminator are gradually improved through either alternating or simultaneous gradient descent minimization of the loss function defined as:

$$\min_G \max_D \mathop{\mathbb{E}}_{x\sim f_X(x)}[D(x)] + \mathop{\mathbb{E}}_{x\sim G(z)}[-D(x)] + p(\lambda), \tag{1}$$

where $p(\lambda) = \lambda(||\nabla_{\tilde{x}} D(\tilde{x})|| - 1)^2$, $\tilde{x} = \varepsilon x + (1-\varepsilon)G(Z)$, $\varepsilon \sim \text{Uniform}(0,1)$, and $Z \sim f_Z(z)$. This gradient penalized loss (Gulrajani et al., 2017) is now widely used to enforce the Lipschitz continuity constraint. Note that setting $\lambda = 0$ recovers the original WGAN objective.

3.2 Choice of embedding

Multiple embeddings have been proposed to encode language entities such as words, ideograms, sentences and even documents. A more recent successor of word2vec, emoji2vec aims to encode groups of words represented by visual symbols (ie ideograms or emojis). This representation is a fine-tuned version of word2vec which was

trained on roughly 1,600 emojis to output a 300-dimensional real-valued vector. We experimented with both word2vec and emoji2vec by encoding each emoji through a sum of the word2vec representations of its textual description. We observed that both word2vec and emoji2vec embeddings yielded only a mild amount of similarity for most emojis. Moreover, dealing with groups of words requires to design a recurrent layer in the architecture, which can be cumbersome and yield suboptimal results as opposed to restricting the generator network to only Unicode characters. Bearing this in mind, we decided to use the emoji2vec embedding in all of our experiments.

3.3 Learning a skewed distribution

Just like in text analysis, some emojis (mostly emotions such as love, laughter, sadness) occur more frequently than domain-specific pictographs (for example, country flags). The distribution over emojis is hence highly skewed and multimodal. Since such imbalance can lead to a considerable reduction in variance, also known as mode collapse, we propose to re-weight each backward pass with coefficients obtained through either of the following schemes:

- term frequency-inverse document frequency (*tf-idf*) weights, a classical approach used in natural language processing (Salton and Buckley, 1988);

- Exponentially-smoothed raw frequencies:

$$w_s(e) = \frac{\exp^{-k \times freq(e)}}{\sum_{i=1}^{N} \exp^{-k \times freq(e_i)}} \; \forall e, k \geq 0 \quad (2)$$

where k is a smoothing constant and $freq(e) = \frac{count(e)}{N}$ is the frequency of emoji e and N is the total number of emojis.

3.3.1 Algorithm

Our method relies on the conditional version of WGAN-GP which accepts fixed size $(64 \times 64 \times 3)$ RGB image tensors. Our approach is presented in Algorithm. 1, shown below:

Algorithm 1 Conditional Wasserstein GAN

Input: Tuple of emojis and images (X, Y), the gradient penalty coefficient λ, the number of critic iterations per generator iteration n_{critic}, the batch size m, learning rate l_r and weight vector w.
Initialization: initialize generator parameters θ_{G_0}, critic parameters θ_{D_0}
for epoch $= 1, ..., N$ **do**
  for $t = 1, ..., n_{critic}$ **do**
    {Updating Discriminator}
    for $n = 1, ..., n_{disc}$ **do**
      Sample $\{x\}_{i=1}^{m} \sim X$, $\{y\}_{i=1}^{m} \sim Y$, $\{z\}_{i=1}^{m} \sim \mathcal{N}(0,1)$, $\{\epsilon\}_{i=1}^{m} \sim U[0,1]$
      $\tilde{x_i} \leftarrow \epsilon x_i + (1 - \epsilon_i) G(z_i|y_i)$
      $\mathcal{L}^{(i)} \leftarrow D(G(z_i|(y_i)) - D(x_i|y_i) + \lambda(|\nabla_{\tilde{x_i}} D(\tilde{x_i}|y_i)| - 1)^2$
      $\theta_D \leftarrow \text{Adam}(\nabla_{\theta_D} \sum_{i=1}^{m} w_i \mathcal{L}^{(i)}, l_r)$
    end for
    {Updating Generator}
    for $n = 1, ..., n_{gen}$ **do**
      sample a batch of $\{z^{(i)}\}_{i=1}^{m} \sim N(0,1)$
      $\theta_G \leftarrow \text{Adam}(-\nabla_{\theta_G} \sum_{i=1}^{m} w_i \mathcal{L}^{(i)}, l_r)$
    end for
  end for
end for

4 Experiments

4.1 Data collection

We used the (soon to be deprecated) Instagram API to collect posts from top influencers within the following categories: fashion, fitness, health and weight loss; we believe that user data across those domains share similar patterns. Here, *influencers* are defined as accounts with the highest combined count of followers, posts and user reactions; 166 influencers were selected from various ranking lists put together by Forbes and Iconosquare. The final dataset has 80,000 (image, pictograph) tuples and covers a total of 753 distinct symbols.

4.2 Architecture

Inspired from (Reed et al., 2016), we performed experiments using the following architecture: the generator has 4 convolutional layers with kernels of size 4 which output a 4×4 feature matrix with a fully connexted layer; the discriminator is identical to G but outputs a scalar softmax instead of a 300-dimensional vector. The structure of both D and G is shown in Fig. 1.

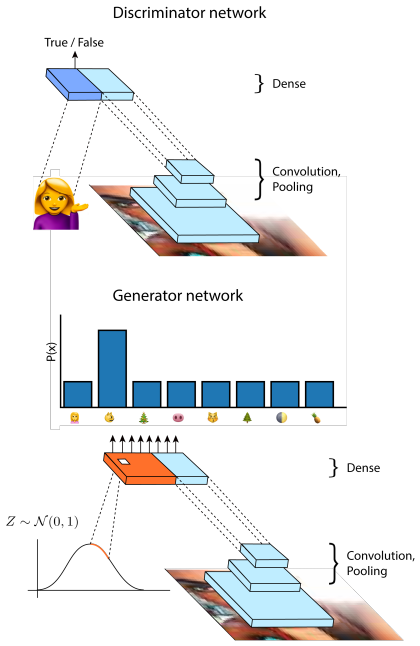

Figure 1: Illustration of how EmojiGAN learns a distribution. The generator learns the conditional distribution of emojis given a set of pictures while the discriminator assigns a score to each generated emoji.

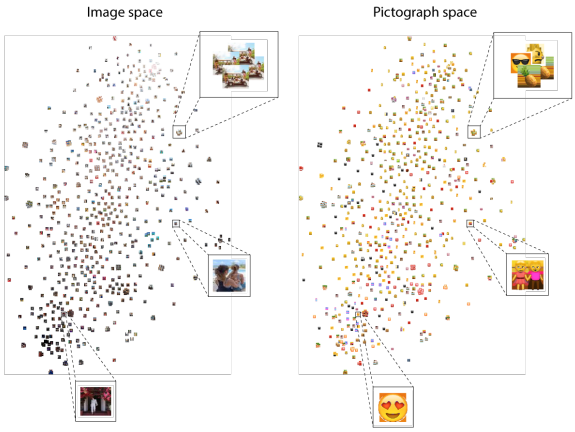

Figure 2: Visualization of t-SNE reduced images and their corresponding most frequent pictographs (emojis). The most popular emoji for each picture was obtained by sampling 50 observations from the generator and taking the mode of the sample. Note that even this technique has a stochastic outcome, meaning that if an image has a rather flat distribution, its mode will not be consistent across runs. The described behaviour can be observed in the upper right area of both space representations.

5 Results

A series of experiments were conducted on the data collected from Instagram. The best architecture was selected through cross-validation and hyperparameter grid search and has been previously discussed. The training process used minibatch alternating gradient descent with the popular Adam optimizer (Kingma and Ba, 2014) with a learning rate $l_r = 0.0001$ and $\beta_1 = 0.1$, $\beta_2 = 0.9$. We trained both G and D until convergence after aproximatively 10 epochs. Empirically, we saw that exponentially-smoothed raw frequencies weights (2) performed better than *tf-idf* weights.

In order to assess how closely the generator network approximates the true data distribution, we first sampled 750 images and obtained their respective emoji distribution by performing 50 forward passes through G. The *mode*, that is the most frequent observation in the sample, of the resulting distribution is considered as the most representative pictograph for the given image. We used t-SNE on the image tensor in order to visualize both the image and the emoji spaces (see Fig. 2). The purpose of the performed experiment was to assert whether two entities close to each other in the image space will also yield similar emojis. The top right corner of both clouds exposes a shortcoming of the algorithm: if the distribution is flat (i.e. is multimodal), even large samples will yield different modes just by chance. This phenomenon is clearly present throughout the cloud of pictographs: four identical images yield three distinct emojis. On the other hand, the two remaining examples correctly capture the presence of two people in a single photo (middle section), as well expression of amazement (bottom section).
The performance of generative models is difficult to assess numerically, especially when it comes to emojis. Indeed, the Fréchet Inception Distance (Heusel et al., 2017) is often used to score generated images but to the best of our knowledge, no such measure exists for ideograms. As an alternative way to assess the performance of our method, we plotted the true and generated distributions over 30 randomly chosen emojis for 1000 random images (see Fig. 3). While our algorithm relied on raw (i.e. uncleaned and unprocessed) data, we still observe a reasonable match between both distributions.

Fig. 4 reports the fitted distribution of the top 10 most frequent observations for three randomly sampled images. The top image represents a fashion model in an outfit; our model correctly captures the concepts of woman, love, and overall

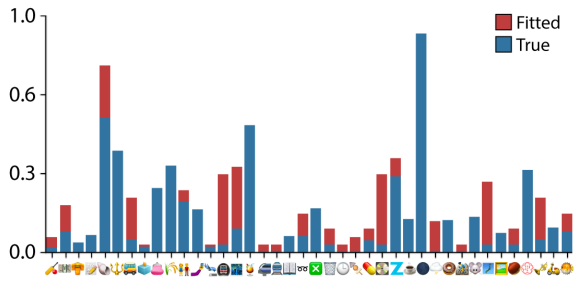

Figure 3: True and fitted distributions over 30 randomly sampled emojis for 500 randomly sampled images. Probabilities are normalized by the maximal element of the set.

positive emotion in the image. However, EmojiGAN can struggle with filtering out unrealistic emojis (in this case, pineapple and pig nose) for images with very few distinct ideograms. The bottom subfigure outlines another very common problem seen in GANs: mode collapse. While the generated emoji fits in the context of the image, the variance in this case is nearly zero and results in G learning a Dirac distribution at the most frequent observation.

The middle image also suffers from the above

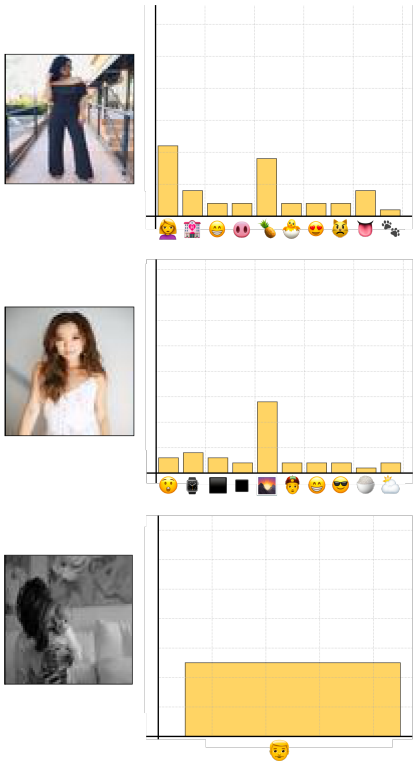

Figure 4: Emojis sampled for some Instagram posts: observe the mode collapse in the bottom subfigure as opposed to more equally spread out distributions.

problems (the sunset pictograph dominates the distribution). We note how algorithms based on unfiltered data from social networks are prone to ethical fallacies, as illustrated in the middle image. This situation is reminiscent of the infamous Microsoft chatbot Tay which started to pick up racist and sexist language after being trained on uncensored tweets and had to be shut down (Neff and Nagy, 2016). We ourselves experienced a similar behaviour when assessing the performance of EmojiGAN. One plausible explanation of this phenomenon would be that while derogatory comments are quite rare, the introduction of *exponential weight* or similar scores in the hope of preventing mode collapse to the most popular emoji has the side effect of overfitting least frequent pictographs.

6 Conclusion and Discussion

In this work, we proposed a new way of modeling social media posts through a generative adversarial network over pictographs. EmojiGAN managed to learn the emoji distribution for a set of given images and generate realistic pictographic representations from a picture. While the issue of noisy predictions still remains, our approach can be used as an alternative to classical image annotation methods. Using a modified attention mechanism (Xu et al., 2015) would be a stepping stone to correctly model the context-dependent connotations (Jibril and Abdullah, 2013) of emojis. However, the biggest concern is of ethical nature: training any algorithm on raw data obtained from social networks without filtering offensive and derogatory ideas is itself a debate (Islam et al., 2016; Davidson et al., 2017).

Future work on the topic should start with a thorough analysis of algebraic properties of emoji2vec similar to (Arora et al., 2016). For example, new Unicode formats support emoji composition, which is reminiscent of traditional word embeddings' behaviour and could be explicitly incorporated into a learning algorithm. Finally, the ethical concerns behind deep learning without limits are not specific to our algorithm but rather a community-wide discourse. It is thus important to work together with AI safety research groups in order to ensure that novel methods developed by researchers learn our better side.

References

Khaled Almgren, Jeongkyu Lee, et al. 2016. Predicting the future popularity of images on social networks. In *Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on SocialInformatics 2016, Data Science 2016*, page 15. ACM.

Martín Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN. *CoRR*, abs/1701.07875.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. Linear algebraic structure of word senses, with applications to polysemy. *arXiv preprint arXiv:1601.03764*.

Spencer Cappallo, Thomas Mensink, and Cees G.M. Snoek. 2015. Image2emoji: Zero-shot emoji prediction for visual media. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 1311–1314, New York, NY, USA. ACM.

Bo Dai, Dahua Lin, Raquel Urtasun, and Sanja Fidler. 2017. Towards diverse and natural image descriptions via a conditional GAN. *CoRR*, abs/1703.06029.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009.

Shaunak De, Abhishek Maity, Vritti Goel, Sanjay Shitole, and Avik Bhattacharya. 2017. Predicting the popularity of instagram posts for a lifestyle magazine using deep learning.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bosnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. *CoRR*, abs/1609.08359.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *CoRR*, abs/1708.00524.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial networks. *CoRR*, abs/1406.2661.

Amit Goyal, Francesco Bonchi, and Laks VS Lakshmanan. 2010. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 241–250. ACM.

Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. 2017. Improved training of wasserstein gans. *CoRR*, abs/1704.00028.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500.

Aylin Caliskan Islam, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. *CoRR*, abs/1608.07187.

Tanimu Ahmed Jibril and Mardziah Hayati Abdullah. 2013. Relevance of emoticons in computer-mediated communication contexts: An overview. *Asian Social Science*, 9(4):201.

Ryan Kelly and Leon Watts. 2015. Characterising the inventive appropriation of emoji as relationally meaningful in mediated close personal relationships. *Experiences of Technology Appropriation: Unanticipated Users, Usage, Circumstances, and Design*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Alex Krizhevsky. 2012. Learning multiple layers of features from tiny images.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Kevin Matzen, Kavita Bala, and Noah Snavely. 2017. Streetstyle: Exploring world-wide clothing styles from millions of photos. *CoRR*, abs/1706.01869.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Hannah Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2016. Blissfully happy or ready to fight: Varying interpretations of emoji. *Proceedings of ICWSM*, 2016.

Gina Neff and Peter Nagy. 2016. Automation, algorithms, and politics— talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication*, 10:17.

Jaehyuk Park, Giovanni Luca Ciampaglia, and Emilio Ferrara. 2016. Style in the age of instagram: Predicting success within the fashion industry using social media. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 64–73. ACM.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. A machine learning approach to twitter user classification. *Icwsm*, 11(1):281–288.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM.

Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. 2016. Variational autoencoder for deep learning of images, labels and captions. In *Advances in neural information processing systems*, pages 2352–2360.

Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. *CoRR*, abs/1605.05396.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659.

Luda Zhao and Connie Zeng. 2017. Using neural networks to predict emoji usage from twitter data.