From Images to Emotions and Emotions to Images: Toward Emotion-Driven Scenes

Anna Lai
Stanford University
alai2@stanford.edu

Michelle Lam
Stanford University
mlam4@stanford.edu

Emily Ling
Stanford University
eling8@stanford.edu

Abstract

Recent methods for semantic image synthesis require a full segmentation map of the scene. We are interested in using the abstracted input of emotion to generate image content. We build upon related work in image emotion classification and image content prediction to enable prediction of scene graph components and composition based on an input emotion. Our baseline approaches map images to a single emotion and we extend upon this to predict more detailed emotion distributions building upon state-of-the-art architectures. We explore object prediction and detection methods to synthesize scene graphs with ground truth emotion distributions and vice-versa. To accomplish our task we use the Emotion6 dataset which maps image to emotional responses and Visual Genome to gather images annotated with scene graphs. From our experiments, we developed an Emotion Stimuli Map predictor and integrate it into an enhanced model to predict emotions distributions with an MSE loss of 0.0159*, a model that to predict emotions distributions from objects only with an MSE loss of* 0.042*, and models to generate objects from emotion distributions on the Visual Genome dataset.*

1. Introduction

Our project was motivated by an interest in extending the flexibility of image synthesis models, particularly to give users the ability to incorporate *emotion* into the synthesized image. Towards this end, we focused on the relationship between the emotion conveyed by an image and the image's visual content, in the form of scene graph data. We envision a system that, given a user's emotional intent, could generate relevant scene components and eventually a full image based on that emotion. One potential application of this work would be to extend semantic image synthesis models such as GauGAN [18], which enables users to control the semantics and style of a scene using a segmentation mask and example style images (Figure 1).

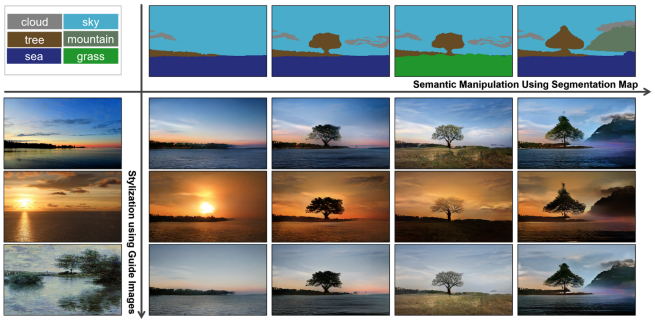

Figure 1. Summary of semantic and style-based inputs to GauGAN

For the scope of this project, we focus on predicting an image's emotion distribution given the image and its scene graph, and then reversing that mapping to predict elements of an image's content given an emotion distribution. To do this, we propose the following process. First, we train an emotion distribution predictor on an image emotion classification dataset, Emotion6 [19]. Then, we once we achieve a satisfactory model, we run this model on a dataset with human-annotated scene graphs, such as Visual Genome, to learn a mapping between scene graphs and emotions. Finally, we reverse this mapping to enable prediction of scene graph components based on an input emotion distribution.

2. Related Work

2.1. Image Emotion Classification

In computer vision, there has been increased interest in image classification that focuses on less-concrete, abstract concepts such as aesthetic quality and image emotion (or affect) [8]. In particular, image emotion prediction is a challenging task because, while aspects like aesthetics are quite directly tied to general image-based features like color, composition, and resolution, image emotion can arise from higher-level concepts and associations that may not interact logically and may be difficult to glean from traditional processing on a raw image. Much of the work in affective image classification has focused on classification of a dominant image emotion [9, 17, 26, 20, 2]. However, more

recently there has been a shift toward more nuanced and human-oriented representations of image emotion. Peng et al. introduced the Emotion6 dataset that labeled images not just with a single emotion label, but with a distribution corresponding to several core emotion categories [19]. Further work has built upon the Emotion6 dataset to include a heatmap of the regions that most strongly influenced the image emotion judgment [21]. More recently, work has investigated the ties between human judgment of image emotion, the level of attention paid to an image, and the relative level of attention within an image [3]. In this work, we aim to continue these trends toward more rich representations of image emotion that capture the way a human attends to the components of an image as we aim to flip this process to enable emotion-driven image creation that must capture the way humans experience and reason about emotion.

2.2. Image Content Prediction

2.2.1 Object Detection

A large variety of different models have been proposed for segmenting and labeling objects within an image [5, 14, 15]. This task involves not only correctly detecting all objects in a scene, but also labeling them and precisely determining the bounding box for each object. There has also been an effort towards increasing the speed of object detection systems, such as the YOLO real-time object detection system [24, 23]. While many object detection combine multiple localized classifiers across different locations and scales of the image, YOLO applies a single neural network to the image in entirety, making it extremely fast.

2.2.2 Scene Graph Generation

Scene graph generation is an extension of the object detection task, predicting not only the object labels and locations, but also the pair-wise relations between objects within an image. A number of different scene graph generations models have been proposed [13, 12, 7, 27]. The current state-of-the-art system on the Visual Genome dataset is Graph R-CNN, which uses an attentional Graph Convolutional Network (aGCN) and Relation Proposal Network (RePN) to effectively and efficiently consider objects and potential relations [27].

3. Methods

3.1. Emotion classification prediction

The baseline predictor uses an SVM method using extracted features from the image data set regarding texture and color, a subset of features used by Peng et al. The texture-related features were based on the Gray-Level Co-occurrence Matrix (GLCM), including features such as the contrast, energy, homogeneity, correlation, dissimilarity, and ASM. For the color-related features, we used cascaded CIECAM02 color histograms for the whole image, which included hue, lightness, brightness, chroma, colorfulness, and saturation [19].

The secondary approach uses a CNN to directly map images to emotion predictions. We experimented with CNN models that purely take in image input to predict the final emotion distribution.

For our first CNN model, we experiment with a design with 4 CONV layers, ReLU activations, dropout, and two fully-connected layers. In preprocessing, we resize the images to a standard 32x32 dimension. We utilized an Adam optimizer with a learning rate of $1e-3$, dropout probability of 0.2, and batch size of 16.

For our second model, we followed the example provided by Peng et al. that achieved state-of-the-art performance and utilize AlexNet with pre-trained weights (trained on ImageNet) and substitute the last fully-connected layer with an output dimension of 7 for the 7 emotional classes [19, 11, 1]. We again used an Adam optimizer with learning rate of $1e-4$.

3.2. Emotion distribution prediction

Next, since our task is motivated by the goal of enabling greater expressivity in image construction/synthesis, we worked on predicting emotion *distributions* rather than just the maximum emotion class. A distribution of emotions, while still more limited than the full range of human emotions, better captures the nuanced set of emotions that an image may convey. Thus, while this task is more challenging than the max-emotion task, it is an important step that can bring us closer to a richer, emotion-driven mode of image interaction.

For this task, we first experimented with a model that built upon AlexNet. We again adjust the final fully-connected layer to have an output dimension of 7 to match the number of emotional classes, but add a softmax layer to transform these outputs into probabilities. To train our model to learn the ground truth distributions, we also change to a mean squared error (MSE) loss function. As before, we performed hyperparameter tuning on our validation set. Next, we experimented with similar model architectures that built upon ResNet and VGG models also pretrained on ImageNet; for our ResNet-based model, we utilized ResNet-101, and for our VGG-based model, we utilized VGG-19 with batch normalization [6, 25].

3.3. Emotion Stimuli Map (ESM) prediction

In addition to these models, we also wanted to explore methods that incorporated some notion of *attention* to particular regions of an image. While the Emotion6 dataset correlates an emotion distribution to an entire image as a whole, the emotion of an image is oftentimes tied to a par-

ticular *region* of the image. Thus, we turned to the EmotionROI dataset, another image dataset which pairs Emotion6 images with a ground truth pixel map of the *regions* that evoked the indicated emotion [21]. This pixel map is termed the Emotion Stimuli Map (ESM); these maps were drawn by Amazon Mechanical Turk (AMT) workers who were asked to draw a rectangle around the portion of the image that most influenced the evoked emotion distribution.

Using this dataset, we aimed to create a supplementary model that, given an input image, would predict its ESM. Then, once we have sufficiently trained this model, we incorporated this model into our emotion distribution prediction model as a feature to capture some notion of "attention" to regions of the image that may be more relevant to the emotion distribution.

For our ESM prediction model, we took an approach inspired by the Fully *Convolutional* Networks (FCN) introduced to predict semantic segmentation [16]. This work outperformed state-of-the-art semantic segmentation models by utilizing convolutional networks *by themselves* to transform from pixels to pixels. Since this ESM prediction task aims to map from raw images pixels to ESM intensity pixels, this model setup seemed promising. However, since we aim to predict pixel values across a continuous range rather than predict per-pixel class belongingness, we made several adapatations as used in Peng et al.'s follow-up work [21]. We built upon FCN-8s at-once, which is fine-tuned from a pretrained VGG-16 model, but we changed the output layer to predict just one class, and we utilized a mean-squared error (MSE) loss during training to compare the model output with the ground-truth ESMs.

For our ESM-enhanced emotion distribution prediction model, we built upon our ResNet-based emotion distribution model. In this enhanced model, we passed each batch of image inputs into our trained ESM prediction model to generate a predicted heatmap of emotionally-relevant image regions. Then, this predicted ESM was concatenated with the layer prior to the fully-connected portion of the model and we again used MSE loss for training.

3.4. Object prediction

As part of our goal to link an image's emotion distribution to image content, we needed to identify objects based on an image. Since the primary purpose of this subtask is to generate a dataset for other tasks further down the pipeline, we opted to run a pretrained model rather than build and train our own model from scratch.

We attempted to use a number of different pretrained scene graph generation models, including Multi-level Scene Description Network (MSDN) [13], FactorizableNet [12], and Scene Graphs with Permutation-Invariant Structured Prediction [7]. However, we encountered significant technical and versions issues and were unable to successfully run any of these models on the Emotion6 dataset required for this task. Additionally, we decided to reduce the scope of this project from incoporating a full scene graph to considering only the objects, so thus we shifted our focus to image object detection instead.

For object detection, we used the pretrained YOLO object detection system [24, 23], to predict object labels and bounding boxes given an input image. The YOLO system can detect over 9000 object categories, and scores 55.3 mAP on the COCO dataset, which is close to state-of-the-art. We chose this system because it is extremely fast, as it only requires a single network evaluation per image, making it several orders faster than similar object detection systems such as R-CNN [4].

3.5. Emotion prediction from objects

From the data synthesized from running the emotion distribution predictor on Visual Genome images, we look at generating emotion distributions from only the objects found in the Visual Genome images. For the baseline model, we used pretrained Global Vectors for Word Representation (GloVe) [22] word embeddings. GloVe uses a co-occurence matrix to obtain a vector representation of words and can capture linguistic or semantic similarities of the corresponding words and well as linear substructures that quantify the relationship between words. The GloVe embeddings used in our models convert words to a 50-dimensional vector. These vector embeddings of the objects are averaged over each object found in the training example which is passed into a single fully-connected layer and mapped to emotion distribution outputs using the MSE loss on the output of the softmax.

Next we incorporate scene graph data of the relationships between objects in the image. From the Visual Genome [10] dataset these relationships consist of subject-predicate-object phrases such as "cat on fence." Using these relationships from each image, we concatenate these phrases to form a paragraph which is then embedded using GloVe vectors and passed into an RNN and LSTM architecture using the MSE loss on the last softmax layer to output an emotion distribution.

3.6. Object prediction from emotion distribution

Now, reversing the mapping above, we want to predict objects based on only the emotion distribution of an image.

3.6.1 Baseline

For the baseline model, we compute a score for each emotion and object, with the score defined as follows:

$$\text{score}(e, o) = \sum_{i \in \text{Images}} i_o * d_e$$

Here, i_o is the number of occurrences of object o in image i, and d_e is the value of the emotion distribution d at emotion class e. We then compute the average score for each object across all 7 emotion classes:

$$\text{avg-score}(o) = \sum_{e \in \text{Emotions}} \text{score}(o)/7$$

The final score for each emotion and object in the training set is the difference between the object score for that emotion and the average object score across all emotions: $\text{final-score}(e, o) = \text{score}(e, o) - \text{avg-score}(o)$.

An initial attempt at ranking based strictly on the object frequency per emotion class returned the same highest-frequency objects across all seven classes, which was an undesirable result. So, this metric is designed to give higher scores to objects that are especially salient for a particular emotion class, while giving lower scores to objects that appear at high frequencies across all emotion classes.

To evaluate this baseline model on an emotion distribution d, we compute $\sum_e d_e \text{final-score}(e, o)$ for all objects o, then return a vector containing the ranking scores for the k highest-ranked objects.

3.6.2 Neural Network Model

We construct a neural network that takes the 7-dimensional emotion distribution vector as input, and outputs a vector containing a confidence rating for each object in the possible object set. The ground truth output is a vector where the confidence rating is the object occurrence count in the image divided by the total object occurrence count in the training set. This normalization was done in an attempt to encourage less frequent but more salient objects to be predicted. Other forms of normalization were also tried, including: dividing by the square root or log of the total occurrence count, dividing by the average occurrence count per image that contains the object, multiplying by (the square root, or log of) the area of each object bounding box.

We use the following neural network architecture: $FC(7, 500) \rightarrow ReLU \rightarrow Dropout(0.5) \rightarrow FC(500, 2000) \rightarrow ReLU \rightarrow Dropout(0.5) \rightarrow FC(2000, 1500) \rightarrow ReLU \rightarrow FC(1500, 1500) \rightarrow Sigmoid$. The network uses cross entropy loss, and is trained for 50 epochs using an Adagrad optimizer with learning rate of 0.005 and learning rate decay of 0.01.

4. Dataset and Features

4.1. Existing datasets

The emotion dataset we use for this task is the Emotion6 dataset, which consists of 1980 images that are labeled with scores for Ekman's six basic emotions: anger, disgust, joy, fear, sadness, and surprise) and a *neutral* category [19]. These images were gathered from Flickr using those six emotion keywords, and emotional responses were collected using Amazon Mechanical Turk (AMT).

Our image scene graph dataset is Visual Genome, a dataset of images densely annotated with scene graphs (*objects* connected by *relationships* and modified by *attributes*) [10]. This dataset enables us to work not only at the pixel level, but to bridge this low-level information with human-interpretable, semantic image representations. We hypothesize that by incorporating this information into our models, we will be able to make greater progress toward emotion classification since emotion is often tied to higher-level concepts rather than pixel-oriented attributes.

4.2. Synthesized datasets

Our goal was to learn a mapping from scene graph objects to an emotion distribution, and vice versa. In order to do this, we needed to synthesize our own datasets that contained both scene graph and emotion distribution labels for the same images. We did this by merging the Emotion6 and Visual Genome datasets with two different approaches:

4.2.1 Dataset A: Ground truth emotion distribution, Synthesized scene graphs

For our first attempt at a combined image, emotion distribution, and objects dataset, we ran the YOLO object detection system on the Emotion6 dataset, which already contains images and ground truth emotion distribution labels. Emotion6 images were fed to the YOLO system one by one using a detection threshold of 20% confidence, resulting in a list of object names, confidences, and bounding boxes for each image. Figure 2 shows a sample output from running the YOLO system on an Emotion6 dataset image.

4.2.2 Dataset B: Synthesized emotion distribution, Ground truth scene graphs

To map emotion distributions to scene graphs, we run our model trained on the Emotion6 dataset to predict emotion distributions for the Visual Genome images with scene graphs. We used a 70-15-15 train-val-test split, which gave us 1386 training examples, 297 validation examples, and 297 test examples using the Emotion6 dataset. The average emotion distribution over this 2000-image subset of Visual Genome is shown in Figure 3.

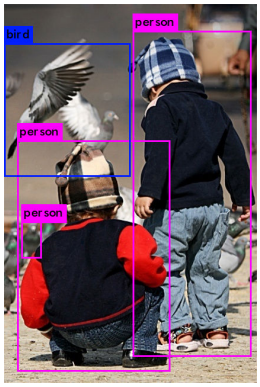

Figure 2. Example object detection from running YOLO on an Emotion6 image with label "joy." Objects were predicted with the following confidences: person: 100%, person: 100%, bird: 89%, person: 64%. We see that there is an incorrect prediction of "person" for what should be "bird," but this incorrect prediction has lower confidence and thus can potentially be distinguished.

Figure 3. Average emotion distribution predicted across all images in the Visual Genome subset.

5. Experiments/Results/Discussion

5.1. Emotion classification prediction

5.1.1 Baseline

For the baseline, we used an SVM to predict the maximum class label for each image. In preprocessing, we concatenate all texture and color features, as described in Methods, to form the input. We also convert each emotion distribution in the Emotion6 dataset into a single label representing the maximum emotion class for each input image. We use a radial basis function (RBF) kernel for the SVM.

We achieve a 33.33% validation accuracy and a 31.31% test accuracy. This accuracy is over twice the accuracy of randomly guessing the class, which is 14.3% since there are 7 emotion classes. However, further examination of the prediction results reveal that the model is predicting the same class for nearly every test image.

Examining the dataset, we see that the classes are quite imbalanced. "Joy" is by far the most common emotion class, so predicting "joy" for all images allows us to achieve over 30% accuracy. In Figure 4, we have the number of images in the full dataset with each emotion class.

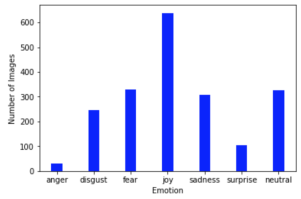

Figure 4. Emotion6 class distribution

5.1.2 Model 1

For the first simpler CNN-based model architecture on this maximum-emotion classification task, we achieve a validation accuracy of 39.06% and a test accuracy of 29.63%.

5.1.3 Model 2

For the AlexNet-based architecture, We performed hyperparameter tuning on the validation set and achieved a validation accuracy of 50.51% and a test accuracy of 48.48%.

5.2. Emotion distribution prediction

To evaluate the performance of our best emotion distribution prediction models in each category, we investigated several metrics designed to measure similarity between distributions. The summary of these metrics with respect to our models are shown in Figure 6. First, we investigated Kullback-Leibler (KL) Divergence, or relative entropy, a metric that captures how different one probability distribution is from another. What we've termed "max-emotion distance" is the difference between the distributions for the emotion that had the maximum value in the ground truth emotion distribution. We felt that this would be an important metric to capture as, for each image, the dominant emotion is the most critical to get right. Finally, we looked at the Bhattacharyya Coefficient (BC), which approximates the amount of overlap between two samples.

Based on these metrics, we see that overall, the ESM-enhanced model outperformed the ResNet-based model, which outperformed the VGG-based model, which in turn outperformed the AlexNet-based model. The MSE loss acheived on the test-set is shown in Figure 7 and follows this same performance trend. All of these models (especially those other than the AlexNet-based version) appear to outperform Peng et al.'s 2015 model, which achieved a CD of 0.265 and a BC of 0.847.

To gain a better understanding of these models, we wanted to investigate examples of the emotion distributions themselves. We created visualizations that display the input image, predicted emotion distribution, and ground truth emotion distribution as shown in Figure 5. We see that on these non-training-set images, both the ResNet-based (left bar in blue) and VGG-based (right bar in cyan) emotion distributions correlate quite well with the ground truth emotion

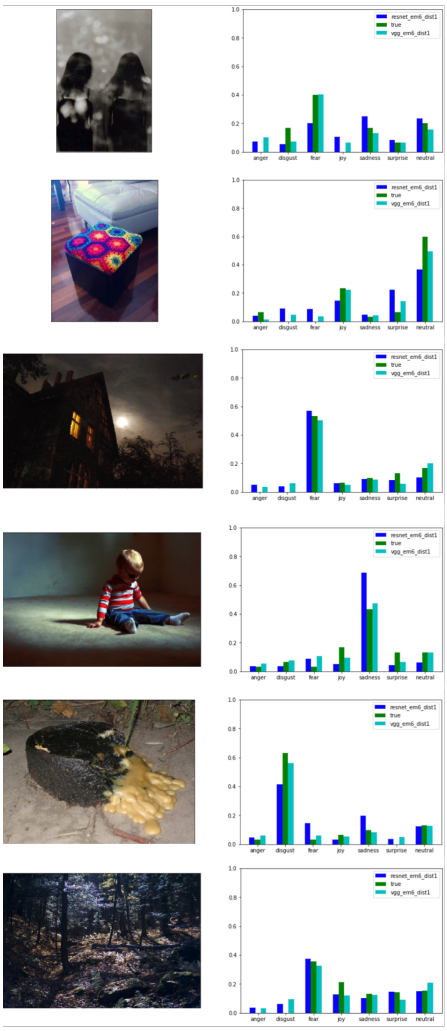

Figure 5. A comparison between the true emotion distribution and the distributions predicted by the ResNet-based and VGG-based models for a variety of non-training images.

	AlexNet	VGG	ResNet	ESM-enhanced
KL-Divergence	3.394	2.967	2.760	**2.732**
Chebyshev dist	0.265	0.247	0.239	**0.229**
Max-emotion dist	0.231	0.218	0.209	**0.200**
Bhattacharyya coeff	0.851	0.871	0.877	**0.880**

Figure 6. Metrics comparing the distance (KL-divergence, Chebyshev distance, Max-emotion distance) or similarity (Bhattacharyya coefficient) of the predicted Emotion6 distribution and the true distribution on the test set. The degree of yellow tint indicates the relative degree of *similarity* of the predicted and true distributions (darkest colors correspond to min values for the first three distance metrics; max values for the fourth similarity metric).

distributions (center bar in green). For most images with a strong dominant emotion, these two models correctly mirror the presence of this primary emotion and mostly match its magnitude. While there is greater variation between the model-predicted values and the ground truth value for some of the lower-probability emotions, in general the predictions are able to match these emotion categories as well.

	AlexNet	VGG	ResNet	ESM-enhanced
Test-set MSE Loss	0.0210	0.0177	0.0168	**0.0159**

Figure 7. Comparison of mean-squared error loss on the test set for each category of the emotion distribution prediction models.

5.3. Emotion Stimuli Map (ESM) prediction

For the ESM prediction task, we aimed to predict the emotion stimuli map for input images. We experimented with models that transformed the input images to a standard 32x32 size or a standard 64x64 size. After hyperparameter tuning and bringing down the validation-set loss from 0.05107 to 0.0374, we were able to achieve a test-set MSE loss of 0.0372. Representative examples of the inputs, ground truth ESMs, predicted ESMs, and a visualization of their differences are displayed in Figure 8. Based on this qualitative evaluation, we find that our predicted ESMs generally align well with the most dominant portions of the ground truth ESMs. However, the predicted ESMs tend to be less well-defined than the ground truth ESMs and oftentimes gravitated more toward the center of the image rather than the extremities. Overall though, we had enough confidence in the model to incorporate it into our emotion distribution prediction model and found that even in this form, our emotion distribution model acheived performance gains with this added image region-based emotion information.

5.4. Emotion prediction from objects

Using the pretrained GloVe word embeddings of the just objects to train a network to predict the emotion distribution of the image proved effective in predicting the "ground truth" emotion distributions we synthesized achieving an MSE loss of 0.042. An example prediction on a test example is shown in Figure 9. This baseline model outperformed the RNN and LSTM networks that trained on richer scene graph data as they had a tendency to over-fit to the full paragraphs of scene relationships in the training set. However, adding more LSTM layers with dropout improved the results by regularizing model. Examining the paragraphs, simply concatenating the subject-predicate-object relationship data also results in some information loss about how the objects are connected which would influence how well the model would be able to generalize. The performance metrics comparing the emotion distributions generated from object and scene graph data and the "ground truth" emotion

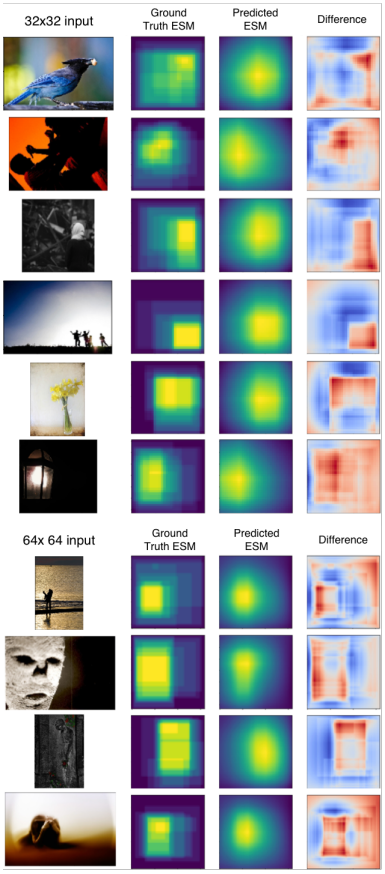

Figure 8. A comparison between the ground truth and predicted emotion stimuli maps (ESMs). The differences between the ground truth and predicted maps are displayed with red indicating regions with larger magnitude of difference. Model versions that convert the input image to 32x32 and 64x64 are compared.

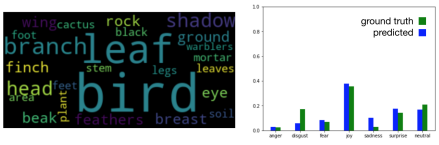

Figure 9. Wordcloud of objects from the test set and predicted emotion distributions from the baseline word embedding model compared to predicted "ground truth" distributions

	FC (object emb)	RNN	LSTM	LSTM (w/ dropout)
Kullback–Leibler div	**0.198**	1.794	1.738	1.728
Bhattacharyya coeff	**0.955**	0.393	0.399	0.400
Chebyshev dist	**0.066**	0.211	0.177	0.163

Figure 10. Performance metrics, defined in Section 5.2, of the emotion prediction from object and scene graph models.

distributions from image data are summarized in Figure 10.

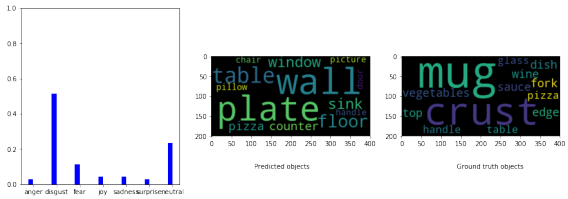

Figure 11. Input emotion distribution along with word clouds of objects from the ground truth scene graph, and objects predicted by a model trained on synthesized dataset B. This emotion distribution heavily weights "disgust."

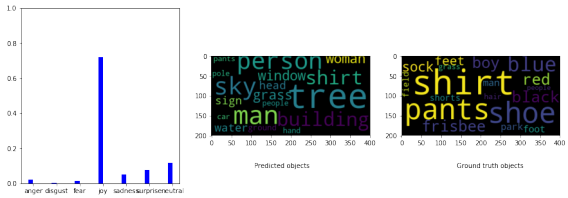

Figure 12. A second example of input emotion distribution with word clouds of ground truth and predicted objects. This emotion distribution heavily weights "joy."

5.5. Object prediction from emotion distribution

Both the baseline and neural network models were trained on both datasets A and B. A training/validation/test split 70%/15%/15% was used.

5.5.1 Baseline

Validation set loss values for datasets A and B can be found in Figure 14. Qualitatively, the baseline model performed quite well, particularly on dataset B, predicting reasonable objects for many emotion distributions.Figures 11 and 12 show two sample emotion distributions from dataset B, along with word clouds representing the ground truth and predicted object arrays. These emotion distributions are predicted using our model on Visual Genome, and the scene graph objects form the ground truth object array. For both of these examples, the baseline model correctly predicts several scene objects, and the general category of objects is aligned between the predicted and ground truth results.

Another qualitative metric we used to evaluate our model was to examine the top 15 objects predicted for a one-hot vector of each emotion (where the emotion distribution value is set to 1 for one emotion class, and 0 for all other classes). Results for dataset B are shown in Figure 14.

We can see that the predicted objects per class are quite relevant – the objects corresponding to "disgust" are generally associated with the bathroom or food; objects associated with "joy" are mostly outdoor objects, while those associated with "sadness" are typically indoor objects. The top 15 objects for dataset A are somewhat less relevant – for example, the top five objects in the "disgust" category are

anger	sideburn, stubble, moon, gray tshirt, white tshirt, ticker, power button, gaming controller, blood, blood spatter, left corner, copyright, black tie, dark background, black beard
disgust	sink, toilet, tile, towel, faucet, bathroom, counter, food, bowl, cabinet, mirror, tub, plate, lid, shower
fear	key, black mark, buttons, remote control, feather, device, steam, tooth, comfort, finger tip, children's book, presidents, engineer, station, bottle of wine
joy	man, tree, person, sky, shirt, grass, building, water, woman, ground, head, window, sign, car, trees
sadness	woman, wall, window, hair, shirt, bench, pillow, glasses, jeans, man, hand, couch, lamp, shoe, tie
surprise	sky, man, person, tree, water, plane, shirt, tail, grass, building, head, boat, trees, clouds, surfboard
neutral	window, building, wall, tree, sign, car, light, table, chair, pole, sky, plate, person, door, floor

Figure 13. Top 15 objects predicted for a one-hot emotion distribution of each emotion.

Dataset	Model	Loss
Dataset A	Baseline	0.12882
	NN Model	0.03479
Dataset B	Baseline	0.00466
	NN Model	0.00355

Figure 14. Validation set loss values for the baseline and neural network values, run on datasets A and B.

cow, fire hydrant, diningtable, scissors.

5.5.2 Neural Network Model

Training the model for 50 epochs resulted in a training loss of 0.003544 for dataset A and 0.003495 for dataset B. See Figure 14 for corresponding validation set loss values.

From a qualitative perspective, the baseline significantly outperformed the neural network model. The top objects predicted for each emotion class were generally very similar across all emotion classes, with only slight variation in the relative ordering. Depending on the normalization used, these top objects were either the most common objects in the dataset (objects such as man, window, building, tree, and person) or uncommon, specific objects (such as right ear, grapes, clock face, apple, lace). Although we experimented with many different forms of object frequency normalization, we couldn't seem to strike the right balance between predicting likely objects and bringing out salient, more uncommon objects per class.

The neural network model likely did not perform well because the cross-entropy loss function encourages the output to match the ground-truth object distributions as much as possible, but the most frequent object set across the different emotion classes is the same. Thus, the model is encouraged to predict only the most frequent objects. Ideally, the loss function needs to somehow prioritize predicting objects that are specifically salient to the class. A potential solution would be to use a GAN, where the generator generates a list of objects given an emotion distribution, and the discriminator predicts an emotion distribution given a list of objects. This would encourage the generator to output objects associated with particular emotion distribution rather than generic objects and encourage a variety of equally valid object predictions for an emotion distribution.

Comparing the two synthetic datasets we created, we found that Dataset A, with ground truth emotion labels and predicted objects, seems to be a noisier and less effective dataset compared to dataset B. The results for dataset A seem qualitatively less relevant than the results for dataset B – this difference in quality is likely due to the noisiness and data sparsity of the object prediction compared to the ground-truth scene graphs. The object detection model we used struggled to identify blurry, small, or close-up objects, which occurred quite frequently in the Emotion6 dataset, making these object labels much less accurate or rich than the ground-truth scene graph labels.

6. Conclusion/Future Work

6.1. Future Work

Our eventual goal is to extend the flexibility and functionality of semantic image synthesis models like GauGAN [18]. First, we'd like to recommend scene graph compositions based on emotion distribution input so that users could interface with semantic image synthesis systems by specifying along dimensions of higher-order concepts like *emotions* rather than just concrete scene elements. In order to do so, we would need to map emotion distributions to not only objects but also object positions. Then, we would need to map object positions to object shapes to generate reasonable segmentation masks for given emotions and objects.

Further, we might work to to enable users to specify loose, blob-like object regions by learning to generate full high-quality segmentation masks given low-quality mask input. Additionally, we could relax the user input requirements by enabling users to simply specify a few objects (by text input) by learning to predict other objects that could realistically co-occur in the scene.

6.2. Conclusion

In this work, we have explored an initial approach towards a translation from emotional intent to scene component creation. We aimed to leverage emotion-labeled image datasets and scene graph-labeled image datasets to learn mappings between these domains. To do so, we generated proxy datasets based on ground-truth emotions with predicted scene components (objects) and based on ground-truth scene graphs with predicted emotion distributions. This step comprised a substantial portion of our task: learning to predict emotion distributions based on input images. Using CNN-based approaches and building a supplementary model that predicted emotion stimuli maps for images, we achieved our best-performing model which exceeded the performance of prior published work.

7. Contributions & Acknowledgements

7.1. Team member contributions

Anna Lai: Implemented and tuned FC, RNN, LSTM networks for using GloVe word embeddings for emotion distribution prediction from Visual Genome object and scene graph data. Wrote corresponding sections of the report, abstract.

Michelle Lam: Implemented and tuned CNN-based Emotion classification prediction, implemented and tuned Emotion distribution prediction (AlexNet, ResNet, VGG), generated model-based emotion predictions for Visual Genome images, implemented and tuned Emotion Stimuli Map (ESM) prediction and the ESM-enhanced emotion distribution prediction model, implemented visualizations of emotion distributions and ESMs. Wrote corresponding portions of Related Work, Data, Model, and Results sections of the report, conclusion.

Emily Ling: Implemented emotion classification SVM baseline. Tested various scene graph generation and object detection pretrained models, then ran object detection model on Emotion6 dataset (dataset A). Implemented baseline and neural network models for object detection prediction from emotion distribution on both synthesized datasets, experimenting with various metrics and normalization types. Wrote corresponding portions of the report, and introduction.

7.2. Acknowledgements

For emotion classification and distribution prediction:

- Pytorch (https://pytorch.org/)
- Pytorch-FCN (https://github.com/wkentaro/pytorch-fcn)

For visual genome interface:

- Visual Genome Python Driver (https://github.com/ranjaykrishna/visual_genome_python_driver)

For object prediction:

- Darknet (https://github.com/pjreddie/darknet)
- YOLO: Real-Time Object Detection (https://pjreddie.com/darknet/yolo/)

For scene graph prediction:

- FactorizableNet(https://github.com/yikang-li/FactorizableNet)
- Scene Graphs with Permutation-Invariant Structured Prediction (https://github.com/shikorab/SceneGraph)

For word embeddings:

- GloVe (https://nlp.stanford.edu/projects/glove/)
- CS230: Emojify Assignment

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 2

[2] Y. Deng, C. C. Loy, and X. Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4):80–106, 2017. 1

[3] S. Fan, Z. Shen, M. Jiang, B. L. Koenig, J. Xu, M. S. Kankanhalli, and Q. Zhao. Emotional attention: A study of image sentiment and visual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7521–7531, 2018. 2

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*, 2014. 3

[5] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017. 2

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2

[7] R. Herzig, M. Raboh, G. Chechik, J. Berant, and A. Globerson. Mapping images to scene graphs with permutation-invariant structured prediction. In *Advances in Neural Information Processing Systems (NIPS)*, 2018. 2, 3

[8] D. Joshi, R. Datta, E. A. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo. Aesthetics and emotions in images. *IEEE Signal Processing Magazine*, 28:94–115, 2011. 1

[9] B. Jou, S. Bhattacharya, and S.-F. Chang. Predicting viewer perceived emotions in animated gifs. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 213–216. ACM, 2014. 1

[10] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and F. Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *CoRR*, abs/1602.07332, 2016. 3, 4

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 2

[12] Y. Li, W. Ouyang, Z. Bolei, S. Jianping, Z. Chao, and X. Wang. Factorizable net: An efficient subgraph-based framework for scene graph generation. In *ECCV*, 2018. 2, 3

[13] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2, 3

[14] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016. 2

[15] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017. 2

[16] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 3

[17] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 83–92. ACM, 2010. 1

[18] T. Park, M. Liu, T. Wang, and J. Zhu. Semantic image synthesis with spatially-adaptive normalization. *CoRR*, abs/1903.07291, 2019. 1, 8

[19] K. Peng, T. Chen, A. Sadovnik, and A. C. Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 860–868, 2015. 1, 2, 4

[20] K.-C. Peng, K. Karlsson, T. Chen, D.-Q. Zhang, and H. Yu. A framework of changing image emotion using emotion prediction. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 4637–4641. IEEE, 2014. 1

[21] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen. Where do emotions come from? predicting the emotion stimuli map. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 614–618. IEEE, 2016. 2, 3

[22] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. 3

[23] J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. *CoRR*, abs/1612.08242, 2016. 2, 3

[24] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018. 2, 3

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[26] X. Wang, J. Jia, J. Yin, and L. Cai. Interpretable aesthetic features for affective image classification. In *2013 IEEE International Conference on Image Processing*, pages 3230–3234. IEEE, 2013. 1

[27] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh. Graph R-CNN for scene graph generation. *CoRR*, abs/1808.00191, 2018. 2