Exploring Visual Engagement Signals for Representation Learning

Menglin Jia\textsuperscript{1,2}, Zuxuan Wu\textsuperscript{2,3}, Austin Reiter\textsuperscript{2}, Claire Cardie\textsuperscript{1}, Serge Belongie\textsuperscript{1}, Ser-Nam Lim\textsuperscript{2}
\textsuperscript{1}Cornell University \textsuperscript{2}Facebook AI \textsuperscript{3}Fudan University

Abstract

Visual engagement in social media platforms comprises interactions with photo posts including comments, shares, and likes. In this paper, we leverage such Visual Engagement clues as supervisory signals for representation learning. However, learning from engagement signals is non-trivial as it is not clear how to bridge the gap between low-level visual information and high-level social interactions. We present VisE, a weakly supervised learning approach, which maps social images to pseudo labels derived by clustered engagement signals. We then study how models trained in this way benefit subjective downstream computer vision tasks such as emotion recognition or political bias detection. Through extensive studies, we empirically demonstrate the effectiveness of VisE across a diverse set of classification tasks beyond the scope of conventional recognition 1.

1. Introduction

People post photos on social media to invite engagement and seek connections. A photo of a cute dog can resonate with other dog lovers and trigger reactions such as the “like” or “love” button or comments including “what an adorable dog” and “look at those blue eyes!” The widely available interactions with the photos posted on social media, which we call visual engagement, contain rich semantic descriptions (“dog”, “blue eyes”) and are far less expensive to obtain than manual annotations in standard computer vision tasks, including coarse and fine-grained class labels [11, 80, 95], bounding boxes [51, 23], and image captions [8].

More importantly, visual engagement, including comments, replies, likes, and shares, provides emotional and cultural context that goes beyond the image content therein. For example, the image in Fig. 1 could be described in a standard captioning task as “a dog sits next to a stuffed animal.” The social media audience of this post may react to the cuteness of the dog, comment on the torn stuffed animal with whimsical responses, or initiate a conversation.

The resulting textual descriptions depart from the exactly what it says on the tin approach of standard image captioning tasks and express private states [65, 82]: opinions, emotions, and speculations, for example. We argue that visual engagement can also serve as supervisory signals for representation learning and transfer well to subjective downstream computer vision tasks like emotion recognition or political bias classification.

Motivated by this observation, we propose to learn image representations from semantically and contextually rich Visual Engagement signals (VisE). We hypothesize that such learned representations, as a byproduct of mapping image content to human reactions, are able to infer private states expressed by images. This is beneficial and could serve as a nice addition to current computer vision research which in general focuses on the objectively present factual information from images (e.g., “this is a dog” vs. “what a cute dog!”).

Open-world visual engagement contains dense subjectivity clues, but is inherently noisy in nature. How to properly leverage such signals for representation learning is a challenging question. Inspired by recent studies on feature learning from proxy tasks [19, 3, 84], we cluster each type of visual engagement and obtain cluster assignment indices for all the responses associated with a training image. These cluster assignments are used as supervisory...
Learning image representations using natural language supervisory signals for representation learning. In the former case, we conduct the clustering on representations encoded by a textual model. Unlike most existing multi-modal methods that perform pre-training for both language and vision modules [12, 66] with hundreds of millions of parameters, we simply use an off-the-shelf encoder to embed comments, which is computationally efficient. We then evaluate representations learned from engagement signals on downstream tasks.

Our main contribution is to demonstrate that social media engagement can provide supervision for learning image representations that benefit subjective downstream tasks. To this end, we explore VisE pre-trained on 250 million publicly available social media posts. Through extensive experiments, we show that in three downstream tasks related to private states detection, the learned VisE models can outperform the ImageNet-supervised counterpart by a substantial margin in some cases. These results highlight that VisE broadens the current representation learning paradigms, thereby narrowing the gap between machine and human intelligence.

2. Related Work

Visual representation learning Learning discriminative features for targeted datasets is a core problem for many computer vision research efforts. Feature learning on small datasets is particularly difficult since high-capacity deep neural networks suffer from overfitting when the amount of target training data is limited. One popular way to mitigate this issue is to pre-train on large-scale image datasets with manually curated class labels such as ImageNet and COCO [13, 15]. However, training on ImageNet requires manually labeled data, which are expensive to obtain and hard to scale. This motivates a plethora of work investigating weakly-supervised [73, 55, 41, 42, 47], semi-supervised [86, 87, 84] and self-supervised learning [3, 96, 14, 4, 22, 48, 90]. These methods tap into alternative forms of supervision, such as user-provided tags [76, 55] and hand-crafted pretext tasks (e.g., inpainting [64], colorization [91, 92], predicting jigsaw permutations [60], rotations of inputs [17]). More recently, contrastive learning [24, 25] is used for feature learning by bringing images closer to their augmented versions than other samples in the training set [83, 89, 26, 58, 6]. In this paper, we use visual engagement, which encompasses high-level semantics, as supervisory signals for representation learning.

Learning image representations using natural language There is growing interest in learning joint visual-language representations [16, 72, 39, 49, 54, 5, 9]. Other studies convert language to discrete labels or continuous probability distributions, such as individual word, part-of-speech (POS) tags, clustering assignments of sentence features, and topic modeling [38, 46, 19, 2]. Some approaches learn visual representations from a pretext task that predicts the natural language captions from images [2, 12, 69]. Recently introduced methods including ConViRT [93], CLIP [66] and ALIGN [32] allow one to learn visual representations with contrastive objectives using image-text pairs. These works use objective natural language, which describes and informs the content of the images, and mostly evaluate their methods on conventional recognition datasets. We instead utilize the density of subjectivity clues in social media engagement and explore the transferability of the learnt representation to alternative downstream tasks.

Beyond conventional recognition Traditional computer vision tasks focus on the recognition of tangible properties of images, such as objects (both entry-level [11, 51] and subordinate categories [80, 78, 33]) and scenes [95]. The research on representation learning mentioned above focuses on this type of task. Relatively little attention has been paid to tasks that involve private states [65, 82] where subjectivity analysis is relevant. This area includes (1) detecting cyberbullying and hate speech [29, 71, 20, 40], (2) identifying emotions [43, 1, 59, 62, 81], (3) understanding rhetoric and intentions [36, 37, 70, 30, 31, 88, 74, 45, 34]. The present work aims to advance research in this area by learning effective features from high-level engagement signals.

3. Approach

With the aim of learning representations that capture the relationships between image content and human responses, we introduce a simple yet effective framework, VisE. It infers visual engagement signals from images (Sec. 3.1). VisE is trained on large scale image-engagement pairs from a social media platform in a multi-task fashion, which will be described in Sec. 3.2 and Sec. 3.3.

3.1. From Engagement to Labels

How people react to images is more telling than the content itself. In this work, we propose to predict raw social engagement signals by converting them into a bag-of-word multi-label multi-task classification task. Fig. 2 illustrates the pipeline we use in this work.

More formally, let \((x, \{e\})\) be an image and a set of corresponding engagement clues (e.g., comments, replies, likes, etc.). Let \(\phi\) be a general engagement feature extractor that transforms \(e\) into a numerical representation \(\phi(e) \in \mathbb{R}^D\). We describe the proposed method to preprocess and obtain pseudo-label for \(e\) below.
Figure 2. Learning from visual engagement. (a) We learn visual representations by predicting the pseudo-labels derived from a set of engagement signals including comments and raw reactions. (b) The pseudo-label is the clustered assignment index that is computed by first transforming the raw engagement \(e\) to a numerical representation \(\phi(e)\). See the main text for more details.

Step 1 (Cluster Generation) We first collect \(e\) from a randomly sampled subset of the entire image-engagement pairs and generate \(k\) clusters using \(K\)-means algorithm. This portion of the dataset will not be used during pre-training.

Step 2 (Label Creation) Given the \(i^{th}\) example from the unprocessed subset, \((x_i, \{e_i\}_i)\), we obtain a set of features \(\{\phi(e)\}_i\) from the engagement set of this example, using the same function \(\phi\). Next we collect the resulting cluster assignments indices as a set of labels \(\{y_c\}_i\) for this image \(x_i\). Fig. 2(a) summarizes this step.

3.2. Engagement Types

We use textual comments \(C\) and raw reactions \(R\) from publicly available Facebook posts as a first step to investigate visual engagement signals for representation learning.

Comments Comments are direct human responses from image posts. Maximal 100 comments are randomly sampled from a post. We use the bag-of-word approach with the term frequency-inverse document frequency (TF-IDF) weighted document embeddings [68] as comment feature extractor \(\phi_C\). The cluster assignments derived from the comment set are used as multi-label classification targets, one label for each sampled comment associated with this specific image.

Raw reactions Reactions are encoded as a normalized distribution over the five reaction buttons (“haha”, “sorry”, “angry”, “wow”, “love”). More specifically, we count the total occurrences of the 5 reaction buttons for each image post and normalize (L2) them to account for differences in followers and popularity of the post. Each post is mapped to a single cluster centroid index.

3.3. Training VisE Models

VisE learning objective VisE is trained in a multi-task learning framework by minimizing the following loss function:

\[
\arg\min_W \mathbb{E}_{(x_i, \{y_c\}, \{y_R\}) \sim \mathcal{D}} \mathcal{L}_C(f(x_i), \{y_c\}; W) + \mathcal{L}_R(f(x_i), \{y_R\}; W),
\]

where \(\mathcal{L}_C\) is a cross-entropy loss with soft probability [55], \(\mathcal{L}_R\) is the standard cross-entropy loss, \(f\) represents an image feature encoder parameterized by \(W\). \(\mathcal{D}\) represents the training data which the image-engagement pairs are sampled from. In addition, \(\{y_C\}\) and \(\{y_R\}\) denote the pseudo labels of comments and reactions, respectively.

Pre-training details In our experiments, we train convolutional networks for engagement prediction. For comparison with previous work, the CNN model architectures we use are: ResNet-50, ResNet-101 [28], and ResNeXt-101 32 \(\times\) 16d [85]. VisE model with the largest backbone, ResNeXt-101 32 \(\times\) 16d, took around 10 days to train on 32 NVIDIA V100 GPUs with 250 million images with mini-batch size of 1536. Other pre-training details are included in the Appendix C.

4. Experiments

We evaluate the effectiveness of feature representations learned by VisE for a wide range of downstream tasks. In our experiment, we aim to show that image representations learned from engagement signals are beneficial for image classification that beyond the scope of conventional recognition tasks. We begin by describing our experimental setups, including a comparison of alternative representation learning methods (Sec. 4.1), implementation details (Sec. 4.2), and a summary of evaluated downstream tasks (Sec. 4.3). Finally we present results and discussion in Sec. 4.4 and Sec. 4.5.

4.1. Compared methods

To train VisE, we collect a total of 270 million public image posts from a social media platform with 20 million used for cluster generation (see Sec. 3.1). To facilitate a fair comparisons with the ImageNet-supervised method, we also randomly sample 1.23 million images for pre-training. We compare VisE pre-trained on 1.23 million (VisE-1.2M) and 250 million data (VisE-250M) with other feature representation learning methods.

Uni-modal learning methods We first compare VisE trained with pseudo labels derived from clustering assignments with networks that are trained with a pre-defined set of object labels:
Cross-modality pre-training: Learning methods that use natural language as supervisory signals are also considered:

- **ImageNet-supervised (IN-Sup):** the image encoder is pre-trained on ILSVRC 2012 train split (1.28M images). The dataset has 1000 classes, which is based on the concepts in WordNet [56].

- **IG-imagenet (IG-940M-IN) [55]:** the visual encoder is pre-trained on 940 million public images with 1.5K hashtags in weakly-supervised fashion; the encoder is further fine-tuned on ImageNet dataset.

- **VQAGrid [35]:** a pre-training method primarily for visual question answering and image captioning tasks. It learns visual representations by training a Faster-RCNN [67] on the Visual Genome dataset [44] which has 1600 object categories and 400 attributes. We use the outputs from the last bottleneck block of the ResNet-50 as the pre-trained image representations.

- **MoCo-v2 [27]:** a self-supervised contrastive method using a momentum-based encoder and a memory queue trained on ImageNet. Given an image sample, it is trained to be closer to its randomly augmented version on a hypersphere than other samples in the dataset. We use the improved version [7] that is trained with 800 epochs. Note that this method uses image as supervision labels instead of the ImageNet class labels.

- **VisE-250M 250M**

<table>
<thead>
<tr>
<th>Method</th>
<th>Input Type</th>
<th>Annotation Type</th>
<th>Noisy Labels</th>
<th>Pre-trained Data</th>
<th>Data Size</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Init</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>train from scratch</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IN-Sup</td>
<td>images</td>
<td>object labels</td>
<td>-</td>
<td>ImageNet [11]</td>
<td>1.28M</td>
<td>CNN</td>
</tr>
<tr>
<td>VQAGrid [35]</td>
<td>images</td>
<td>object + attribute</td>
<td></td>
<td>VisualGenome [44]</td>
<td>103k</td>
<td>Faster-RCNN</td>
</tr>
<tr>
<td>ours</td>
<td>images</td>
<td>pseudo labels</td>
<td>✓</td>
<td>VisE-12M</td>
<td>1.23M</td>
<td>CNN</td>
</tr>
</tbody>
</table>

Table 1. VisE vs. alternative methods compared. The seven representation learning approaches are grouped into: (1) Uni-modal pre-training: similar to VisE, these methods use visual encoders only during pre-training; (2) Cross-modal pre-training: like VisE, these approaches learn from natural language as inputs or supervisory signals. All three architectures involve a Transformer-based model [79] for the textual module. *We include the negative image-image / image-text pairs when counting the total data size for MoCo-v2 and CLIP. See Appendices for more details. We also acknowledge that the effective data size is also affected by other factors such as data augmentations in different approaches. For simplicity, we use the actual dataset size for non-contrastive learning methods.

- **VirTex [12]:** this method also pre-trains on COCO-captions but on a different task: generating captions based on images.

- **CLIP [66]:** Contrastive Language-Image Pre-training method (CLIP) utilizes an image encoder and a text encoder to predict which images are paired with which textual descriptions in a large-scale dataset of 400M image-text pairs.

It is worth pointing out all of these approaches train a text encoder to learn better natural language representations during pre-training stage, while VisE simply uses an off-the-shelf text encoder to compute representations for clustering purposes. We expect better performance of VisE if these textual representations are further fine-tuned [3] as these aforementioned methods.

We also report results of “Random Init”, where no pre-trained features are used. Table 1 summarizes the differences between VisE and all of the baseline methods used in the experiment.

4.2. Evaluation Protocols and Details

We adopt two common protocols for evaluating the effectiveness of feature representations [55, 21, 57, 27]: (1) Linear evaluation: all pre-trained models are used as visual feature extractors, where the weights of the image encoders are fixed. This protocol is preferred for applications where computational resources or training data for target tasks are limited. The test performance indicates how effective the learned representations are for specific tasks. (2) Fine-tuning: the parameters of the pre-trained image encoders are used as an advanced weight initialization method; these encoders are fine-tuned in an end-to-end manner for downstream tasks. Prior studies [18, 67] have shown that the latter protocol outperforms the linear evaluation approach due to its flexibility and adaptability to a wider range of downstream tasks.
See the Appendix C for more implementation details, including a full list of hyperparameters used (batch sizes, learning rates, decay schedules, etc.) and sensitivity to hyperparameters for both linear and fine-tuned experiments.

4.3. Downstream Tasks

We evaluate these visual representation methods on four downstream tasks, including sentiment classification, political bias, hate speech detection and fine-grained bird species classification.

**UnbiasedEmotion Dataset** This dataset [62] contains 3045 images annotated into six emotional categories. To reduce object biases in the dataset, different emotion labels contain the same set of objects/scenes. Since there is no official split of this dataset, we random split the images into train (70%), val (10%), test (20%) set five times and report mean and standard deviation of the resulting accuracy.

**Politics** The task of this dataset [75] is to predict the political leaning (left and right) of images from news media. This dataset contains 749,932 images in total. Since only train and test set are publicly available, we randomly split the training set into train (90%) and val (10%) and report accuracy scores.

**Hateful Memes** Hateful Memes dataset [40] contains multimodal memes including images and text. The task is to detect each meme is hate speech or not. We use the data from Phase 1 of the Hateful Memes challenge, which has 8500 training and 500 validation data. We obtain the sentence embeddings from a pre-trained RoBERTa model [52], and concatenate the text features and image features together before linear evaluation that map the feature to the label space. We report the macro averaged ROC AUC score and accuracy score on the val set.

**Caltech-UCSD Birds-200-2011 (CUB-200-2011)** In addition to the above subjective classification tasks, we also evaluate our approach on standard image classification tasks. To this end, we use the CUB-200-2011 dataset. CUB-200-2011 [80] has a total of 11,788 images allocated over 200 (mostly North American) bird species. It is a benchmarking dataset for subordinate categorization. We train on the publicly available train set and report top-1 accuracy on the val set.

4.4. VisE for Subjective Recognition Tasks

Following the two protocols described in Sec. 4.2, we compare the transfer-learning capabilities of VisE with other baseline approaches for three subjective tasks. Fig. 3 presents the results for the linear evaluation and fine-tuned protocols, respectively.

**VisE vs. other uni-modal methods** From Fig. 3, we can see that: (1) VisE is consistently better than other image-encoder only methods on three datasets over two visual backbone choices, except for MoCo-v2 in Politics. Even pre-trained with similar amounts of data (1.28M vs. 1.23M), VisE-1.2M still achieves better performance across all three tasks with a ResNet-50 backbone than models trained with labels from In-Sup. (2) VisE-250M substantially outperforms IG-940M-IN, a method trained with a substantially larger amount of pre-trained data (950M vs. 250M). (3) MoCo-v2, a self-supervised approach that does not require object category annotations, yields the best accuracy scores on Politics among approaches with ResNet-50 backbone. This also highlights the limitation of using object labels as pre-training supervision for such subjective tasks.

**VisE vs. other methods that learn from language** Under both protocols, VisE offers better or comparable results than other methods that leverage textual information during pre-training. VisE achieves consistently better performance across all three tasks. This suggests that features learned with visual engagement signals are more suitable for subjective downstream tasks. We also observe that all other four visual-language approaches obtain better results than IN-Sup on UnbiasedEmotion dataset when fine-tuned, but they are worse than IN-Sup with linear evaluation. Such discrepancy might be caused by the scale of the dataset, since UnbiasedEmotion is the smallest among the other tasks.

4.5. VisE for Standard Recognition Tasks

We compare the effectiveness of features learned from visual engagement signals with those trained using ImageNet labels on CUB-200-2011. This task extends the general object classification from ImageNet and focuses on distinguishing fine-grained differences among 200 bird species. Moreover, 59 out of 1000 classes in ImageNet are already bird categories, including overlapping definitions with CUB-200-2011 [77]. Thus ImageNet based approaches should transfer to this task better than VisE. Table 2 shows the results of both linear evaluation and fine-tuned transfer protocols.

Indeed, IN-Sup and IG-940M-IN achieve decent accuracy scores using linear classifiers alone to map the features to 200 bird species, outperforming VisE by a large margin. This is foreseeable since visual engagement signals do not necessarily contain object information. It is understandable that VisE features are not as transferable to this task as models trained on ImageNet.

When fine-tuning is performed, VisE with ResNet-50 have comparable or better performance than IN-Sup. This highlights that fine-tuning the whole network can sometimes compensate for the inflexibility of learned features, which is in line with discussions in [66].

---

3 Hateful Memes: Phase 1 Challenge
Figure 3. Linear evaluation (top) and fine-tuned (bottom) results among VisE, uni-modal (ResNet-50), cross-modal (ResNet-50), and uni-modal (ResNeXt-101 32×16d) methods for all three datasets. For easy comparison, two blue dashed lines represent the performance of VisE-1.2M and VisE-250M respectively.

<table>
<thead>
<tr>
<th>Backbone Method</th>
<th>Linear Evaluation</th>
<th>Fine-tuned Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Init</td>
<td>3.42</td>
<td>62.70</td>
</tr>
<tr>
<td>VisE-1.2M</td>
<td>9.95</td>
<td>72.65</td>
</tr>
<tr>
<td>VisE-250M</td>
<td>9.92</td>
<td>76.58</td>
</tr>
<tr>
<td>VisE-940M-IN</td>
<td>7.04</td>
<td>62.80</td>
</tr>
<tr>
<td>VisE-1.2M</td>
<td>9.76</td>
<td>73.73</td>
</tr>
<tr>
<td>VisE-250M</td>
<td>10.93</td>
<td>79.54</td>
</tr>
<tr>
<td>VisE-1.2M</td>
<td>58.31</td>
<td>65.79</td>
</tr>
<tr>
<td>VisE-250M</td>
<td>58.39</td>
<td>64.44</td>
</tr>
<tr>
<td>VisE-940M-IN</td>
<td>58.31</td>
<td>65.79</td>
</tr>
<tr>
<td>VisE-1.2M</td>
<td>64.25</td>
<td>67.56</td>
</tr>
<tr>
<td>VisE-250M</td>
<td>63.54</td>
<td>67.56</td>
</tr>
<tr>
<td>VisE-1.2M</td>
<td>59.45</td>
<td>63.64</td>
</tr>
<tr>
<td>VisE-250M</td>
<td>59.42</td>
<td>63.64</td>
</tr>
<tr>
<td>VisE-940M-IN</td>
<td>59.45</td>
<td>63.64</td>
</tr>
<tr>
<td>VisE-1.2M</td>
<td>58.83</td>
<td>65.58</td>
</tr>
<tr>
<td>VisE-250M</td>
<td>58.93</td>
<td>66.41</td>
</tr>
<tr>
<td>VisE-940M-IN</td>
<td>58.86</td>
<td>66.41</td>
</tr>
<tr>
<td>VisE-1.2M</td>
<td>57.31</td>
<td>62.15</td>
</tr>
<tr>
<td>VisE-250M</td>
<td>59.31</td>
<td>64.44</td>
</tr>
<tr>
<td>VisE-940M-IN</td>
<td>57.31</td>
<td>64.44</td>
</tr>
<tr>
<td>VisE-1.2M</td>
<td>0.5335</td>
<td>0.5691</td>
</tr>
<tr>
<td>VisE-250M</td>
<td>0.5517</td>
<td>0.5784</td>
</tr>
<tr>
<td>VisE-940M-IN</td>
<td>0.5659</td>
<td>0.5702</td>
</tr>
<tr>
<td>VisE-1.2M</td>
<td>0.5631</td>
<td>0.5621</td>
</tr>
<tr>
<td>VisE-250M</td>
<td>0.5631</td>
<td>0.5631</td>
</tr>
<tr>
<td>VisE-940M-IN</td>
<td>0.5631</td>
<td>0.5631</td>
</tr>
<tr>
<td>VisE-1.2M</td>
<td>0.5642</td>
<td>0.5642</td>
</tr>
<tr>
<td>VisE-250M</td>
<td>0.5642</td>
<td>0.5642</td>
</tr>
<tr>
<td>VisE-940M-IN</td>
<td>0.5642</td>
<td>0.5642</td>
</tr>
</tbody>
</table>

Table 2. Validation accuracy on CUB-200-2011. Features learnt from Image-Net class labels transfer well to CUB-200-2011.

5. Analysis

To better understand the values of visual engagement signals and our pre-trained VisE models, we conduct ablation studies and qualitative analysis using the same set of subjective target tasks. All experiments use the fine-tuned protocol unless otherwise specified. Additional results and analysis are included in the Appendix B.

5.1. Pre-training ablation

Effect of pre-training data size. We additionally train VisE models using randomly sampled \{123k, 308k, 615k\} images using fine-tuned setting. Fig. 4 presents the results for three datasets compared with other baselines for easy reference. From Fig. 4(a) and 4(b), we can see that there is a positive correlation between training data size and model performance for VisE. VirTex and ICMLM achieve similar results as VisE with less training data (118k vs. \{615k, 1.23M\}). This shows the noisy pseudo-labels from visual engagement might need more data to compensate its weakly-supervised nature.

For a multi-modal dataset like Hateful Memes, the correlation between data size and performance is less clear than the other two tasks, as shown in Fig. 4(c). When fine-tuned, VisE is able to perform better than other baselines, demonstrating the effectiveness of visual engagement signals.
(a) UnbiasedEmotion. (b) Politics. (c) Hateful Memes.

Figure 4. Data size ablation using ResNet-50 (top) and ResNeXt-101 32×16d (bottom) backbones. We only present one of ICMLM_{att-fc} and ICMLM_{tfm} due to space constraint. The differences between both ICMLM methods are within 1% with each other.

<table>
<thead>
<tr>
<th>Backbone Method</th>
<th>UnbiasedEmotion</th>
<th>Politics</th>
<th>Hateful Memes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear Fine-tuned</td>
<td>45.74 ± 2.15</td>
<td>74.20 ± 1.03</td>
<td>64.100</td>
</tr>
<tr>
<td>Comments</td>
<td>49.15 ± 1.30</td>
<td>72.16 ± 1.09</td>
<td>64.005</td>
</tr>
<tr>
<td>Reaction</td>
<td>33.05 ± 1.75</td>
<td>70.03 ± 2.64</td>
<td>63.052</td>
</tr>
</tbody>
</table>

Table 3. Task ablation of VisE-1.2M with ResNet-50. Colored text with ↑ and ↓ indicate the differences with results from the VisE model trained with both reaction and comments.

Effect of pre-training tasks VisE is trained using cluster assignments from both comments and raw reactions. Table 3 shows the ablation study where we evaluate the contribution of different engagement formats. In general, VisE obtains the best performance when trained with a multi-task objective.

Effect of pre-training visual backbone Fig. 5 presents ablation studies using different visual backbones. VisE-1.2M is better than Random Init and IN-Sup across all three backbone choices and three downstream tasks. We also note that the advantages of VisE is diminishing as the number of parameters of backbone getting larger possibly due to overfitting on the downstream training set.

5.2. Multimodal fine-tuning ablation

We used an image encoder and a frozen textual encoder for experiments on Hateful Memes dataset in Sec 4. To isolate the effects of both module for this dataset, we compare VisE with other baselines under the following setups: (1) Image + Text (Fine-tuned): parameters from both encoders are updated during transfer learning. (2) Image + Text-Frozen (Fine-tuned): the same fine-tuned setting as the experiment in Sec 4. (3) Image Only (Fine-tuned): we only use image encoder in fine-tuned setting. (4) Image Only (Linear): image encoder is used as feature extractor only. Note the linear evaluation of Hateful Memes in Sec. 4.4 (Fig. 3(c)) uses concatenated features from both encoders. Fig. 6 presents the results.

Methods with VisE are able to achieve better or comparable results compared with other baselines. And visual-language methods outperform IN-Sup in Fig. 6(a) except for CLIP. This shows the visual backbones that are pre-trained with a textual module are better when both modules are fine-tuned together.
Figure 6. Multi-modal ablation for Hateful Memes val split. For reference, using text encoder alone give ROC AUC scores: 0.6363 (fine-tuned), 0.5983 (linear).

Figure 7. Qualitative results on UnbiasedEmotion dataset using ResNeXt-101 $32 \times 16d$.

From Figs. 6(a)-6(c), the ROC AUC scores are getting smaller and smaller as the textual encoder makes less and less contribution. This demonstrates that the text module has dominant effect on predicting if a multi-modal meme is hateful or not. If using image encoders only, linear evaluation obtains better results than the fine-tuned protocol (Fig. 6(c) vs. 6(d)). This also suggests textual information is more important on this dataset.

5.3. Qualitative Analysis

We also conduct qualitative studies using UnbiasedEmotion to further understand the benefit of visual engagement. Fig. 7 shows sample predictions produced by our VisE-250M and 3 other approaches. The predicted classes for each approach are color coded (green as correct, red as incorrect). We also use class activation mappings [94] to visualize the discriminate image regions for the predicted emotion. Although all methods that are initialized with pre-trained models can detect the object of interest in the image, IN-Sup and IG-940M-IN are more likely to predict “joy” and “love” for the cats and dogs photos in Fig. 7, while VisE-250M yields more diverse predictions (see row 2 left vs. row1 right as an example). This seems to suggest that IN-Sup and IG-940M-IN map dogs and cat to positive emotions. VisE-250M, on the other hand, does not rely on the ImageNet object labels during pre-training and is able to distinguish the subtle emotional differences among different images with dogs and cats. However, all methods failed to infer “surprise” from the bottom left image, possibly due to imbalanced training data of UnbiasedEmotion. Additional visualization can be found in the Appendix A.

6. Conclusion

We explored social media visual engagement as supervisory signals for representation learning. We presented VisE, a streamlined pre-training method that uses pseudo-labels derived from human responses to social media posts, including reactions and comments. Experiments and analysis show that visual engagement signals transfer well to various downstream tasks that go beyond conventional visual recognition. VisE is able to outperform various representation learning models on these datasets. We therefore hope that VisE could inspire and facilitate future research that focuses on the cognitive aspects of images. Pre-trained models will be released upon acceptance of the work.

Acknowledgement We thank Marseille who is featured in Figs. 1 and 2. This work is supported by a Facebook AI research grant awarded to Cornell University.
References


using crowdsourced dense image annotations. Li Fei-Fei. Visual genome: Connecting language and vision.


Lapedriza. Emotion recognition in context. In ECCV.

Big transfer (bit): General visual representation learning. In NeurIPS.


[72] Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. Grounded compositional semantics for finding and describing images with sentences. Transactions of the Association for Computational Linguistics, 2014. 2


[74] Christopher Thomas and Adriana Kovashka. Predicting the politics of an image using webly supervised data. In NeurIPS, pages 3625–3637, 2019. 2


[76] Burt Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. Communications of the ACM, 2016. 2


[78] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In CVPR, 2018. 2


[81] Zijun Wei, Jianming Zhang, Zhe Lin, Joon-Young Lee, Niranjan Balasubramanian, Minh Hoai, and Dimitris Samaras. Learning visual emotion representations from web data. In CVPR, June 2020. 2


[84] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In CVPR, 2020. 1, 2


[93] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of
medical visual representations from paired images and text. 


[96] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *ICCV*, 2019. 2