We Are Humor Beings: Understanding and Predicting Visual Humor

Arjun Chandrasekaran[1] Ashwin K. Vijayakumar[1] Stanislaw Antol[1] Mohit Bansal[2]
Dhruv Batra[1] C. Lawrence Zitnick[3] Devi Parikh[1]
[1]Virginia Tech [2]TTI-Chicago [3]Facebook AI Research
[1]{carjun, ashwinkv, santol, dbatra, parikh}@vt.edu [2]mbansal@ttic.edu [3]zitnick@fb.com

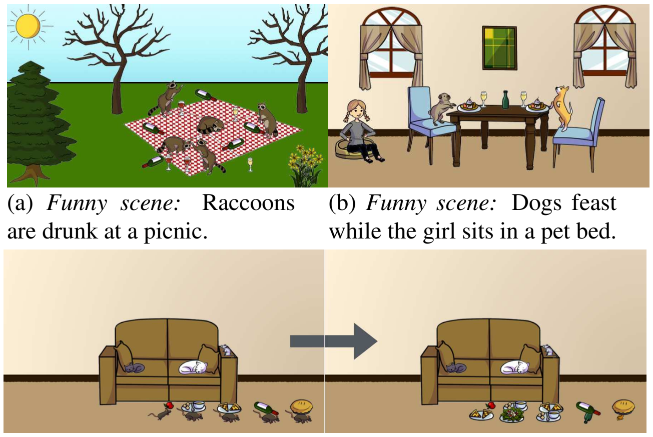
Abstract

Humor is an integral part of human lives. Despite being tremendously impactful, it is perhaps surprising that we do not have a detailed understanding of humor yet. As interactions between humans and AI systems increase, it is imperative that these systems are taught to understand subtleties of human expressions such as humor. In this work, we are interested in the question – what content in a scene causes it to be funny? As a first step towards understanding visual humor, we analyze the humor manifested in abstract scenes and design computational models for them. We collect two datasets of abstract scenes that facilitate the study of humor at both the scene-level and the object-level. We analyze the funny scenes and explore the different types of humor depicted in them via human studies. We model two tasks that we believe demonstrate an understanding of some aspects of visual humor. The tasks involve predicting the funniness of a scene and altering the funniness of a scene. We show that our models perform well quantitatively, and qualitatively through human studies. Our datasets are publicly available.

1. Introduction

An adult laughs 18 times a day [25] on average. A good sense of humor is related to communication competence [13, 14], helps raise an individual's social status [43], popularity [17, 26], and helps attract compatible mates [8, 10, 35]. Humor in the workplace improves camaraderie and helps workers cope with daily stresses [38] and loneliness [52]. *fMRI* [40] studies of the brain reveal that humor activates the components of the brain that are involved in reward processing [53]. This probably explains why we actively seek to experience and create humor [33].

Despite the tremendous impact that humor has on our lives, the lack of a rigorous definition of humor has hindered humor-related research in the past [4, 46]. While verbal humor is better understood today [41, 44], visual humor remains unexplored. As vision and AI researchers we are interested in the following question – what content in an image causes it to be funny? Our work takes a step in the direction of building computational models for visual humor. Computational visual humor is useful for a number of applications: to create better photo editing tools, smart cameras that pick the right moment to take a (funny) picture, recommendation tools that rate funny pictures higher (say, to post on social media), video summarization tools that summarize only the funny frames, automatically generating funny scenes for entertainment, identifying and catering to personalized humor, *etc*. As AI systems interact more with humans, it is vital that they understand subtleties of human emotions and expressions. In that sense, being able to identify humor can contribute to their *common sense*.



(a) *Funny scene:* Raccoons are drunk at a picnic.

(b) *Funny scene:* Dogs feast while the girl sits in a pet bed.

(c) *Funny scene:* Rats steal food while the cats are asleep.

(d) *Funny Object Replaced (unfunny) counterpart:* Rats in *(c)* are replaced by food.

Figure 1: (a), (b) are selected funny scenes in the Abstract Visual Humor dataset. (c) is an originally funny scene in the Funny Object Replaced dataset. The objects contributing to humor in (c) are replaced by a human with other objects, to create an unfunny counterpart.

Understanding visual humor is fraught with challenges such as having to detect all objects in the scene, observing the interactions between objects, and understanding context, which are currently unsolved problems. In this work, we argue that, by using scenes made from clipart [1, 2, 15, 22, 23, 50, 57, 58], we can study visual humor without having to wait for these detailed recognition problems to be solved. Abstract scenes are inherently densely annotated (*e.g*. all objects and their locations are known), and so enable us to learn fine-grained semantics of a scene that causes it to be funny. In this paper, we collect two datasets of abstract scenes that facilitate the study of humor at both the scene-level (Fig. 1a, Fig. 1b) and the object-level (Fig. 1c, Fig. 1d). We propose a model that predicts how funny a scene is using semantic visual features of the scene such as occurrence of objects, and their relative locations. We also build computational models for a particular source of humor, *i.e*., humor due to the presence of objects in an unusual context. This source of humor is explained by the *incongruity theory* of humor which states that a playful violation of the subjective expectations of a perceiver causes humor [28]. *E.g*., Fig. 1b is funny because our expectation is that people eat at tables and dogs sit in pet beds and this is violated when we see the roles of people and dogs swapped.

The scene-level Abstract Visual Humor (AVH) dataset contains funny scenes (Fig. 1a, Fig. 1b) and unfunny scenes with human ratings for funniness of each scene. Using the ground truth rating, we demonstrate that we can reliably predict a *funniness score* for a given scene. The object-level Funny Object Replaced (FOR) dataset contains scenes that are originally funny (Fig. 1c) and their unfunny counterparts (Fig. 1d). The unfunny counterparts are created by humans by replacing objects that contribute to humor such that the scene is not funny anymore. The ground truth of replaced objects is used to train models to alter the funniness of a scene – to make a funny scene unfunny and vice versa. Our models outperform natural baselines and ablated versions of our system in quantitative evaluation. They also demonstrate good qualitative performance via human studies.

Our main contributions are as follows:

1. We collect two abstract scene datasets consisting of scenes created by humans which are publicly available.
   i. The scene-level Abstract Visual Humor (AVH) dataset consists of funny and unfunny abstract scenes (Sec. 3.2). Each scene also contains a brief explanation of the humor in the scene.
   ii. The object-level Funny Object Replaced (FOR) dataset consists of funny scenes and their corresponding unfunny counterparts resulting from object replacement (Sec. 3.3).
2. We analyze the different sources of humor techniques depicted in the AVH dataset via human studies (Sec. 3.2).
3. We learn distributed representations for each object category which encode the context in which an object naturally appears, *i.e*., in an unfunny setting. (Sec. 4.1).
4. We model two tasks to demonstrate an understanding of visual humor:
   i. Predicting how funny a given scene is (Sec. 5.1).
   ii. Automatically altering the funniness of a given scene (Sec. 5.2).

To the best of our knowledge, this is the first work that deals with understanding and building computational models for visual humor.

2. Related Work

Humor Theories. Humor has been a topic of study since the time of Plato [37], Aristotle [3] and Bharata [5]. Over the years, philosophical studies and psychological research have sought to explain why we laugh. There are three theories of humor [55] that are popular in contemporary academic literature. According to the incongruity theory, a perceiver encounters an incongruity when expectations about the stimulus are violated [24]. The two stage model of humor [48] further states that the process of discarding prior assumptions and reinterpreting the incongruity in a new context (resolution) is crucial to the comprehension of humor. Superiority theory suggests that the misfortunes of others which reflects our own superiority is a source of humor [34]. According to the relief theory, humor is the release of pent-up tension or mental energy. Feelings of hostility, aggression, or sexuality that are expressed bypassing any societal norms are said to be enjoyed [16].

Previous attempts to characterize the stimuli that induce humor have mostly dealt with linguistic or verbal humor [28] *e.g*., script-based semantic theory of humor [44] and its revised version, the general theory of verbal humor [41].

Computational Models of Humor. A number of computational models are developed to recognize language-based humor *e.g*., one-liners [30], sarcasm [11] and *knock-knock* jokes [49]. Other work in this area includes exploring features of humorous texts that help detection of humor [29], and identifying the set of words or phrases in a sentence that could contribute to humor [56].

Some computational humor models that generate verbal humor are JAPE [7] which is a pun-based riddle generating program, HAHAcronym [47] which is an automatic funny acronym generator, and an unsupervised model that produces "*I like my X like I like my Y, Z*" jokes [36]. While the above works investigate detection and generation of verbal humor, in this work we deal purely with *visual* humor.

Recent works predict the best text to go along with a given (presumably funny) raw image such as a meme [51] or a cartoon [45]. In addition, Radev *et al*. [39] develop unsupervised methods to rank funniness of captions for a cartoon. They also analyze the characteristics of the funniest captions. Unlike our work, these works do not predict whether a *scene* is funny or which components of the scene contribute to the humor.

Buijzen and Valkenburg [9] analyze humorous commercials to develop and investigate a typology of humor. Our contributions are different as we study the sources of humor in static images, as opposed to audiovisual media. To the best of our knowledge, ours is the first work to study *visual* humor in a computational framework.

Human Perception of Images. A number of works investigate the intrinsic characteristics of an image that influence human perception *e.g.*, memorability [20], popularity [21], visual interestingness [18], and virality [12]. In this work, we study what content in a scene causes people to perceive it as funny, and explore a method of altering the funniness of a scene.

Learning from Visual Abstraction. Visual abstractions have been used to explore high-level semantic scene understanding tasks like identifying visual features that are semantically important [57, 59], learning mappings between visual features and text [58], learning visually grounded word embeddings [22], modeling fine-grained interactions between pairs of people [2], and learning (temporal and static) common sense [15, 23, 50]. In this work, we use abstract scenes to understand the semantics in a scene that cause humor, a problem that has not been studied before.

3. Datasets

We introduce two new abstract scenes datasets – the Abstract Visual Humor (AVH) dataset (Sec. 3.2) and the Funny Object Replaced (FOR) dataset (Sec. 3.3) using the interfaces described in Sec. 3.1. The AVH dataset (Sec. 3.2) consists of both funny and unfunny scenes along with funniness ratings. The FOR dataset (Sec. 3.3) consists of funny scenes and their altered unfunny counterparts. Both the datasets are made publicly available on the project webpage.

3.1. Abstract Scenes Interface

Abstract scenes enable researchers to explore high-level semantics of a scene without waiting for low-level recognition tasks to be solved. We use the clipart interface[1] developed by Antol *et al*. [1] which allows for indoor and outdoor scenes to be created. The clipart vocabulary consists of 20 deformable human models, 31 animals in various poses, and around 100 objects that are found in indoor (*e.g*., chair, table, sofa, fireplace, notebook, painting) and outdoor (*e.g*., sun, cloud, tree, grill, campfire, slide) scenes. The human models span different genders, races, and ages with 8 different expressions. They have limbs that are adjustable to allow for continuous pose variations. This combined with the large vocabulary of objects result in diverse scenes with rich semantics. Fig. 1 (*Top Row*) shows scenes that AMT workers created using this abstract scenes interface and vocabulary. Additional details, example scenes, and a sample of clipart objects are available on the project webpage.

3.2. Abstract Visual Humor (AVH) Dataset

This dataset consists of funny and unfunny scenes created by AMT workers, facilitating the study of visual humor at the scene level.

Collecting Funny Scenes. We collect 3.2K scenes via AMT by asking workers to create funny scenes that are meaningful, realistic, and that other people would also consider funny. This is to encourage workers to refrain from creating scenes with inside jokes or catering to a very personalized form of humor. A screenshot of the interface used to collect the data is available on the project webpage. We provide a random subset of the clipart vocabulary to each worker out of which at least 6 clipart objects are to be used to create a scene. In addition, we also ask the worker to give a brief description of why the scene is funny in a short phrase or sentence. We find that this encourages workers to be more thoughtful and detailed regarding the scene they create. Note that this is different from providing a caption to an image since this is a simple explanation of what the worker had in mind while creating the scene. Mining this data may be useful to better understand visual humor. However, in this work we focus on the harder task of understanding purely *visual* humor and do not use these explanations.

We also use an equal number (3.2K) of abstract scenes from [1] which are realistic, everyday scenes. We expect most of these scenes to be mundane (*i.e*., not funny).

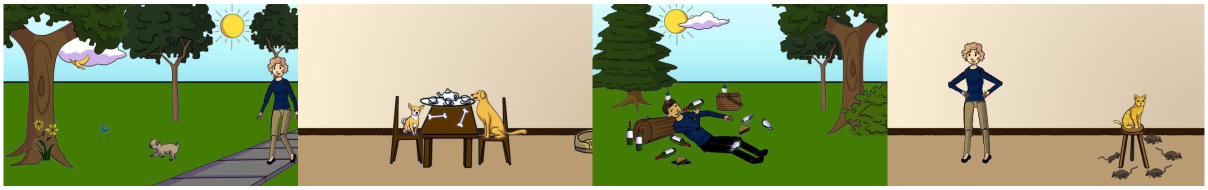
Labeling Scene Funniness. Anyone who has tried to be funny knows that humor is a subjective notion. A well-intending worker may create a scene that other people do not find very funny. We obtain funniness ratings for each scene in the dataset from 10 different workers on AMT who do not see the creator's explanation of funniness. The ratings are on a scale of 1 to 5, where 1 is not funny and 5 is extremely funny. We define the *funniness score* F_i of a scene i, as the average of the 10 ratings for the scene. We found 10 ratings to be sufficient for good inter-human agreement. Further analysis is provided on the project webpage.

By plotting a distribution of these scores, we determine the optimal threshold that best separates scenes that were intended to be funny (*i.e*., workers were specifically asked to create a funny scene) and other scenes (*i.e*., everyday scenes from [1], where workers were not asked to create funny scenes). We label all scenes that have a $F_i \geqslant$ threshold as *funny* and all scenes with a lower F_i as *unfunny*. This re-labeling results in 522 *unintentionally funny* scenes (i.e., scenes from [1], which were determined to be funny), and 682 *unintentionally unfunny* scenes (*i.e*., well-intentioned worker outputs which were deemed not funny by the crowd).

In total, this dataset contains 6,400 scenes (3,028 funny scenes and 3,372 unfunny scenes). We randomly split these scenes into train, val, and test sets having 60%, 20%, and 20% of the scenes, respectively. We refer to this dataset as the AVH dataset.

Humor Techniques. To better understand the different sources of humor in our dataset, we collect human annotations of the different techniques are used to depict humor in each scene. We create a list of humor techniques that

[1] www.github.com/VT-vision-lab/abstract_scenes_v002



(a) 0.1 (b) 1.5 (c) 4.0 (d) 4.0

Figure 2: Spectrum of scenes *(left to right)* in ascending order of funniness score, F_i (Sec. 3.2) as rated by AMT workers.

are motivated by existing humor theories, based on patterns that we observe in funny scenes, and the audio-visual humor typology by Buijzen *et al*. [9]: *person doing something unusual*, *animal doing something unusual*, *clownish behavior (*i.e.*, goofiness)*, *too many objects*, *somebody getting hurt*, *somebody getting scared* and *somebody getting angry*.

We choose a subset of 200 funny scenes from the AVH dataset. We show each of these scenes to 10 different AMT workers and ask them to choose all the humor techniques that are depicted. Our options also included *none of the above reasons*, which also prompted workers to briefly explain what other unlisted technique depicted in the scene made it funny. However, we observe that this option was rarely used by workers. This may indicate that most of our scenes can be explained well by one of the listed humor techniques. Fig. 3 shows the top voted images corresponding to the 4 most popular techniques of humor. We find that the techniques that involve animate objects – *animal doing something unusual* and *person doing something unusual* are voted higher than any other technique by a large margin. For 75% of the scenes, at least 3 out of 10 workers picked one of these two techniques. We observe that this *unusualness* or *incongruity* is generally caused by objects occurring in an unusual context in the scene.

Introducing or eliminating incongruities can alter the funniness of a scene. An elderly person kicking a football while simultaneously skateboarding (Fig. 4, *bottom*) is incongruous and hence considered funny. However, when the person is replaced by a young girl, this is is not incongruous and hence not funny. Such incongruities that can alter the funniness of a scene serves as our motivation to collect the Funny Object Replaced dataset which we describe next.

3.3. Funny Object Replaced (FOR) Dataset

Replacing objects in a scene is a technique to manipulate incongruities (and hence funniness) in a scene. For instance, we can change funny interactions (which are unexpected by our common sense) to interactions that are *normal* according to our mental model of the world. We use this technique to collect a dataset which consists of funny scenes and their altered unfunny counterparts. This enables the study of humor in a scene at the *object-level*.

We show funny scenes from the AVH dataset and ask AMT workers to make the least number of replacements in the scene to render the originally funny scene unfunny. The motivation behind this is to get a precise signal of which objects in the scene contribute to humor and what they can be replaced with to reduce/eliminate humor, while keeping the underlying structure of the scene the same. We ask workers to replace an object with another object that is as similar as possible to the first object and keep the scene realistic. This helps us understand fine-grained semantics that causes a specific object category to contribute to humor. There could be other ways to manipulate humor, *e.g*., by adding, removing, or moving objects in a scene, *etc*. but in our work we employ only the technique of replacing objects. We find that this technique is very effective in altering the funniness of a scene. Our interface did not allow people to add, remove, or move the objects in the scene. A screenshot of the interface used to collect this dataset is available on the project webpage.

For each of the 3,028 funny scenes in the AVH dataset, we collect *object-replaced* scenes from 5 different workers resulting in 15,140 unfunny counterpart scenes. As a sanity check, we collect funniness ratings (via AMT) for 750 unfunny counterpart scenes. We observe that they indeed have an average F_i of 1.10, which is smaller than that of their corresponding original funny scenes (whose average F_i is 2.66). Fig. 4 shows two pairs of funny scenes and their object-replaced unfunny counterparts. We refer to this dataset as the FOR dataset.

Given the task posed to workers (altering a funny scene to make it unfunny), it is natural to use this dataset to train a model to reduce the humor in a scene. However, this dataset can also be used to train flipped models that can increase the humor in a scene as shown in Sec. 5.2.3.

4. Approach

We propose and model two tasks that we believe demonstrate an understanding of some aspects of visual humor:

1. Predicting how funny a given scene is.
2. Altering the funniness of a scene.

The models that perform the above tasks are described in Sec. 4.2 and Sec. 4.3, respectively. The features used in the models are described first (Sec. 4.1).

4.1. Features

Abstract scenes are trivially densely annotated which we use to compute rich semantic features. Recall that our interface allows two types of scenes (indoor and outdoor) and

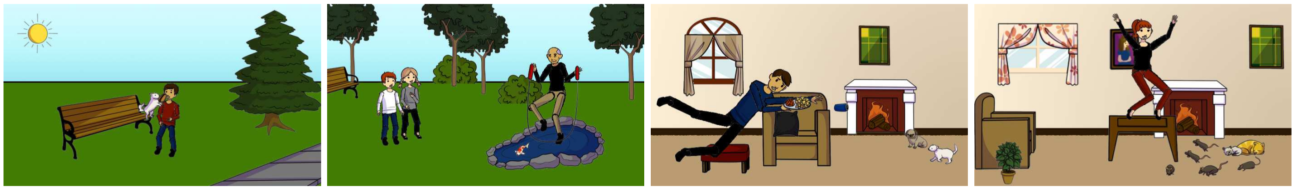


Figure 3: Top voted scenes by humor technique (Sec. 3.2). From *left* to *right*: *animal doing something unusual*, *person doing something unusual*, *somebody getting hurt*, and *somebody getting scared*.

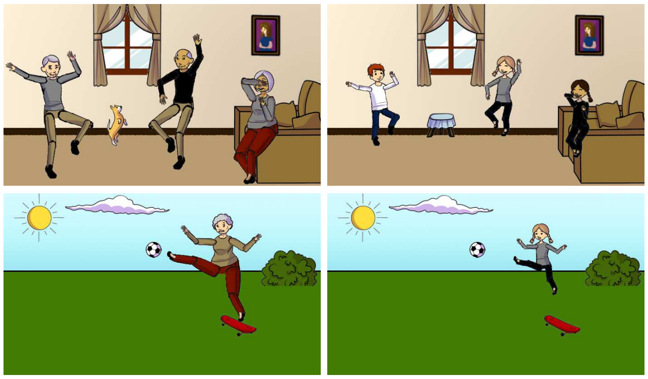


Figure 4: Funny scenes (*left*) and *one* among the 5 corresponding object-replaced unfunny counterparts (*right*) from the FOR dataset (see Sec. 3.3). For each funny scene, we collect an unfunny counterpart from a different worker.

our vocabulary consists of 150 object categories. We compute both scene-level and instance-level features.

1. **Instance-Level Features**

(a) **Object embedding (150-d)** is a distributed representation that captures the context in which an object category usually occurs. We learn this representation using a word2vec-style continuous Bag-of-Words model [32]. The model tries to predict the presence of an object category in the scene, given the context provided by other instances of objects in the scene. Specifically, in a scene, given 5 (randomly chosen) instances, the model tries to predict the object category of the 6th instance. We train the single-layer (150-d) neural network [31] with multiple 6-item subsets of instances from each scene. The network is trained using Stochastic Gradient Descent (SGD) with a momentum of 0.9. We use 11K scenes (that were not intended to be funny) from the dataset collected in [1] to train the model. Thus, we learn representations of objects occurring in natural contexts which are not funny. A visualization of the object embeddings is available on the project webpage.

(b) **Local embedding (150-d)** For each instantiation of an object in the scene, we compute a weighted sum of object embeddings of all the other instances in the scene. The weight of every other instance is its inverse square-root distance w.r.t. the instance under consideration.

2. **Scene-Level Features**

(a) **Cardinality (150-d)** is a Bag-of-Words representation that indicates the number of instances of each object category that are present in the scene.

(b) **Location (300-d)** is a vector of the horizontal and vertical coordinates of every object in the scene. When multiple instances of an object category are present, we consider location of the instance closest to the center of the scene.

(c) **Scene Embedding (150-d)** is the sum of object embeddings of all objects present in the scene.

4.2. Predicting Funniness Score

We train a Support Vector Regressor (SVR) that predicts the funniness score, F_i for a given scene i. The model regresses to the F_i computed from ratings given by AMT workers (described in Sec. 3.2) on scenes from the AVH dataset (Sec. 3.2). We train the SVR on the scene-level features (described in Sec. 4.1) and perform an ablation study.

4.3. Altering Funniness of a Scene

We learn models to alter the funniness of a scene – from funny to unfunny and *vice versa*. Our two-stage pipeline involves:

1. Detecting objects that contribute to humor.
2. Identifying suitable replacement objects from 1. to make the scene unfunny (or funny), while keeping it realistic.

Detecting Humor. We train a multi-layer perceptron (MLP) on scenes from the FOR dataset to make a binary prediction on each object instance in the scene – whether it should be replaced to alter the funniness of a scene or not. The input is a 300-d vector formed by concatenating object embedding and local embedding features. The MLP has two hidden layers comprising of 300 and 100 units respectively, to which ReLU activation is applied. The final layer has 2 neurons and is used to perform binary classification (replace or not) using cross-entropy loss. We train the model using SGD with a base learning rate of 0.01 and momentum of 0.9. We also trained a model with skip-connections that considers the predictions made on other objects when making a prediction on a given object. However, this did not result in significant performance gains.

Altering Humor. We train an MLP to perform a 150-way classification to predict potential replacer objects (from the clipart vocabulary), given an object predicted to be replaced in a scene. The model's input is a 300-d vector formed by concatenating local embedding and object embedding features. The classifier has 3 hidden layers of 300 units each, with ReLU non-linearities. The output layer has 150 units over which we compute soft-max loss. We train the model using SGD with a base learning rate of 0.1, momentum of 0.9, and a dropout ratio of 0.5. The label for an instance is the index of the replacer object category used by the worker. Due to the large diversity of viable replacer objects that can alter humor in a scene, we also analyze the top-5 predictions of this model. We train two models – one on funny scenes, and another on their unfunny counterparts from the FOR dataset. Thus, we learn models to alter the funniness in a scene in one direction – funny to unfunny or vice versa. Although we could train the pipeline end-to-end, we train each stage separately so that we can evaluate them separately and isolate their errors (for better interpretability).

5. Results

We discuss the performance of our models in the two visual humor tasks of:

1. Predicting how funny a given scene is (Sec. 5.1)
2. Altering funniness of a scene (Sec. 5.2).

We discuss the quantitative results of our model in altering an unfunny scene to make it funny in Sec. 5.2.2), and the *vice versa* in Sec. 5.2.3. In Sec. 5.3, we report qualitative results through human studies.

5.1. Predicting Funniness Score

This section presents performance of the SVR (Sec. 4.2) that predicts the funniness score F_i of a scene.

Metric. We use average relative error to quantify our model's performance computed as follows:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{|Predicted\ F_i - Ground\ Truth\ F_i|}{Ground\ Truth\ F_i} \tag{1}$$

where N is the number of test scenes and F_i is the funniness score for the test scene i.

Baseline: The baseline model always predicts the average funniness score of the training scenes.

Model. As shown in Table 1, we observe that our model trained using combinations of different scene-level features (described in Sec. 4.1) performs better than the baseline model. We see that Location features perform slightly better than Cardinality. This makes sense because Location features also have occurrence information. The Embedding does not have location information and hence does worse. Due to some redundancy (all features have occurrence information), combining them does not improve performance.

Features	**Avg. Rel. Err.**
Avg. Prediction Baseline	0.3151
Embedding	0.2516
Cardinality	0.2450
Location	0.2400
Embedding + Cardinality + Location	0.2400

Table 1: Performance of different feature combinations in predicting funniness score F_i of a scene.

5.2. Altering Funniness of a Scene

We discuss the performance in the tasks of identifying objects in a scene that contribute to humor (Sec. 4.2) and replacing those objects with other objects to reduce (or increase) humor (Sec. 4.3).

5.2.1 Predicting Objects to be Replaced

We train this model to detect objects instances that are funny in the scene. It makes a binary prediction whether each instance should be replaced or not.

Metric. Along with naïve accuracy (% of correct predictions, *i.e.*, Acc.), we also report average class-wise accuracy (*i.e.*, Avg. Cl. Acc.) to determine the performance of our model for this task. As the data is skewed, with the majority class being *not-replace*, we require our model to perform well both class-wise and as a whole.

Baselines:

1. **Priors.** We always predict that an instance should not be replaced. We also compute a stronger baseline that replaces an object if it is replaced at least T% of the time in training data. T was set to 20 based on the validation set.
2. **Anomaly Detection.** We use cosine similarity between object embedding (of each instance in the scene) and the scene embedding to predict anomalous objects in the scene. This is similar to finding the odd-one-out given a group of words [31]. Objects that have a cosine similarity less than a threshold T with the scene are predicted as anomalous objects and are replaced. A modification to this baseline is to replace K objects that are least similar to the scene. Based on performance on the validation set, T and K are determined to be 0.8 and 4, respectively.

Model. Table 2 compares the performance of our model with the baselines described above. We observe that the baseline based on priors performs better than anomaly detection. This is perhaps not surprising because the prior-based baseline, while naïve, is "supervised" in the sense that it relies on statistics from the training dataset of which objects tend to get replaced. On the other hand, anomaly detection is completely unsupervised since it only captures the context of objects in *normal* scenes. Our approach performs better than the baseline approaches in identifying objects that contribute to humor.

Method	Avg. Cl. Acc.	Acc.
Priors (do not replace)	39.93 %	**79.86**%
Priors (object's tendency to be replaced)	73.13 %	71.5%
Anomaly detection (threshold distance)	62.16 %	58.30%
Anomaly detection (top-K objects)	63.01 %	64.31%
Our model	**74.45**%	74.74%

Table 2: Performance of predicting whether an object should be replaced or not, for the task of altering a funny scene to make it unfunny. As the data is skewed with the majority class being "not-replace", we require our model to perform well both class-wise and as a whole.

On average, we observe that our model replaces 3.67 objects for a given image as compared to an average of 2.54 objects replaced in the ground truth. This bias to replace more objects ensures that a given scene becomes significantly less funny than the original scene. We observe that the model learns that in general, animate objects like humans and animals are potentially stronger sources of humor compared to inanimate objects. It is interesting to note that the model also learns fine-grained detail, *e.g.*, to replace older people playing outdoors (which may be considered funny) with younger people (Fig. 5, top row).

5.2.2 Making a Scene Unfunny

Given that an object is predicted to be replaced in the scene, the model has to also predict a suitable replacer object. In this section, we discuss the performance of the model in predicting these replacer objects. This model is trained and evaluated using ground truth annotations of objects that are replaced by humans in a scene. This helps us isolate performance between predicting *which objects to replace* and predicting *suitable replacers* .

Metric. In order to evaluate the performance of the model on the task of replacing funny objects in the scene to make it unfunny, we use the top-5 metric (similar to ImageNet [42]), *i.e.*, if any of our 5 most confident predictions match the ground truth, we consider that as a correct prediction.

Baselines:

1. **Priors.** Every object is replaced by one of its 5 most frequent replacers in the training set.
2. **Anomaly Detection.** We subtract the embedding of the object that is to be replaced from the scene embedding. The 5 objects from the clipart vocabulary that are most similar (in the embedding space) to this resultant scene embedding are the ones that contextually "fit in".

Model. We observe that the performance trend in Table 3 is similar to that observed in the previous section (Sec. 5.2.1), *i.e.*, our model performs better than priors, which performs better than anomaly detection. By qualitative inspection, we find that our top prediction is intelligent, but lazy. It eliminates humor in most scenes by choosing to replace objects contributing to humor with other objects that blend well into the background. By relegating an object to the background, it is rendered inactive and hence, cannot be contribute to humor in the scene. For *e.g.*, the top prediction is frequently "plant" in indoor scenes and "butterfly" in outdoor scenes. The 2nd prediction is both intelligent and creative. It effectively reduces humor while also ensuring diversity of replacer objects. Subsequent predictions from the model tend to be less meaningful. Qualitatively, we find the 2nd most confident prediction to be the best compromise.

Method	Top-5 accuracy
Priors (top 5 GT replacers)	24.53%
Anomaly detection (object that "fits" into scene)	7.69%
Our model	**29.65**%

Table 3: Performance of predicting which object to replace with, for the task of altering a funny scene to make it unfunny.

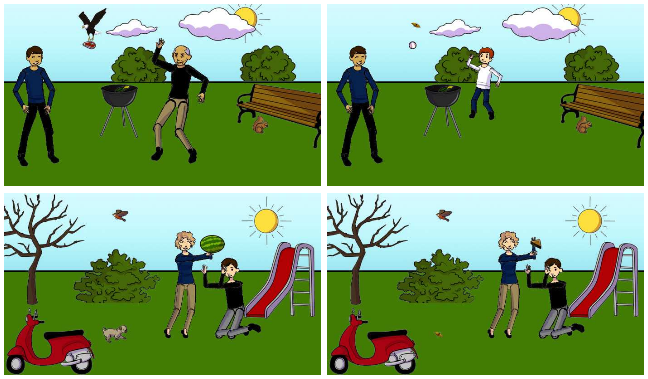


Figure 5: Fully automatic result of altering an input funny scene *(left)* into an unfunny scene *(right)*.

Full pipeline. Fig. 5 shows qualitative results from our full pipeline (predicting objects to replace and predicting their replacers) using the 2nd predictions made by our model.

5.2.3 Making a Scene Funny

We train our full pipeline model used in Sec. 5.2.2 on scenes from the FOR dataset to perform the task of altering an unfunny scene to make it funny. Some qualitative results are shown in Fig. 6.

5.3. Human Evaluation

We conducted two human studies to evaluate our full pipeline:

1. **Absolute:** We ask 10 workers to rate the funniness of the scene predicted by our model on a scale of 1-5. We then compare this with the F_i of the input funny scene.
2. **Relative:** We show 5 workers the input scene and the predicted scene (in random order) and ask them to indicate which scene is funnier.

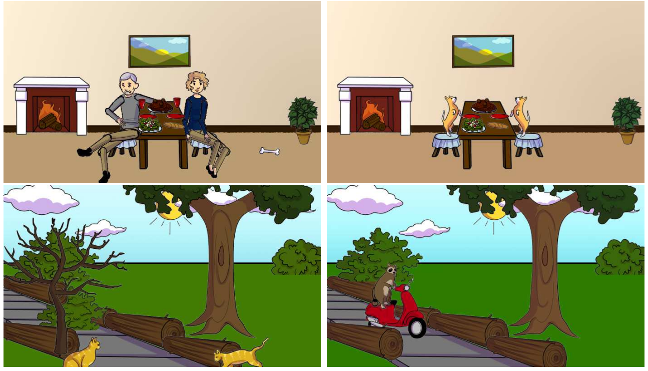


Figure 6: Fully automatic result of altering an input unfunny scene *(left)* into a funny scene *(right)*.

Funny to unfunny. As expected, the output scenes from our model are less funny than the input funny scenes on average. The average F_i of the input funny test scenes is 2.69. This is 1.05 points higher than the output unfunny scenes whose average F_i is 1.64. Unsurprisingly, in relative evaluation, workers find our output scenes to be less funny than the input funny scenes 95% of the time.

Unfunny to funny. During absolute evaluation, we find that the average F_i of scenes made funny by our model is 2.14. This is a relatively high score, considering that the average F_i score of the corresponding originally funny scenes that were created by workers is 2.69. Interestingly, the relative evaluation can be perceived as a *Turing test* of sorts, where we show workers the model's output funny scene and the original funny scene created by workers. 28% of the time, workers picked the model's scenes to be *funnier*.

6. Discussion

Humor is a subtle and complex human behavior. It has many forms ranging from slapstick which has a simple physical nature, to satire which is nuanced and requires an understanding of social context [54]. Understanding the entire spectrum of humor is a challenging task. It demands perception of fine-grained differences between seemingly similar scenarios. *E.g.*, a teenager falling off his skateboard (such as in America's Funniest Home Videos[2]) could be considered funny but an old person falling down the stairs is typically horrifying. Due to these challenges some people even consider computational humor to be an "AI-complete" problem [6, 19].

While understanding fine-grained semantics is important, it is interesting to note that there exists a qualitative difference in the way humor is perceived in abstract and real scenes. Since abstract scenes are not photorealistic, they afford us "suspension of reality". Unlike real images, the content depicted in an abstract scene is benign. Thus, people are likely to find the depiction more funny [27]. In our everyday lives, we come across a significant amount of humorous content in the form of comics and cartoons to which our computational models of humor are directly applicable. They can also be applied to learn semantics that can extend to photorealistic images as demonstrated by Antol *et al*. [2].

Recognizing funniness involves violation of our mental model of how the world "ought to be" [28]. In verbal humor, the first few lines of the joke (set-up) build up the world model and the last line (punch line) goes against it. It is unclear what forms our mental model when we look at images. Is it our priors about the world around us formed from our past experiences? Is it because we attend to different regions of the image when we look at it and gradually build an expectation of what to see in the rest of the image? These are some interesting questions regarding visual humor that remain unanswered.

7. Conclusion

In this work, we take a step towards understanding and predicting visual humor. We collect two datasets of abstract scenes which enable the study of humor at different levels of granularity. We train a model to predict the *funniness score* of a given scene. We also explore the different sources of humor depicted in the funny scenes via human studies. We train models using incongruity-based humor to alter a scene's funniness. The models learn that in general, animate objects like humans and animals contribute more to humor compared to inanimate objects. Our model outperforms a strong anomaly detection baseline, demonstrating that detecting humor involves something more than just anomaly detection. In human studies of the task of making an originally funny scene unfunny, humans find our model's output to be less funny 95% of the time. In the task of making a normal scene funny, our evaluation can be interpreted as a *Turing test* of sorts. Scenes made funny by our model were found to be funnier 28% of the time when compared with the original funny scenes created by workers. Note that our model would match humans at 50%. We hope that addressing the problem of studying visual humor using abstract scenes and the two datasets that are made public would stimulate further research in this new direction.

Acknowledgements. We thank the anonymous reviewers for their valuable comments and suggestions. This work was supported in part by the Paul G. Allen Family Foundation via an award to D.P. DB was partially supported by the National Science Foundation CAREER award, the Army Research Office YIP award, and an Office of Naval Research grant N00014-14-1-0679. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government or any sponsor. We thank Xinlei Chen for his work on earlier versions of the clipart interface.

[2] www.afv.com

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 1, 3, 5

[2] S. Antol, C. L. Zitnick, and D. Parikh. Zero-Shot Learning via Visual Abstraction. In *European Conference on Computer Vision, ECCV*, 2014. 1, 3, 8

[3] Aristotle and R. McKeon. *The Basic Works of Aristotle*. Modern Library, 2001. 2

[4] S. Attardo. *Linguistic theories of humor*. Walter de Gruyter, 1994. 1

[5] Bharata-Muni and M. Ghosh. Natya shastra (with english translations). 1951. 2

[6] K. Binsted, B. Bergen, D. O'Mara, S. Coulson, A. Nijholt, O. Stock, C. Strapparava, G. Ritchie, R. Manurung, and H. Pain. Computational humor. *IEEE Intelligent Systems*, 2006. 8

[7] K. Binsted and G. Ritchie. Computational rules for generating punning riddles. *Humor: International Journal of Humor Research*, 1997. 2

[8] E. R. Bressler, R. A. Martin, and S. Balshine. Production and appreciation of humor as sexually selected traits. *Evolution and Human Behavior*, 2006. 1

[9] M. Buijzen and P. M. Valkenburg. Developing a typology of humor in audiovisual media. *Media Psychology*, 2004. 2, 4

[10] D. M. Buss. The evolution of human intrasexual competition: Tactics of mate attraction. *Journal of Personality and Social Psychology*, 1988. 1

[11] D. Davidov, O. Tsur, and A. Rappoport. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Conference on Computational Natural Language Learning*, 2010. 2

[12] A. Deza and D. Parikh. Understanding image virality. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2015. 3

[13] R. L. Duran. Communicative adaptability: A measure of social communicative competence. *Communication Quarterly*, 1983. 1

[14] R. L. Duran. Communicative adaptability: A review of conceptualization and measurement. *Communication Quarterly*, 1992. 1

[15] D. F. Fouhey and C. L. Zitnick. Predicting object dynamics in scenes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2014. 1, 3

[16] S. Freud. *The Joke and Its Relation to the Unconscious*. Penguin, 2003. 2

[17] J. D. Goodchilds, J. Goldstein, and P. McGhee. On being titty: Causes, correlates, and consequences. *The Psychology of Humor: Theoretical Perspectives and Empirical Issues*, 1972. 1

[18] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Gool. The interestingness of images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2013. 3

[19] M. M. Hurley, D. C. Dennett, and R. B. Adams. *Inside jokes: Using humor to reverse-engineer the mind.* MIT Press, 2011. 8

[20] P. Isola, D. Parikh, A. Torralba, and A. Oliva. Understanding the intrinsic memorability of images. In *NIPS*, 2011. 3

[21] A. Khosla, A. Das Sarma, and R. Hamid. What makes an image popular? In *International Conference on World Wide Web*, 2014. 3

[22] S. Kottur, R. Vedantam, J. M. Moura, and D. Parikh. Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. 2015. 1, 3

[23] X. Lin and D. Parikh. Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2015. 1, 3

[24] A. Mahapatra and J. Srivastava. Incongruity versus incongruity resolution. In *Proceedings of the 2013 International Conference on Social Computing*, 2013. 2

[25] R. A. Martin and N. A. Kuiper. Daily occurrence of laughter: Relationships with age, gender, and type a personality. *Humor*, 1999. 1

[26] P. E. McGhee. Chapter 5: The contribution of humor to children's social development. *Journal of Children in Contemporary Society*, 1989. 1

[27] A. P. McGraw and C. Warren. Benign violations making immoral behavior funny. *Psychological Science*, 2010. 8

[28] R. Mihalcea. The multidisciplinary facets of research on humour. In *International Workshop on Fuzzy Logic and Applications*, 2007. 2, 8

[29] R. Mihalcea and S. Pulman. Characterizing humour: An exploration of features in humorous texts. *Computational Linguistics and Intelligent Text Processing*, 2007. 2

[30] R. Mihalcea and C. Strapparava. Making computers laugh: Investigations in automatic humor recognition. In *EMNLP*, 2005. 2

[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 5, 6

[32] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 2013. 5

[33] D. Mobbs, M. D. Greicius, E. Abdel-Azim, V. Menon, and A. L. Reiss. Humor modulates the mesolimbic reward centers. *Neuron*, 2003. 1

[34] M. P. Mulder and A. Nijholt. *Humor Research: State of the Art*. University of Twente, Centre for Telematics and Information Technology, 2002. 2

[35] B. I. Murstein and R. G. Brust. Humor and interpersonal attraction. *Journal of Personality Assessment*, 1985. 1

[36] S. Petrovic and D. Matthews. Unsupervised joke generation from big data. In *ACL*, 2013. 2

[37] Plato, E. Hamilton, and H. Cairns. *The Collected Dialogues of Plato, Including the Letters.* Pantheon Books, 1961. 2

[38] B. Plester. Healthy humour: Using humour to cope at work. *New Zealand Journal of Social Sciences Online*, 2009. 1

[39] D. Radev, A. Stent, J. Tetreault, A. Pappu, A. Iliakopoulou, A. Chanfreau, P. de Juan, J. Vallmitjana, A. Jaimes, and R. Jha. Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest. *arXiv preprint arXiv:1506.08126*, 2015. 2

[40] P. Rinck. Magnetic resonance in medicine. the basic textbook of the european magnetic resonance forum. 8th edition; 2014. 1

[41] W. Ruch, S. Attardo, and V. Raskin. Toward an empirical verification of the general theory of verbal humor. *Humor: International Journal of Humor Research*, 1993. 1, 2

[42] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015. 7

[43] P. Salovey, A. J. Rothman, J. B. Detweiler, and W. T. Steward. Emotional states and physical health. *American Psychologist*, 2000. 1

[44] A. Salvatore and V. Raskin. Script rheory revisited: Joke similarity and joke representation model. *Humor-International Journal of Humor Research*, 1991. 1, 2

[45] D. Shahaf, E. Horvitz, and R. Mankoff. Inside jokes: Identifying humorous cartoon captions. In *SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015. 2

[46] G. Sinicropi. La struttura della parodia— avvero: Bradamante in arli. *Strumenti Critici Torino*, 1981. 1

[47] O. Stock and C. Strapparava. HAHAcronym: A computational humor system. In *ACL*, 2005. 2

[48] J. M. Suls. A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. *The Psychology of Humor: Theoretical Perspectives and Empirical Issues*, 1972. 2

[49] J. Taylor and L. Mazlack. Computationally recognizing wordplay in jokes. *Proceedings of CogSci*, 2004. 2

[50] R. Vedantam, X. Lin, T. Batra, C. L. Zitnick, and D. Parikh. Learning common sense through visual abstraction. In *ICCV*, 2015. 1, 3

[51] W. Y. Wang and M. Wen. I can has cheezburger? a nonparanormal approach to combining textual and visual information for predicting and generating popular meme descriptions. In *NAACL*, 2015. 2

[52] M. B. Wanzer, M. Booth-Butterfield, and S. Booth-Butterfield. Are funny people popular? an examination of humor orientation, loneliness, and social attraction. *Communication Quarterly*, 1996. 1

[53] K. K. Watson, B. J. Matthews, and J. M. Allman. Brain activation during sight gags and language-dependent humor. *Cerebral Cortex*, 2007. 1

[54] Wikipedia. Humor, November 2015. 8

[55] Wikipedia. Theories of humor, April 2016. 2

[56] D. Yang, A. Lavie, C. Dyer, and E. Hovy. Humor recognition and humor anchor extraction. 2015. 2

[57] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2013. 1, 3

[58] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the visual interpretation of sentences. In *ICCV*, 2013. 1, 3

[59] C. L. Zitnick, R. Vedantam, and D. Parikh. Adopting abstract images for semantic scene understanding. *PAMI*, 2014. 3