

We Are Humor Beings: Understanding and Predicting Visual Humor

Supplementary Material

Overview of Supplementary material

In the following appendix we provide:

- I. Inter-human agreement on funniness ratings in the Abstract Visual Humor (AVH) dataset.
- II. Details of the model architecture used to learn object embeddings and visualizations of its embeddings.
- III. A sample of objects from the abstract scenes vocabulary.
- IV. Examples of scenes from our datasets.
- V. Analysis of occurrences of different object types in scenes from our datasets.
- VI. The user interfaces used to collect scenes for the AVH and Funny Object Replaced (FOR) datasets.

Inter-human Agreement

In this section, we describe our experiment to determine inter-human agreement in funniness ratings of scenes. The Abstract Visual Humor (AVH) dataset contains 3,028 funny scenes and 3,372 unfunny scenes that were created by Amazon Mechanical Turk (AMT) workers. The funniness of each scene in the dataset is rated by 10 different workers on a scale of 1-5. We define the *funniness score* of a scene, as the average of all ratings for a scene. In this section, we investigate the extent to which people agree regarding the funniness of a scene.

Perception of an image differs from one person to another. Moran *et al.* [6] treat humor appreciation by people as a personality characteristic. We investigate to what extent people agree how funny each scene in our dataset is. We split the votes we received for each scene into two groups, keeping each individual worker’s ratings in the same group to the extent possible. We compute the *funniness score* of each scene across workers in each group. We compute Pearson’s correlation between the two groups. Fig. 1 shows a plot of Pearson’s correlation (y-axis) vs. the number of workers (x-axis). We can see that inter-human agreement increases as we increase the number of workers in a group and that the trend is gradually saturating. This indicates that ratings from 10 workers is sufficient to compute a reliable *funniness score*.

We observed that the standard deviation among ratings from 10 different workers for funny scenes is 1.09, and for unfunny scenes is 0.73. *I.e.*, people agree more on scenes

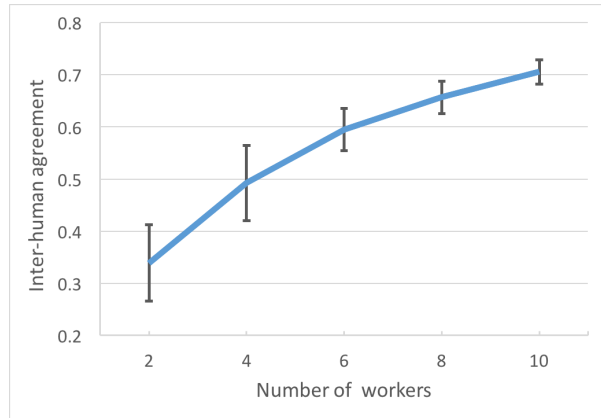


Figure 1: Inter-human agreement (y-axis) as we collect funniness ratings from more workers (x-axis). We see that by 10 ratings, we are starting to saturate with high agreement, indicating that 10 ratings is sufficient for a reliable *funniness score*.

that are clearly not funny than on ones that are funny, matching our intuition that humor is subjective, while the lack thereof is not.

Object Embeddings

In this section, we describe our model that learns embeddings for clipart objects and present visualizations of these embeddings. We learn distributed representations for each object category in the abstract scenes vocabulary using a word2vec-style continuous Bag-of-Words model [5]. During training, subsets of 6 objects are sampled from all of the objects present in a scene and the model tries to predict one of the objects, given the other 5. Each object is assigned a 150-d vector, which is randomly initialized. The vectors corresponding to the 5 context objects are projected to an embedding space via a single layer whose parameters are shared between the 5 objects. This (randomly initialized) layer consists of 150 hidden units without a non-linearity after it. The sum of these 5 object projections is used to compute a softmax over the 150 classes in the object vocabulary. Using the correct label (*i.e.*, the object category of the 6th object), the cross-entropy loss is computed and backpropa-

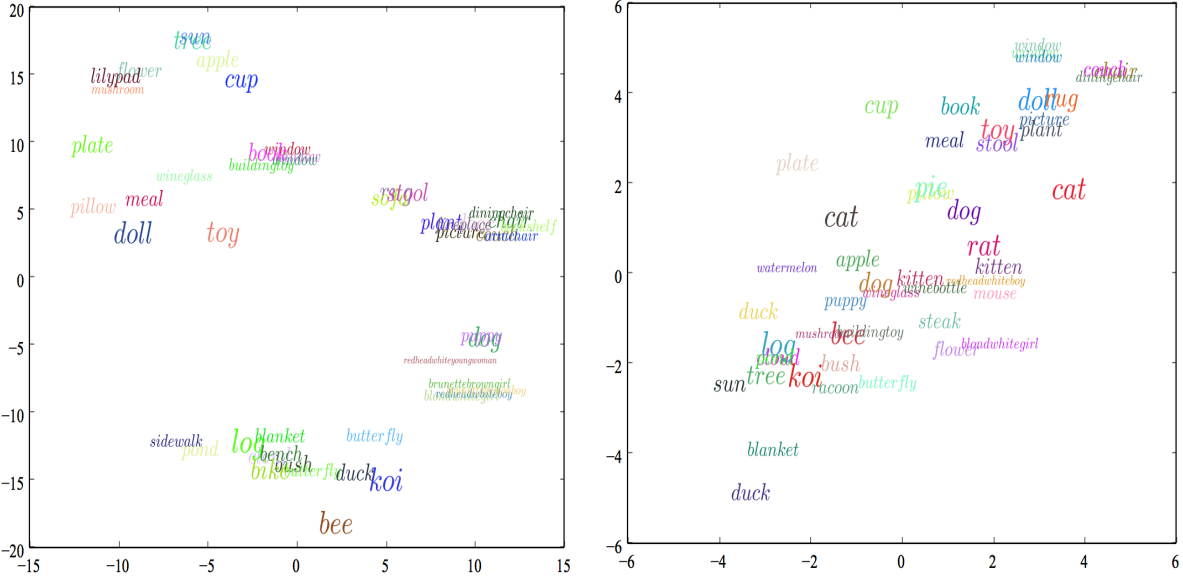


Figure 2: *Left*. Visualization of “normal” object embeddings of 75 most frequent objects in unfunny scenes. We see that closely placed objects have semantically similar meanings. *Right*. Visualization of “humor” embeddings of 75 most frequent objects in funny scenes. We see that objects that are close in the “humor” embedding space may be semantically very different.

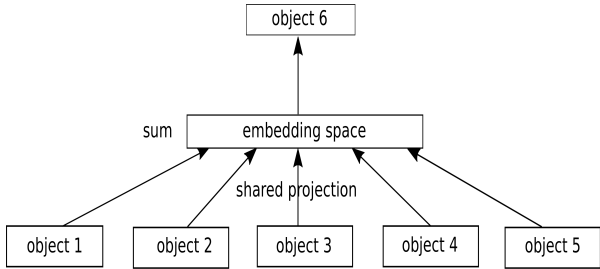


Figure 3: The continuous Bag-of-Words model used to obtain the object embeddings.

gated to learn all network parameters. The model is trained using Stochastic Gradient Descent with a base learning rate of 0.0001 and a momentum update of 0.9. The learning rate was reduced by a factor of two after each epoch. A diagram of the model can be seen in Fig. 3.

The context provided by the 5 objects ensures that the representations learnt reflect the relationships between objects. *I.e.*, objects that are semantically related tend to have similar representations. We learn the “normal” embeddings (*i.e.*, the object embedding instance-level features from the main paper) from 11K scenes collected by Antol *et al.* [1]. As these scenes were not intended to be humorous, the relationships captured in the embeddings are the ones that occur naturally in the abstract scenes world.

Fig. 2 (*left*) is a t-SNE [2] visualization of the “normal” embeddings for the 75 most frequent objects in unfunny scenes. In Fig. 2 (*right*), we also visualize “humor” embeddings, which were not used as features but provide us with insights. These are learnt from the 3,028 funny scenes in the AVH dataset.

We observe that the “normal” embeddings encode a notion for which object categories occur in similar contexts. We also observe that closely placed objects in the “normal” embedding space have semantically similar meanings. For instance, humans are clustered together around coordinates (10, -7). Interestingly, “dog” and “puppy” (coordinates (10, -5)) are placed together and furniture like “chair”, “bookshelf”, “armchair”, *etc.* are placed together (coordinates (10, 5)). This follows from the distributional hypothesis, which states that words which occur in the similar contexts tend to have similar meanings [3, 4].

In contrast, in the “humor” embeddings, visualized in Fig. 2 (*right*), we see that objects that are close in the embedding space may be semantically very different. For instance, “dog” and “wine glass” are placed together at coordinates (0, 0). These are placed far apart (at opposite ends) in the “normal” embedding. However, in the “humor” embedding, these two categories are extremely close to each other; even closer than semantically similar categories like two breeds of dogs. We hypothesize that this because our dataset contains funny scenes consisting of dogs with wine glasses,

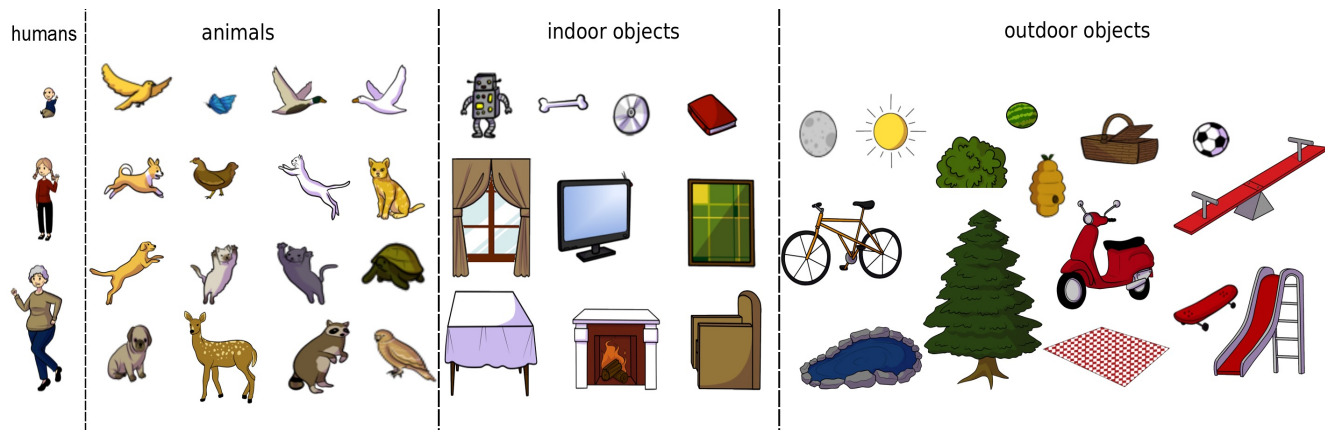


Figure 4: A subset of clipart objects from the abstract scenes vocabulary.

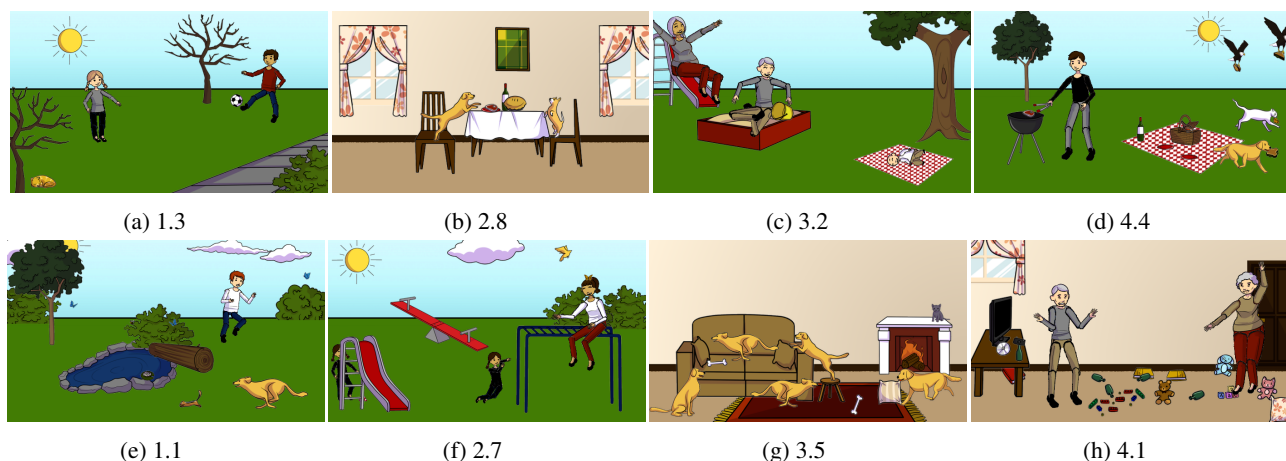


Figure 5: Spectrum of scenes from our AVH dataset that are arranged in ascending order of *funniness score* (shown in the sub-caption)

e.g., Fig. 5b. It is interesting to note that “background” objects that do not contribute to humor in a scene are also placed together. For example, “chair”, “couch”, and “window” are placed together in the “humor” embedding as well (coordinates (4, 5)).

The understanding of semantically similar object categories that can occur in a context, represented by the “normal” embeddings, can be interpreted as a person’s mental model of the world. The “humor” embeddings capture deviations or incongruities from this “normal” view that might cause humor.

Abstract Scenes Vocabulary

The abstract scenes interface developed by Antol *et al.* [1] consists of 20 “deformable” humans, 31 animals in different poses, and about 100 objects that can be found

in indoor scenes (*e.g.*, couch, picture, doll, door, window, plant, fireplace) or outdoor scenes (*e.g.*, tree, pond, sun, clouds, bench, bike, campfire, grill, skateboard). In addition to the 8 different expressions available for humans, the ability to vary the pose of a human at a fine-grained level enables these abstract scenes to effectively capture the semantics of a scene. The large clipart vocabulary (of which only a fraction is shown to a worker during creation of a scene) ensures diversity in the scenes being depicted. A subset of objects from our Abstract Scenes vocabulary is shown in Fig. 4.

Example Scenes

In this section, we present examples of scenes that were created using the abstract scenes interface. Fig. 5, depicts a

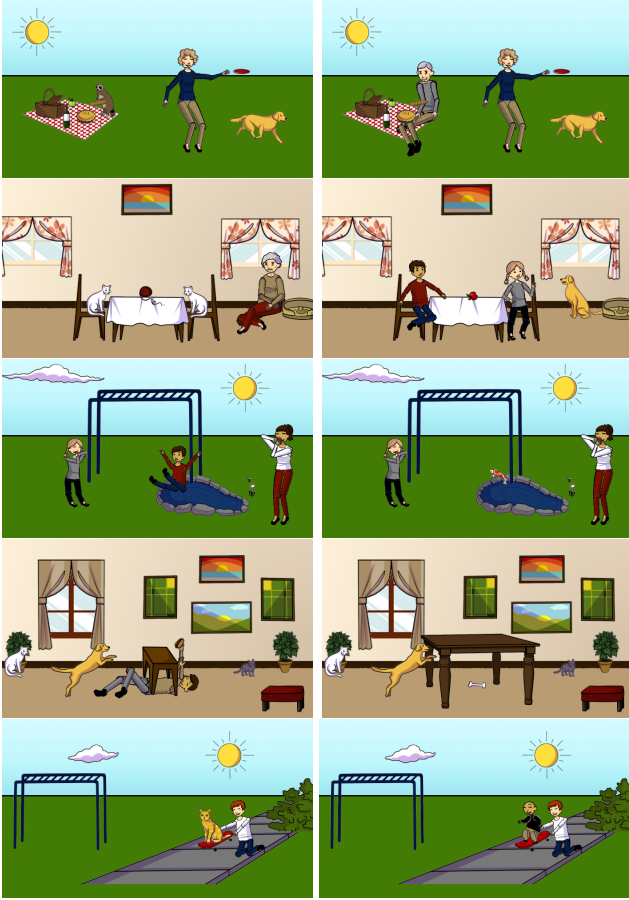


Figure 6: Some example originally funny scenes (*left*) and their object-replaced unfunny counterparts (*right*) from the FOR dataset.

spectrum of scenes from the AVH dataset in ascending order of *funniness score*. These scenes were created by AMT workers using the interface presented in Fig. 9.

Fig. 6 shows originally funny scenes (*left*) and their unfunny counterparts (*right*) from the FOR dataset. AMT workers created the counterparts by replacing as few objects in the originally funny scene such that the resulting scene is not funny anymore. A screenshot of the interface that was used to create the unfunny counterparts is shown in Fig. 10.

Object Type Occurrences

In this section, we first analyze the occurrence of each object type in funny and unfunny scenes. We then analyze the most commonly cooccurring object types in funny scenes as compared to unfunny scenes.

Distribution of Object Types. We analyze the distribution of object types in funny and unfunny scenes across all scenes in our dataset. We compute the frequency of appear-

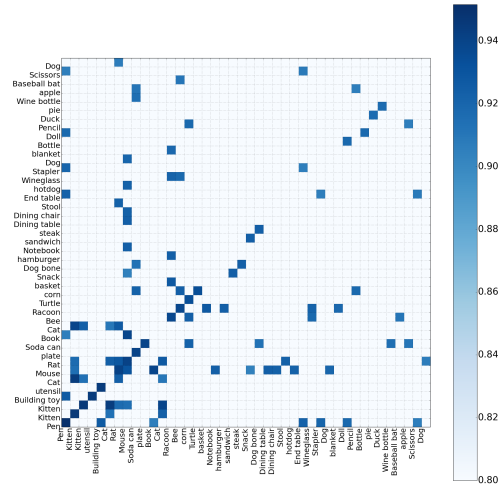


Figure 7: Top 100 object pairs that have the highest probabilities of cooccurring in a funny scene. Please note that repeated entries for an object type (*e.g.*, “dog”), correspond to slightly different versions (*e.g.*, breeds) of the same object type.

ance of each object type in funny and unfunny scenes. We use this to compute the probability of a scene being funny, given that an object is present in the scene, which is shown in *blue* in Fig. 8. Since we have more unfunny scenes than funny scenes, we use normalized counts.

We observe that the humans that most appear in funny scenes are elderly people. This is probably because a number of scenes in our dataset depict old men behaving unexpectedly, *e.g.*, dancing or playing in the park as shown in Fig. 5c, which is funny. Interestingly, we also observe that in general, animals appear more frequently in funny scenes. Animals like “mouse”, “rat”, “raccoon” and “bee” appear in funny scenes significantly more than they do in unfunny scenes. Other objects having a strong bias towards appearing in funny scenes include “wine bottle”, “pen”, “scissors”, “tape”, “game” and “beehive”. Thus, we see that certain object types have a tendency to appear in funny scenes. A possible reason for this is that these objects are involved in funny interactions, or are intrinsically funny, and hence contribute to humor in these scenes.

Funny Cooccurrence Matrix. We populate two object cooccurrence matrices – \mathbf{F} and \mathbf{U} , corresponding to funny scenes and unfunny scenes, respectively. Each element in \mathbf{F} and \mathbf{U} corresponds to the count of the cooccurrence of a pair of objects across all funny and unfunny scenes, respectively. To enable the study of types of cooccurrences that contribute to humor, we compute the probability of a

scene being funny, given that a pair of objects cooccur in the scene as $\frac{F}{F+U}$, which is shown in Fig. 7 for the top 100 probable combinations that exist in a funny scene. Please note that repeated entries for an object type (e.g., “dog”), correspond to slightly different versions (e.g., breeds) of the same object type. An interesting set of object pairs that are present in funny scenes are “rat” appearing alongside “kitten”, “cat”, “stool”, and “dog”. Another interesting set of combinations is “raccoon” cooccurring with “bee”, “hamburger”, “basket”, and “wine glass”. We observe that this matrix captures interesting and unusual combinations of objects that appear together frequently in funny scenes.

User Interfaces

In this section, we present the user interfaces that were used to collect data from AMT. Fig. 9 shows a screenshot of the user interface that we used to collect funny scenes. Objects in the clipart library (on the right in the screenshot) can be dragged on to any part of the empty canvas shown in the figure. The pose, flip (i.e., lateral orientation), and size of all objects can be changed once they are placed in the scene. In the case of humans, one of 8 expressions must be chosen (initially humans have blank faces) and fine-grained pose adjustments are required. Fig. 10 shows the interface that we used to collect “object-replaced” scenes for our FOR dataset. We showed workers an originally funny scene and asked them to replace objects in that scene so that the scene is not funny anymore. On clicking an object in the original scene, the object gets highlighted in green. A replacer object can then be chosen from the clipart library (displayed on the right in the screenshot). Objects that are replaced in the original scene show up in the empty canvas below. At any point, to undo a replacement, a user can click on the object in the below canvas and the corresponding object will be placed at its original position in the scene. The interface does not allow for the movement or the removal of objects.

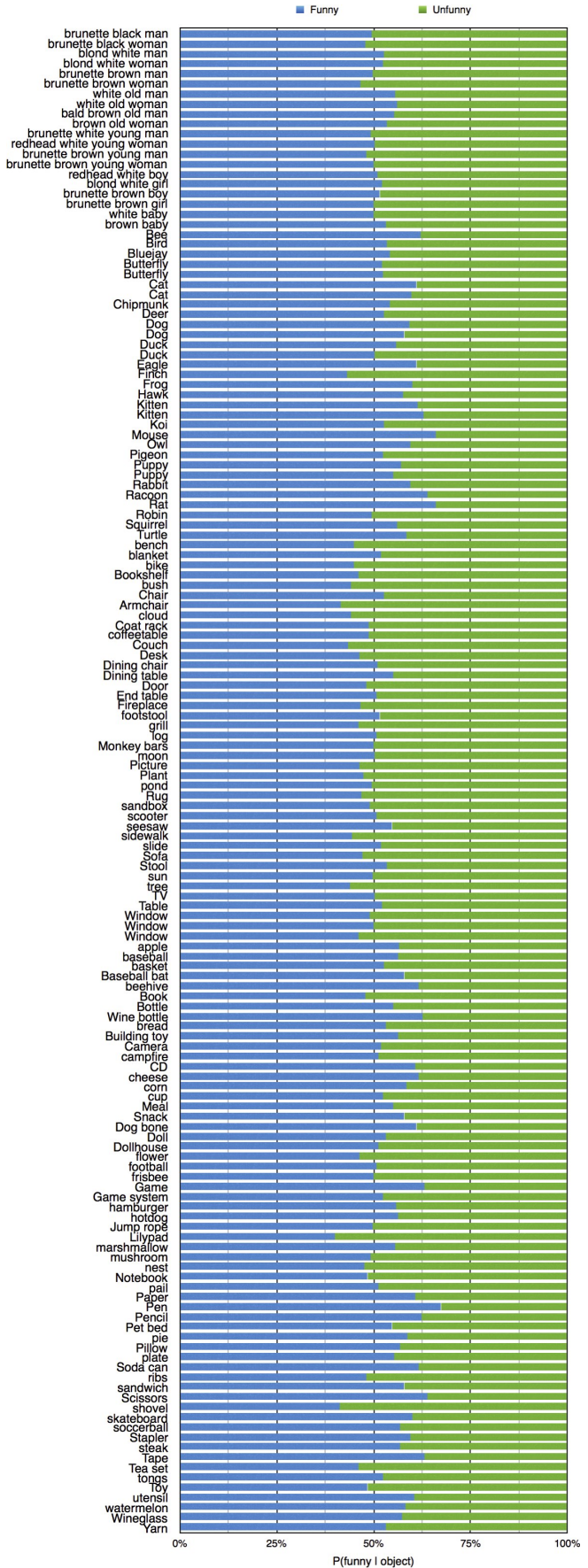


Figure 8: Probability of scene being funny, given object.

Depict Funny Scenarios! (Living/Dining Room)

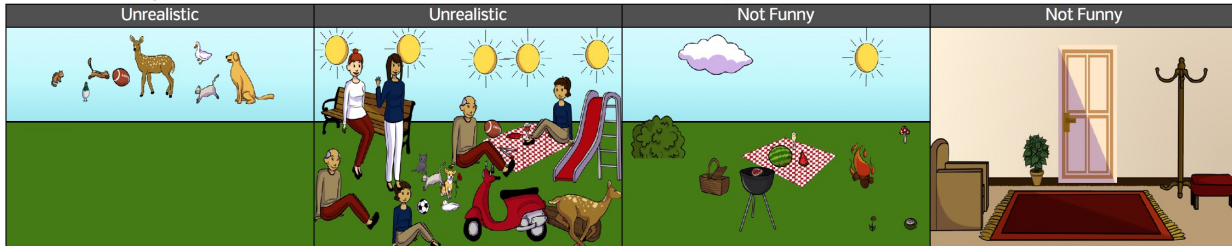
[Images may take some time to load] [Spamming will get blocked]
 [Tested with Chrome, Firefox and Safari. Interface may not work well with Internet Explorer]

Using the clipart interface below, please create scenes where a funny scenario is being depicted.

Please follow the **instructions** carefully, otherwise your work **WILL BE REJECTED**.

1. While funny, make your scenarios **realistic** and **meaningful** (e.g., the scene should **not** contain a **random assortment** of clipart pieces).
2. **Other people** should also find your scenario funny (e.g., no inside jokes).
3. Please use **at least 6 pieces** of clipart in the scene.
4. If you do multiple HITs, please be sure to create very **different scenarios** across HITs and not minor variations of a previously created scenario.
5. Give us a **description** of why you think the scenario is funny. Once you create a scene and click next, you will be asked to provide a **description** of what about the scenario is funny.

Below are examples of **bad scenarios** that are either not realistic or not funny:



Clipart objects (5 instances each) may be added by dragging them onto the scene and removed by dragging them off. They may be **resized (CTRL + a/CTRL + z)**, **flipped (CTRL + c)**, **sent backward (CTRL + s)** or **brought forward (CTRL + x)**.

You will be asked to complete 2 tasks.

You can go back and forth between all of your scenarios by pressing "Prev" and "Next". When you finish your last one, a pop-up will ask you to submit the HIT. We'd love to hear any feedback you have about the usability of the interface, any bugs you encounter, or the HIT in general, so feel free to leave a comment.

Thanks for your work!

Scenario 1/2
Prev Next

Type

Scene Depth

Flip

People	Animals	Large objects	Small objects
5	5	5	5
5	5	5	5
5	5	5	5
5	5	5	5
5	5	5	5

Figure 9: User interface used to create the funny scenes in the AVH dataset.

Make Scenarios Not Funny! (Park)

[Interface **initially** takes a few seconds to load. All scenarios thereafter load **instantly**] [Tested with Chrome, Firefox and Safari. Interface may not work well with Internet Explorer]

Your Task:

We will show you a funny scenario. Your task is to **replace objects from the scenario** such that the scenario goes from **funny to not funny**. Please be sure to follow **all** of these conditions:

1. Replace the **minimum number of objects** from the scenario so that the scenario is not funny anymore.
2. Please replace an object with another object that is **most similar** to the existing object to the extent possible, but makes the scenario go from funny to not funny.
3. The final scenario should be a **realistic scenario that is not funny**.

To replace an object from the scenario:

- a) Click on the object in the scenario to select it
- b) Click on the object in the object library to the right of the screen that you want to replace the object with
- c) The objects you remove will show up in the empty background of the scenario shown at the bottom. To **undo** replacement of an object, click on the object at the bottom. This will bring the object back to the above scenario.

When an object or a human is replaced by another object, you can choose one of the available poses for the added object. When an object is replaced by a human, you can change the human's expression and pose by rotating (clicking and dragging) the limbs of the human.

NOTE: When a human from the original scenario is replaced with another human, the expression and pose of the human added to the scenario **can not** be changed.

You will be asked to complete **5 tasks**.

You can go back and forth between all of your scenarios by pressing "Prev" and "Next". When you finish your last one, a pop-up will ask you to submit the HIT. We'd love to hear any feedback you have about the usability of the interface, any bugs you encounter, or the HIT in general, so feel free to leave a comment.

Thanks for your work!

[SHOW EXAMPLES](#)




















Scenario 1/5

[PREV](#)

[NEXT](#)

Expression

Scene Depth Flip

People	Animals	Large objects	Small objects
 4	 5	 5	 5
 5	 5	 5	 5
 5	 3	 5	 5
 5	 5	 5	 4
 5	 4	 4	

Click on an object below to **undo** your replacement

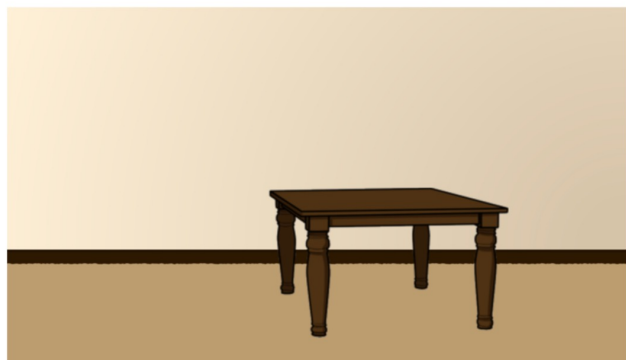


Figure 10: User interface to replace objects for the FOR dataset.

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. [2](#), [3](#)
- [2] L. V. der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008. [2](#)
- [3] J. R. Firth. *A synopsis of linguistic theory*. Blackwell, 1957. [2](#)
- [4] Z. S. Harris. Distributional structure. *word*, 10 (2-3): 146–162. reprinted in fodor, j. a and katz, jj (eds.), *readings in the philosophy of language*, 1954. [2](#)
- [5] 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. [1](#)
- [6] J. M. Moran, M. Rain, E. Page-Gould, and R. A. Mar. Do i amuse you? asymmetric predictors for humor appreciation and humor production. *Journal of Research in Personality*, 2014. [1](#)