

Learning Words by Drawing Images

D'ídac Surís, Adria Recasens, David Bau, David Harwath, James Glass, Antonio Torralba

Elif Çalışkan
2016400183

Introduction

- Goal is to learn the correspondence between spoken words and abstract visual attributes, from a dataset of spoken descriptions of images.

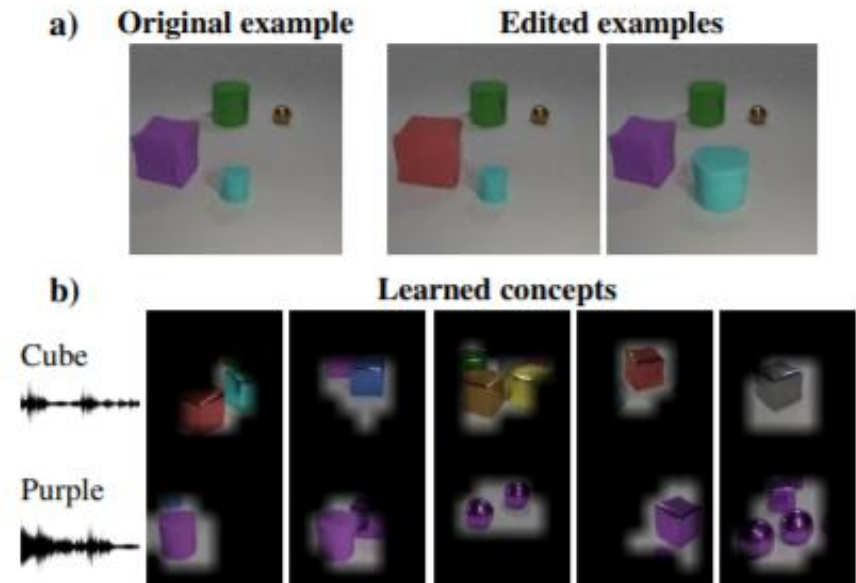


Figure 1: In this paper we propose a framework for learning through generated images. In a) two edited examples are compared to an original generated image at the left. In b) we show the concepts learned by our system when trained with edited examples.

Related Work

- **Concept learning**
- **Curriculum learning:** reformulates the curriculum problem by proposing that the training data can be synthesized by a GAN
- **CLEVR dataset:** as a highly controlled visual domain in which compositional attributes are clear, and that a GAN can learn to draw well
- **Generative Adversarial Networks:** to generate high-resolution photorealistic images for creating training data.
- **Audio and Image**

Images and descriptions

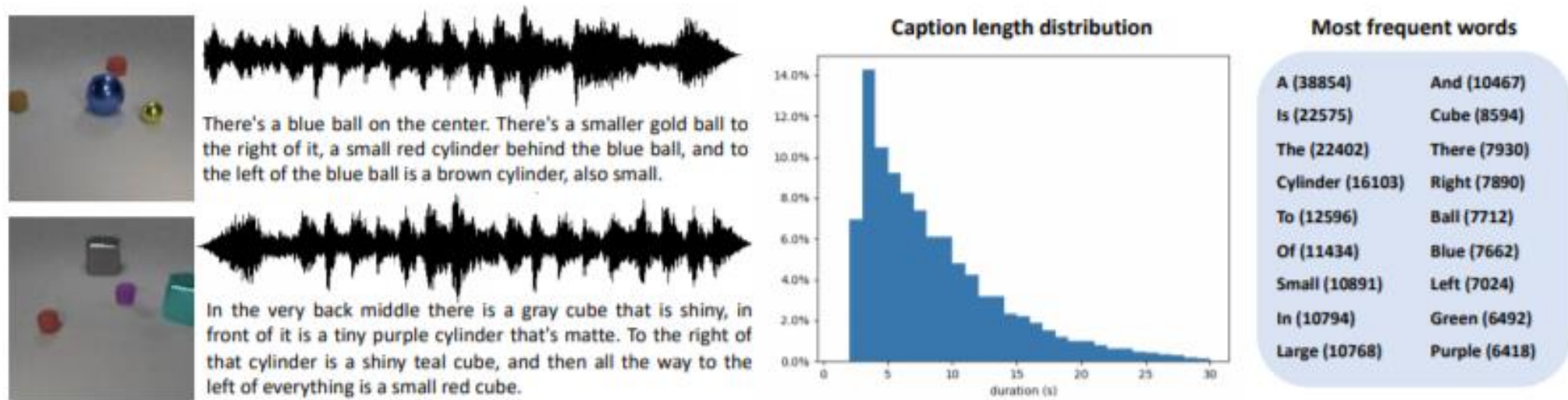


Figure 2: Examples of generated images and human annotated audios. In this figure, the transcriptions of the audio are shown instead of the audio, but no text transcriptions are used at any point during training or evaluation. We also provide some basic statistics of the Audio CLEVRGAN dataset.

Editing training examples

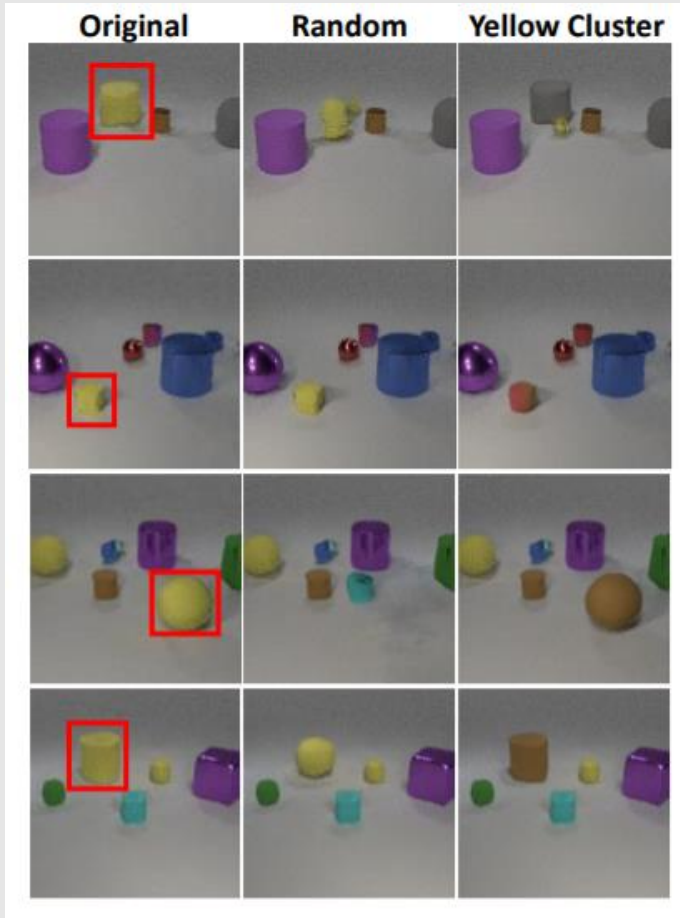
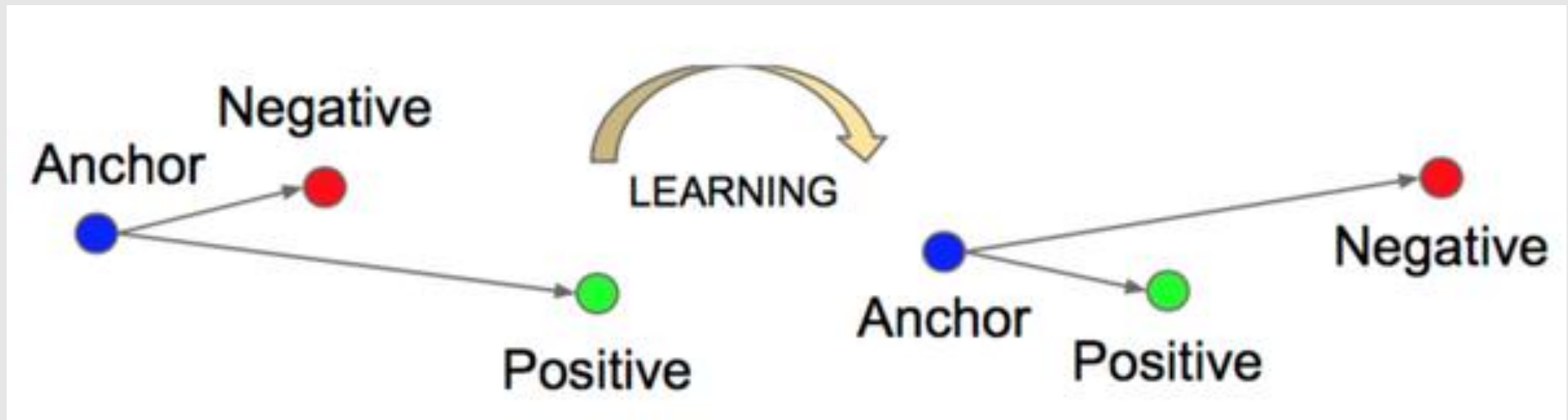


Figure 3: Examples of edited samples created using random editing and targeted interventions. In the left column, the original images with the target object in red. In the second column, randomly ablated units, applied to the same feature maps. Results range from distortions or complete change of the object (first and third rows), through useful semantic changes (fourth row), to barely noticeable changes (second row). In the last column, images generated by ablating the units corresponding to the yellow cluster. Ablating these units makes the yellow color change, as the cluster is representing this attribute.

Triplet Loss



Learning words by drawing images

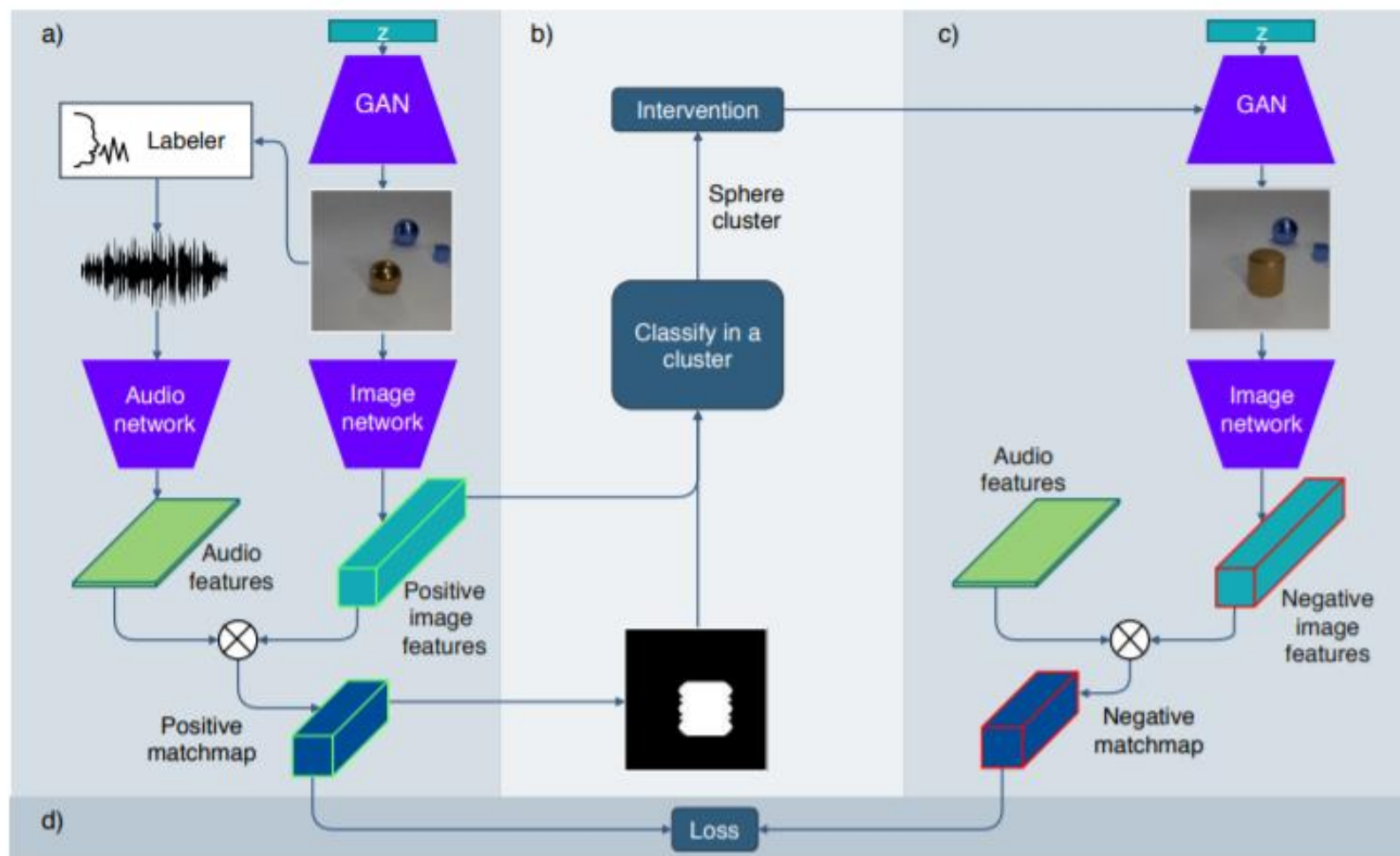


Figure 4: **Intervention schematic:** a) Basic model, where the original image and audio features are computed, as well as their matchmap. b) Clustering: highly activated image features are classified into a cluster, and an intervention is computed to generate an edited example. c) Generation of the edited example. The noise vector z is the same as in a). d) Triplet loss.

Clustering

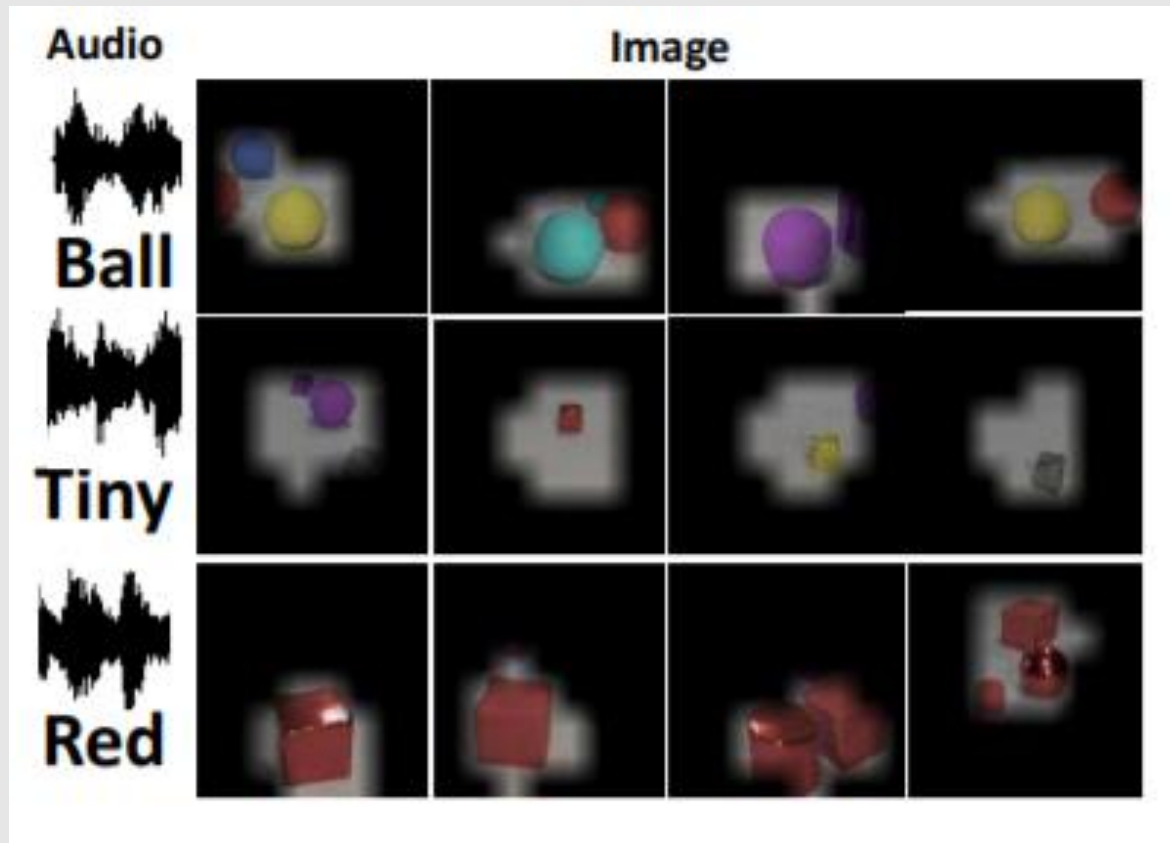


Figure 5: We show three examples of clusters learned by our model, represented by the images that mostly activate each cluster. We represent the audio-cluster with text for clarity, but all the learning is done in the audio domain. As shown, the system is able to learn color, shape and size.

Synthetic dataset creation

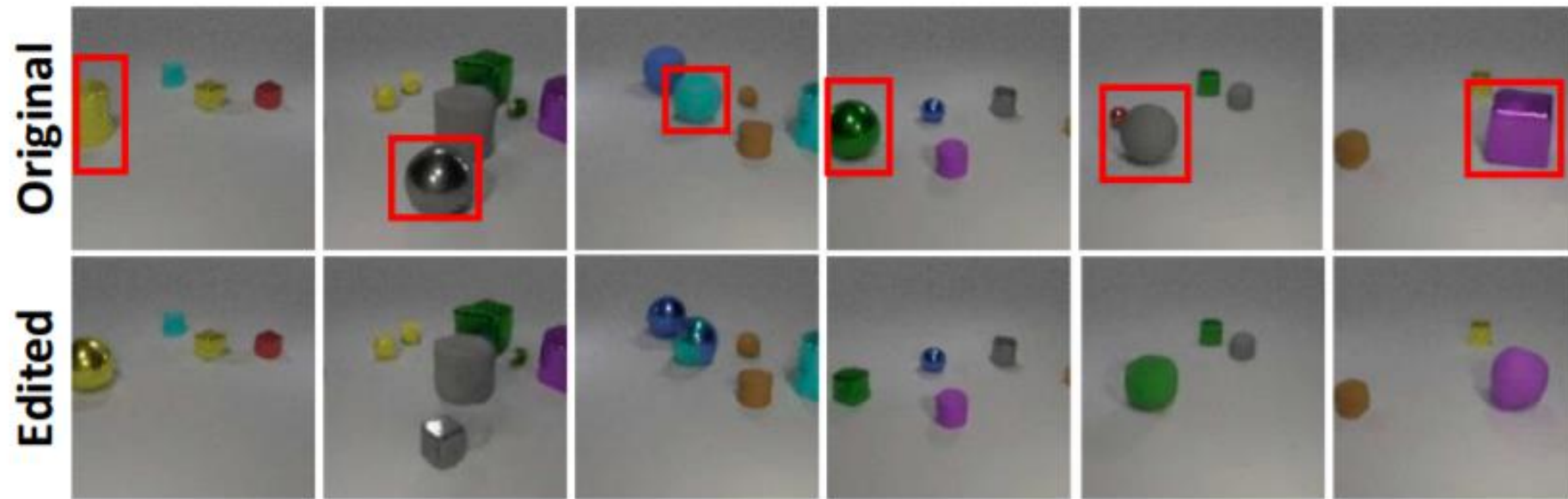


Figure 6: In this figure we show multiple examples of edited images using our targeted algorithm. Note that the system is able to modify particular attributes of the object.

Examples

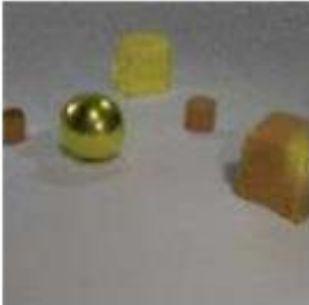

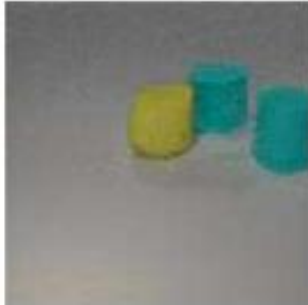





Caption	Selected Image	Ground Truth	Caption	Selected Image	Ground Truth
There is a gold metallic cube. To the left hand side and behind it there is a gold metallic sphere			In this picture I have two cubes in the back yellow and teal and in the front a till cylinder they are all large objects.		
A yellow square next to a large golden ball.			There is a small blue mat ball in front of a large green mat ball.		

Figure 7: Examples of our system selecting images given a caption. Note that the retrieved image usually is closely related with the given description.

Evaluation

		Shape	Material	Color	Size	Mean
Human dataset	DaveNet	50.3	60.8	86.8	72.2	67.6
	Random Edits	52.0	48.9	87.8	91.3	70.0
	Target Edits	54.1	63.0	86.2	91.3	73.7
	Hard Neg	53.6	60.8	88.4	87.8	72.7
	HN+Random Edits	54.8	63.0	87.9	87.8	73.4
	HN+Target Edits	56.2	67.4	87.9	88.7	75.1
Synthetic dataset	DaveNet	72.6	63.3	51.1	98.0	71.2
	Random Edits	70.9	97.8	54.0	96.9	79.9
	Target Edits	69.3	97.5	57.9	95.4	80.1
	Hard Neg	75.6	91.3	62.2	97.6	81.7
	HN+Random Edits	73.3	94.5	70.5	95.1	83.3
	HN+Target Edits	77.7	96.9	66.6	97.1	84.6

Table 1: **Semantic accuracy:** We evaluate the ability of the different models to detect particular attributes in image. Given an audio with only the attribute, we ask the system to discriminate between images with and without the attribute.

	Human Dataset			Synthetic Dataset		
	R@1	R@5	R@10	R@1	R@5	R@10
DaveNet	8.4	26.3	38.5	14.9	43.7	62.2
Random Edits	12.5	33.8	49.8	60.6	89.0	95.1
Targeted Edits	14.1	37.2	52.2	75.1	95.5	98.5
Hard Neg	20.5	45.1	60.7	73.4	94.6	97.6
HN+Random	19.3	48.3	63.0	94.8	99.7	99.9
HN+Targeted	20.3	49.3	61.9	93.4	99.6	99.9

Table 2: **Results in the Audio CLEVRGAN dataset:** Recall results (in %) for the two datasets, for the different methods, showing that more refined interventions get better results. Recall in the random test is over 500 samples.

Thank you for listening!