

CMPE537
Term Paper

LEARNING WORDS BY DRAWING IMAGES

Boğaziçi University

Elif Çalışkan

2016400183

1. Abstract:

Learning Words by Drawing Images paper tries to find a correspondence between correspondence between spoken words and abstract visual attributes, from a dataset of spoken descriptions of images. The problem appears very much alike how babies learn words and their visual meanings. Since the data they face is raw and unannotated, this makes it harder to differentiate each concept. Because this learning process is intriguing, my paper majors on teaching a machine to learn visual attributes based on drawing and hearing raw audio. State of the art was not mentioned in this paper, but I believe the work is closely related to “Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input”. Learning an object with a raw audio is completed but the visual attributes were not covered in this paper. So “Learning Words by Drawing Images” focuses on the correspondence between visual attributes such as size, shape etc. and raw audio.

2. Introduction:

It is a long and challenging process to learn a word for a human being. Since babies face to many difficulties during apprehending an unknown word such as raw and unaligned visual and acoustic information. Babies learn to understand speech and recognize objects in an extremely weakly supervised fashion, aided not by ground-truth annotations, but by observation, repetition, multi-modal context, and environmental interaction. But “Learning Words by Drawing Images” and “Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input” papers focus on the learning process from a raw audio of a machine rather than a human being.



The input to state of the art: images paired with waveforms of speech audio.

This question has been in researchers' minds before the increase in GAN generated images. State of the art generates a framework for pairing an object with its raw audio description using CNNs. “Learning Words by Drawing Images” follows a similar model using GAN images and finds clusters. But this time it also tries to find the closest image to given description. State of the art created “matchmap” neural networks which are capable of directly learning the semantic correspondences between speech frames and image pixels without the need for annotated training data in either modality.

But the process of directly generating images given a spoken description, or generating artificial speech describing a visual scene was not covered. We can see that “Learning Words by Drawing Images” paper has moved the learning model forward by adding this feature. The relations between objects were untouched in state of the art whereas the recent paper also contains those positional relations. Finally, a crucial element of human language learning is the dialog feedback loop which makes the network to fix any possible errors is also added in this paper by creating negative matchmaps.

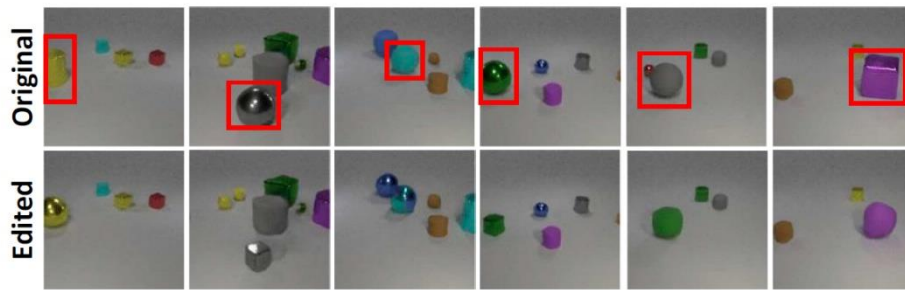


Figure 6: In this figure we show multiple examples of edited images using our targeted algorithm. Note that the system is able to modify particular attributes of the object.

In Problems section, I go through each challenge of this paper and explain the different techniques from previous papers. In Comparison section, 4 papers are compared to each other based on some criteria. Then I tell my ideas about state of the art in the following section. Lastly in the conclusion part, I go through the general process of the improvement in this area.

3 Problems:

3.1 Creating dataset:

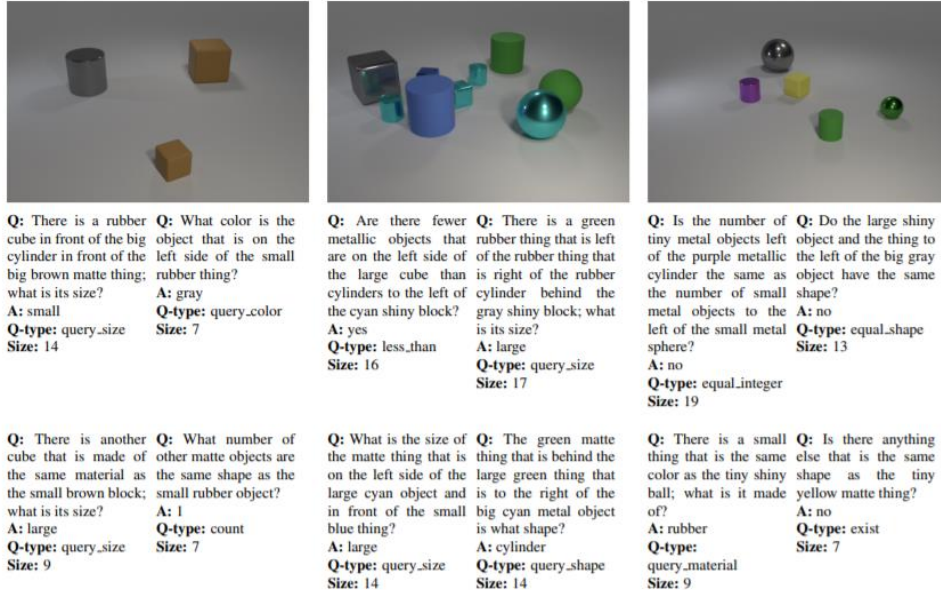
A long-standing goal of artificial intelligence research is to develop systems that can reason and answer questions about visual information. When building artificial intelligence systems that can reason and answer questions about visual data, we need diagnostic tests to analyze our progress and discover shortcomings. In the survey “**Visual Question Answering using Deep Learning: A Survey and Performance Analysis**”[1], many datasets are compared with each other such as:

Visual Madlibs: The Visual Madlibs dataset presents a different form of template for the Image Question Answering task.

Visual7W: The Visual7W dataset contains 47,300 COCO images with 327,939 question-answer pairs.

CLEVR[2]: CLEVR is a synthetic dataset to test the visual understanding of the VQA systems. The dataset is generated using three objects in each image, namely cylinder, sphere and cube. These objects are in two different sizes, two different materials and placed in eight different colors. The questions are also synthetically generated based on the objects placed in the image. The dataset also accompanies the ground truth bounding boxes for each object in the image.

These examples go on in the survey but the dataset that was used in this paper is CLEVR. “CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning” paper introduces a diagnostic dataset that tests a range of visual reasoning abilities without the effect of biasing. Each image is created using CLEVR dataset so that it is easily testable and precise. Below you can see the queries that are used for testing.



3.2 Editing training examples:

Because the visual attribute words are generally tied to other attributes and not in isolation, it becomes difficult for a system to differentiate each individual attribute and isolate the meaning. To overcome this problem, the paper introduces the technique of generating targeted negatives by editing single visual attributes within images which corresponds to the technique that was explained as future work in state of the art.

Beginning with an image paired with a detailed description, they alter a single visual attribute in the image, after which the image no longer matches the original audio description. Such edited training examples are used to guide the system to learn the correspondence between individual visual attributes and relevant audio words. In order to edit these images, they use a technique that was learned in prior work: “GAN Dissection: Visualizing and Understanding Generative Adversarial Networks”.

3.3 Ablating Visual Units:

In order to edit some features of generated images, the same method is followed with previously discussed paper:

3.3.1 Comparing Units Across Datasets, Layers and Models:

Each unit is gathered by using abstractions with emergence of individual unit object detectors, interpretable units for different scene categories, interpretable units for different network layers and interpretable units for different GAN models.

3.3.2 Diagnosing and Improving GANS

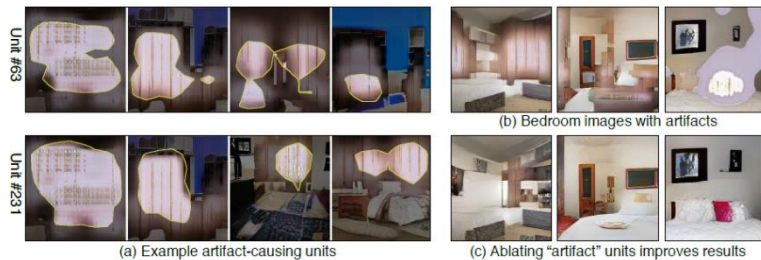


Figure 8: (a) We show two example units that are responsible for visual artifacts in GAN results. There are 20 units in total. By ablating these units, we can fix the artifacts in (b) and significantly improve the visual quality as shown in (c).

3.3.3 Locating Causal Units with Ablation:



A variety of specific object types can also be removed from GAN output by ablating a set of units in a GAN. They find that, by turning off these small sets of units, most of the output of people, curtains, and windows can be removed from the generated scenes.

3.3.4 Characterizing Contextual Relationships via Insertion:

Using the learned information in this part, each object can be added if the context is appropriate and the position is plausible. For example, it is not possible to trigger a door in the sky or on trees. Even if we wanted to add a door to sky, it will be rejected by GAN because of the enforced relationships between objects.

3.4 Generating Image Edits and Editing Specific Attributes:

With the knowledge learned from “GAN Dissection: Visualizing and Understanding Generative Adversarial Networks”, a GAN generator contains different sets of convolutional filters

that specialize in generating different attributes and objects. The activations of these convolutional filters can be modified to change, add, and remove objects in the output image. In this paper, they use this technique to modify certain attributes in particular objects.

3.5 Learning Words by Drawing Images:

This paper uses the technique from previous work [4] which learns concepts from raw audio by using negative matchmaps. After creating these negative matchmaps, they add these results to training process in order to improve the model’s ability to distinguish and isolate particular attributes. This process is done as a control mechanism and done many times.

The training system has following steps:

1. Training the basic system without any edited examples.

In [4], they introduced audio-visual “matchmap” neural networks which are capable of directly learning the semantic correspondences between speech frames and image pixels without the need for annotated training data in either modality. They applied these networks for semantic image/spoken caption search, speech-prompted object localization, audio-visual clustering and concept discovery, and real-time,

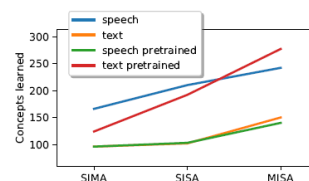
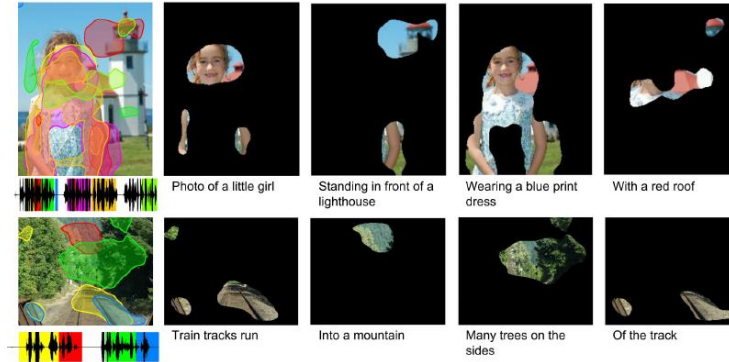


Fig. 7: We study the number of concepts learned by the different networks with different losses, and we find it is consistently lower for SIMA and higher for MISA.

speech-driven, semantic highlighting. Using the same technique, the system is trained with unedited examples.



2. Using edited examples in which neurons are randomly ablated. This improves the internal representations of objects and attributes.

In order to ablate a random or specific object, they use the techniques explained in 3.3 Ablating Visual Units part [3] Using the edited images, negative matchmaps are found with multiplication of the original audial domain matrix.

3. Partitioning the space of audio-visual representation by clustering units according to co-occurrences. Each of these clusters correspond to different concepts present in the captions, such as colors, sizes, shapes, etc. They use these clusters to generate edited examples that are tailored to the mentioned concepts.

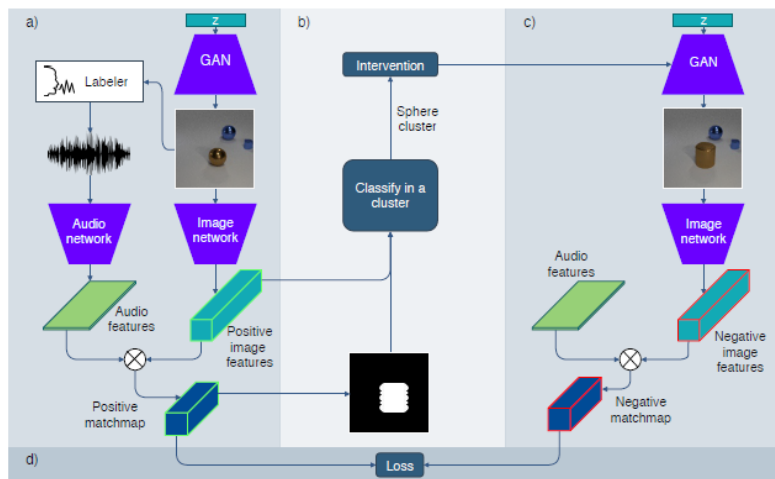
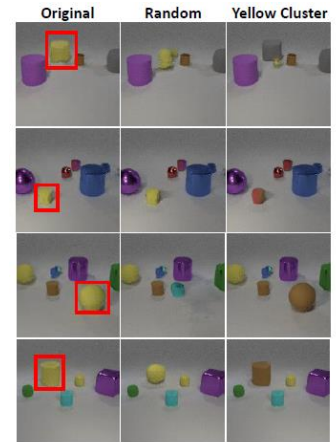


Figure 4: **Intervention schematic:** a) Basic model, where the original image and audio features are computed, as well as their matchmap. b) Clustering: highly activated image features are classified into a cluster, and an intervention is computed to generate an edited example. c) Generation of the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

The selection of negative examples has long been an important topic in computer vision. Previous work [4] proposed using random samples or mismatched samples that the network classifies closest to the threshold. These methods assume a closed set of images from which to choose, but none entertain the

possibility of *creating* mismatched samples to aid learning. The paper uses interventions in the GAN to generate ideal counterexamples to pair with each positive image. The edited negative examples will improve performance on the most confusing cases.

After creating these randomly edited images, their matchmaps are computed and the system is trained with their data. While some of the modified images are informative negatives that falsify a single word in the caption, others may be too similar to the positive image to correspond to any caption change; and others may be different enough to correspond to differences in many words. If the negative image is not informative enough, then it tries to improve the quality of the image.

3.6 Clustering:

A similar technique with [4] is used for clustering. To build the concept clusters, they process the full training set through their audio-visual model and observe the audiovisual features that activate in each training pair. They compute a co-occurrence matrix of the binarized features to measure how much every pair of neurons co-activate. This enables us to partition the neuron space using a dendrogram, grouping units with high co-occurrence. This clustering in the unit space induces a semantic clustering in the shared embedding space of the matchmap. The image clusters are coherent and usually represent a concept in the image space, while the audio usually represents one or a few spoken words with the same meaning.

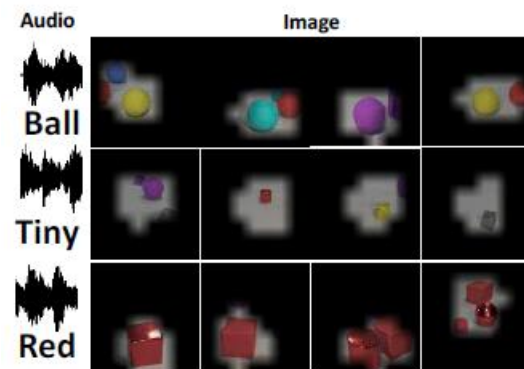


Figure 5: We show three examples of clusters learned by our model, represented by the images that mostly activate each cluster. We represent the audio-cluster with text for clarity, but all the learning is done in the audio domain. As shown, the system is able to learn color, shape and size.

4. Comparison:

Categories:

1. Problem
2. Feature Extraction Technique
3. Dataset
4. Evaluation
5. Benchmark
6. Performance

Papers to Compare:

1. Learning Words by Drawing Images [5]
2. Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input [4]
3. Unsupervised Learning of Spoken Language with Visual Context[6]
4. Learning word-like units from joint audio-visual analysis[7]

Problem:

Learning Words by Drawing Images	Learning the correspondence between spoken words and abstract visual attributes, from a dataset of spoken descriptions of images.
Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input	Exploring neural network models that learn to associate segments of spoken audio captions with the semantically relevant portions of natural images that they refer to.
Unsupervised Learning of Spoken Language with Visual Context	Presenting a deep neural network model capable of rudimentary spoken language acquisition using untranscribed audio training data, whose only supervision comes in the form of contextually relevant visual images. The networks learn visual associations directly from the data, without the use of conventional speech recognition, text transcriptions, or any expert linguistic knowledge whatsoever.
Learning word-like units from joint audio-visual analysis	Automatically discovering words and other elements of linguistic structure from continuous speech has been a longstanding goal in computational linguistics, cognitive science, and other speech processing fields and the purpose of this paper is discovering word-like acoustic units in the continuous speech signal and grounding them to semantically relevant image regions.

Feature Extraction Technique:

Learning Words by Drawing Images	First, they train the basic system without any edited examples. Second, they use edited examples in which neurons are randomly ablated. This improves the internal representations of objects and attributes. Finally, they partition the space of audio-visual representation by clustering units according to co-occurrences. Each of these clusters correspond to different concepts present in the captions, such as colors, sizes, shapes, etc. They use these clusters to generate edited examples that are tailored to the mentioned concepts. With this process, the visual attributes are learned with their audio description.
Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input	They introduced audio-visual “matchmap” neural networks which are capable of directly learning the semantic correspondences between speech frames and image pixels without the need for annotated training data in either modality. They applied these networks for semantic image/spoken caption search, speech-prompted object localization, audio-visual clustering and concept discovery, and real-time, speech-driven, semantic highlighting.
Unsupervised Learning of Spoken Language with Visual Context	They present a deep neural network architecture capable of learning associations between natural image scenes and accompanying free-form spoken audio captions. The networks do not rely on any form of conventional speech recognition, text transcriptions, or expert linguistic knowledge, but are able to learn to recognize semantically meaningful words and phrases at the spectral feature level.
Learning word-like units from joint audio-visual analysis	They demonstrate that a neural network trained to associate images with the waveforms representing their spoken audio captions can successfully be applied to discover and cluster acoustic patterns representing words or short phrases in untranscribed audio data. An analogous procedure can be applied to visual images to discover visual patterns, and then the two modalities can be linked, allowing the network to learn, for example, that spoken instances of the word “train” are

	associated with image regions containing trains. This is done without the use of a conventional automatic speech recognition system and zero text transcriptions, and therefore is completely agnostic to the language in which the captions are spoken. Further, this is done in $O(n)$ time with respect to the number of image/caption pairs, whereas previous state-of-the-art acoustic pattern discovery algorithms which leveraged acoustic data alone run in $O(n^2)$ time.
--	--

Dataset:

Learning Words by Drawing Images	CLEVR dataset: a synthesized visual dataset consisting of scenes of simple objects with composable color, shape and material attributes. CLEVR has been used to study abstract visual reasoning, question answering, and model interpretability. They use CLEVR as a highly controlled visual domain in which compositional attributes are clear, and that a GAN can learn to draw well.
Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input	For training their models, they use the Places Audio Caption dataset. This dataset contains approximately 200,000 recordings collected via Amazon Mechanical Turk of people verbally describing the content of images from the Places 205 image dataset. They augment this dataset by collecting an additional 200,000 captions, resulting in a grand total of 402,385 image/caption pairs for training and a held-out set of 1,000 additional pairs for validation.
Unsupervised Learning of Spoken Language with Visual Context	Since they desire spontaneously spoken audio captions, they collected a new corpus of captions for the Places205 dataset. Places205 contains over 2.5 million images categorized into 205 different scene classes, providing a rich variety of object types in many different contexts. To collect audio captions, they turned to Amazon's Mechanical Turk, an online service which allows requesters to post "Human Intelligence Tasks".
Learning word-like units from joint audio-visual analysis	They employ a corpus of over 200,000 spoken captions for images taken from the Places205 dataset, corresponding to over 522 hours of speech data. The captions were collected using Amazon's Mechanical Turk service, in which workers were shown images and asked to describe them verbally in a free-form manner.

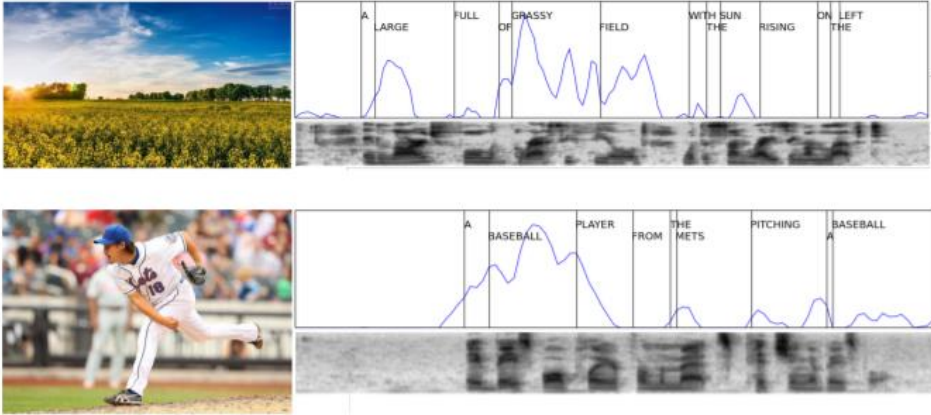
Evaluation:

Learning Words by Drawing Images	For each attribute, they produce pairs of images, one containing the attribute and another without the attribute. They then create an input for the audio network containing the isolated attribute to be evaluated in the form of a spoken word. They compute the accuracy of the system on selecting the image with the attribute against the image without the attribute. In addition to the semantic test, they also show the recall on random negatives, where 500 image-audio pairs of a held-out test set are passed through the network, and the retrieval recalls from audio to image and from image to audio are computed.
Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input	They weight the similarity with w_i , which is proportional to intersection over union of the pixels for that class into the masked region of the image. Using this metric, they then assign one value per dimension, which measures how well both the audio network and the image network agree on that particular concept. Anecdotally, we found $c > 0.6$ to be a good indicator that a concept has been learned. But they did not evaluate their process after all. They don't mention how many times they followed this procedure.

Unsupervised Learning of Spoken Language with Visual Context	They subsample a validation set of 1,000 image/caption pairs from the testing set. To perform image search given a caption, they keep the caption fixed and use their model to compute the similarity score between the caption and each of the 1,000 images in the validation set. Image annotation works similarly, but instead the image is kept fixed and the network is tasked with finding the caption which best fits the image. They experimented with many different variations on our model architecture, including varying the number of hidden units, number of layers, filter sizes, embedding dimension, and embedding normalization schemes. They found that an embedding dimension of $d = 1024$ worked well, and that normalizing the caption embeddings prior to the similarity score computation helped. When only the acoustic embedding vectors were L2 normalized, we saw a consistent increase in performance. However, when the image embeddings were also L2 normalized (equivalent to replacing the dot product similarity with a cosine similarity), the recall scores suffered.
Learning word-like units from joint audio-visual analysis	Their strategy is to forcealign the Google recognition hypothesis text to the audio, and then assign a label string to each acoustic segment based upon which words it overlaps in time. The alignments are created with the help of a speech recognizer based on the standard WSJ recipe and trained using the Google ASR hypothesis as a proxy for the transcriptions. Any word whose duration is overlapped 30% or more by the acoustic segment is included in the label string for the segment. They then employ a majority vote scheme to derive the overall cluster labels. When computing the purity of a cluster, they count a cluster member as matching the cluster label as long as the overall cluster label appears in the member's label string.

Benchmark:

Learning Words by Drawing Images	For evaluation, they compare many different training methods. DaveNet: The training procedure where random negatives are used. Hard Negatives: The negative image and audio are selected as the sample in the minibatch with highest loss. Random Edited Examples: The examples produced by random ablation in of the hidden representation in the GAN. Targeted Edited Examples: The examples produced according to the semantics of the object intervened. Hard Negatives + Random Edits: They combine the random edited examples with the hard negative loss. In training, they use the hardest negative of both methods. Hard Negatives + Targeted Edits: They combine the targeted edited examples with the hard negative loss.
Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input	Because there are a very large number of different words appearing in the speech, and no one-to-one mapping between words and ADE20k objects exists, they manually define a set of 100 word-object pairings. They choose commonly occurring pairs that are unambiguous, such as the word "building" and object "building," the word "man" and the "person" object, etc. For each word-object pair, they compute an average IoU score across all instances of the word-object pair appearing together in an ADE20k image and its associated caption. They then average these scores across all 100 word-object pairs and report results for each model type. They also report the IoU scores for the ASR text-based baseline models.
Unsupervised Learning of Spoken Language with Visual Context	They time aligned the recognition hypothesis to the spectrogram, allowing us to see exactly which words overlapped the audio regions that were highly similar to the image. Figure below displays several examples of these similarity curves along with the overlaid recognition text. In the majority of cases, the regions of the spectrogram which have the highest similarity to the accompanying image turn out to be highly

	<p>informative words or phrases, often making explicit references to the salient objects in the image scenes. This suggests that this network is in fact learning to recognize audio patterns consistent with words using zero linguistic supervision.</p> 
Learning word-like units from joint audio-visual analysis	<p>They first build a list of the 1,000 WordNet synsets associated with the ILSVRC2012 classes. They then take the set of unique majority-vote labels associated with the discovered word clusters for $k = 500$, filtered by setting a threshold on their variance ($\sigma^2 \leq 0.65$) so as to get rid of garbage clusters, leaving 197 unique acoustic cluster labels. They then look up each cluster label in WordNet, and compare all noun senses of the label to every ILSVRC2012 class synset according to the path similarity measure. This measure describes the distance between two synsets in a hyponym/hypernym hierarchy, where a score of 1 represents identity and lower scores indicate less similarity. They retain the highest score between any sense of the cluster label and any ILSVRC2012 synset. Of the 197 unique cluster labels, only 16 had a distance of 1 from any ILSVRC12 class, which would indicate an exact match. A path similarity of 0.5 indicates one degree of separation in the hyponym/hypernym hierarchy</p>

Performance:

Learning Words by Drawing Images

		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
Target 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.

The accuracy of this method and the baselines for the semantic test, both in the human captioned dataset and the synthetic generated dataset are reported. As expected, the basic DaveNet model performs poorly in this test, suggesting that the system is not able to learn particular isolated concepts. Furthermore, the models using targeted edits have a better ability on predicting particular attributes, which reinforces

	<p>the idea that using edited examples for training increases the model understanding of isolated attributes. Finally, human models focus more its attention on discriminating color as they are more mentioned in the audio captions. However, when using the synthetic captions, where attributes are evenly distributed, performance drops.</p>																								
<p>Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input</p>	<p>Since this paper does not have a proper evaluation part, it explains the process over an example. The numerical values for six concept pairs are shown in Figure 6. We see how neurons with higher value are cleaner and more related with its counterpart. The bottom right neuron shows an example of low concept value, where the audio word is “rock” but the neuron images show mountains in general.</p> <table border="1"> <thead> <tr> <th>Word</th> <th>Images</th> <th>Concept Value</th> <th>Word</th> <th>Images</th> <th>Concept Value</th> </tr> </thead> <tbody> <tr> <td>Building</td> <td></td> <td>0.78</td> <td>Table</td> <td></td> <td>0.65</td> </tr> <tr> <td>Furniture</td> <td></td> <td>0.77</td> <td>Flower</td> <td></td> <td>0.65</td> </tr> <tr> <td>Water</td> <td></td> <td>0.72</td> <td>Rock</td> <td></td> <td>0.51</td> </tr> </tbody> </table> <p>Fig. 6: Matching the most activated images in the image network and the activated words in the audio network we can establish pairs of image-word, as shown in the figure. We also define a concept value, which captures the agreement between both networks and ranges from 0 (no agreement) to 1 (full agreement).</p>	Word	Images	Concept Value	Word	Images	Concept Value	Building		0.78	Table		0.65	Furniture		0.77	Flower		0.65	Water		0.72	Rock		0.51
Word	Images	Concept Value	Word	Images	Concept Value																				
Building		0.78	Table		0.65																				
Furniture		0.77	Flower		0.65																				
Water		0.72	Rock		0.51																				
<p>Unsupervised Learning of Spoken Language with Visual Context</p>	<div> </div> <table border="1"> <thead> <tr> <th>System</th> <th>P@N</th> <th>EER</th> </tr> </thead> <tbody> <tr> <td>MFCC baseline</td> <td>0.50</td> <td>0.127</td> </tr> <tr> <td>[9]</td> <td>0.53</td> <td>0.164</td> </tr> <tr> <td>[12]</td> <td>0.63</td> <td>0.169</td> </tr> <tr> <td>This work</td> <td>0.62</td> <td>0.049</td> </tr> </tbody> </table> <p>Table 2: Precision @ N and equal error rate (EER) results for the TIMIT keyword spotting task. The 10 keywords used for the task were: <i>development, organizations, money, age, artists, surface, warm, year, problem, children</i>.</p> <p>Figure 5: t-SNE visualization in 2 dimensions for 1645 spoken instances of 14 different word types taken from the development data.</p> <p>To further examine the high-level acoustic representations learned by their networks, they extracted spectrograms for 1645 instances of 14 different ground truth words from the development set by force aligning the Google recognizer hypotheses to the audio. They did a forward pass of each of these individual words through the audio branch of their network, leaving an embedding vector for each spoken word instance. They performed t-SNE analysis on these points, shown in Figure 5. They observed that the</p>	System	P@N	EER	MFCC baseline	0.50	0.127	[9]	0.53	0.164	[12]	0.63	0.169	This work	0.62	0.049									
System	P@N	EER																							
MFCC baseline	0.50	0.127																							
[9]	0.53	0.164																							
[12]	0.63	0.169																							
This work	0.62	0.049																							

	points form pure clusters, indicating that the top-level activations of the audio network carry information which is discriminative across different words			
Learning word-like units from joint audio-visual analysis	Cluster	ILSVRC synset	Similarity	They demonstrate the success of their methodology on a large-scale dataset of over 214,000 image/caption pairs comprising over 522 hours of spoken audio data. They have shown that the shared multimodal embedding space learned by their model is discriminative not only across visual object categories, but also acoustically and semantically across spoken words.
	snow	cliff.n.01	0.14	
	desert	cliff.n.01	0.12	
	kitchen	patio.n.01	0.25	
	restaurant	restaurant.n.01	1.00	
	mountain	alp.n.01	0.50	
	black	pool_table.n.01	0.25	
	skyscraper	greenhouse.n.01	0.33	
	bridge	steel_arch_bridge.n.01	0.50	
	tree	daisy.n.01	0.14	
	castle	castle.n.02	1.00	
	ocean	cliff.n.01	0.14	
	table	desk.n.01	0.50	
	windmill	cash_machine.n.01	0.20	
	window	screen.n.03	0.33	
	river	cliff.n.01	0.12	
	water	menu.n.02	0.25	
	beach	cliff.n.01	0.33	
	flower	daisy.n.01	0.50	
	wall	cliff.n.01	0.33	
	sky	cliff.n.01	0.11	
	street	swing.n.02	0.14	
	golf course	swing.n.02	0.17	
	field	cliff.n.01	0.20	
	lighthouse	beacon.n.03	1.00	
	forest	cliff.n.01	0.20	
	church	church.n.02	1.00	
	people	street_sign.n.01	0.17	
	baseball	baseball.n.02	1.00	
	car	freight_car.n.01	0.50	
	shower	swing.n.02	0.17	
	people walking	(none)	0.00	
	wooden	(none)	0.00	
	rock	toilet_tissue.n.01	0.20	
	night	street_sign.n.01	0.14	
	station	swing.n.02	0.20	
	chair	barber_chair.n.01	0.50	
	building	greenhouse.n.01	0.50	
	city	cliff.n.01	0.12	
	white	jean.n.01	0.33	
	sunset	street_sign.n.01	0.11	
	Table 5: The 40 lowest variance, uniquely-labeled acoustic clusters paired with their most similar ILSVRC2012 synset.			

5. State of the art:

Based on the knowledge I have gained while reading these papers, I see that each paper has added a new feature to the existing one. The paper I have presented in class is: Learning Words by Drawing Images [7].

This paper represents a learning framework which can learn visual attributes and create relation with its corresponding audio by drawing. This framework uses negative -edited images for training the model. In order to make edited images, some methods learned in previous papers [9] are used. These changes can be made on a random object in the image or can also be made on a salient concept of the image. For finding the salient concept, dot product of audio and visual network is computed. Then using triplet loss, clusters

are made. After feeding the negative generator with the cluster to be edited, a negative image is created. Then this image and the original audio creates a negative matchmap and it is used in triplet loss again. This procedure creates training images and using this feedback mechanism, visual attributes such as shape, color, size and material can be learned by drawing. The experiments are made using various models and the mean accuracy of detecting particular attributes is very high.

Since each paper actually does a different work, I can say that I would choose one of these ways according to the goal of the project. If I want to concentrate on learning visual attributes and their acoustic descriptions, I would choose [5]. If the goal is finding correspondence between any object and raw audio and using CNNs and size or shape doesn't matter, I would choose paper [4]. Because after each paper the process has improved. Each paper is done with the help of the previous ones and it adds a new feature every time. We can see that the similarity function helps us understand the general concept of what the audio tells.

6. Discussion and Conclusion:

In conclusion, there are many papers which focus on the correspondence between a audio domain and visual domain. Since babies learn each word in a noisy and unsupervised way, people have lead their investigations this way. We saw that after each paper, previously learned concepts are used in order to add a new feature. First CNNs were used and only the objects could be recognized with the audio. Then, with the increase in interest about GANs, this idea has formed accordingly. With the help of paper[9], it is learned that an object can be ablated from an image. With this editing technique negative matchmaps are created and the framework is trained using them. Then using triplet loss, the neural network is trained again. Even though matchmap operations have always been made in four papers, triplet loss was not used in them. We can say that the base has been the same but the contributions are made with each paper.

7. References:

- [1] Yash Srivastava, Vaishnav Murali, Shiv Ram Dubey and Snehasis Mukherjee Computer Vision Group, Visual Question Answering using Deep Learning: A Survey and Performance Analysis, 2019
- [2] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, Ross Girshick, CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, 2016
- [3] D. Bau, J.-Y. Zhu, H. Strobelt, B. Zhou, J. B. Tenenbaum, W. T. Freeman, and A. Torralba. Visualizing and understanding generative adversarial networks. In *International Conference on Learning Representations*, 2019.
- [4] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass. Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input. In *European Conference on Computer Vision*, 2018.
- [5] D'ídac Surís, Adrià Recasens, David Bau, David Harwath, James Glass, Antonio Torralba, Learning Words by Drawing Images, CVPR 2019
- [6] D. Harwath, A. Torralba, and J. R. Glass. Unsupervised learning of spoken language with visual context. In *NIPS*, 2016.
- [7] D. Harwath and J. Glass. Learning word-like units from joint audio-visual analysis. In *ACL*, 2017.