Learning English with Peppa Pig

Anonymous TACL submission

Abstract

Attempts to computationally simulate the acquisition of spoken language via grounding in perception have a long tradition but have gained momentum since around 2015. Current neural approaches exploit associations between the spoken and visual modality and learn to represent speech and visual data in a joint vector space. A major unresolved issue from the point of ecological validity is the training data, typically consisting of images or videos paired with spoken descriptions of what is depicted. Such a setup guarantees an unrealistically strong correlation between speech and the visual world. In the real world the coupling between the linguistic and the visual is loose, and often contains confounds in the form of correlations with non-semantic aspects of the speech signal. The current study is a first step towards simulating a naturalistic grounding scenario by using a dataset based on the children's cartoon *Peppa Pig*. We train a simple bi-modal architecture on the portion of the data consisting of naturalistic dialog between characters, and evaluate on segments containing descriptive narrations. Despite the weak and confounded signal in this training data our model succeeds at learning aspects of the visual semantics of spoken language.

1 Introduction

Attempts to model or simulate the acquisition of spoken language via grounding in the visual modality date to the beginning of this century (Roy and Pentland, 2002) but have gained momentum since 2014 with the revival of neural networks (e.g. Synnaeve et al., 2014; Harwath and Glass, 2015; Harwath et al., 2016; Chrupała et al., 2017; Alishahi et al., 2017; Harwath et al., 2018; Merkx et al., 2019; Havard et al., 2019; Rouditchenko et al., 2020; Khorrami and Räsänen, 2021; Peng and Harwath, 2021). Current approaches work well enough from an applied point of view but leave much to be desired as regards ecological validity. Training data typically consist of images or videos paired with spoken descriptions of the scene depicted. The type of input that a child faces when learning a language is much more challenging. Firstly, speech is only loosely coupled with the visual modality. Secondly in addition to correlations between the visual scenes and the *meaning* of spoken utterances, there are also correlations with non-semantic aspects of the speech signal, such as the voice of specific characters, as well as with non-speech ambient sounds. Always it is plausible that such non-semantic correlations can sometimes be useful to the learner in the general endeavor of making sense of the world, for the specific task of learning the semantics of linguistic units they are likely more often an obstacle, as they make it harder to zoom in on the meaning bearing aspects of the audio signal.

In the current study we make a first step towards simulating the acquisition of language via grounding in perception in a naturalistic scenario. We use the well-know children's cartoon *Peppa Pig* as a case study as a source of training and evaluation data. Compared to commonly used video datasets, this data has a number of interesting characteristics. The visual modality is very schematic, and the language is also simple in terms of vocabulary size and syntactic complexity. Crucially, however, most of the speech in the videos consists of naturalistic dialogs between the characters. The utterances are only loosely and noisily correlated to the scenes and actions depicted in the videos.

GC: Can we support this assertion in a quantitative way?

This choice of data thus allows us to directly address the ecological limitations of the current approaches. In addition, the cartoon videos also contain comments interjected by the narrator. We use these for a evaluation a source of more descriptive and less noisy data which allows us to measure performance while controlling for speaker characteristics. Our contributions are the following:

- We implement a simple bi-modal architecture which learns spoken language embeddings from videos;
- We evaluate model performance in terms of video fragment retrieval and additionally design controlled evaluation protocols inspired by the intermodal preferential looking paradigm (Hirsh-Pasek and Golinkoff, 1996);
- We carry out ablations of model components in order to understand the effects of pre-training for the audio and video encoders, the role of temporal information, and of segmentation strategies while training.

We show that despite the challenges of our naturalistic training data our model succeeds at learning aspects the visual semantics of spoken language. Our findings include the fact that temporal information contributes substantially to video modeling, and that unsupervised pre-training of the audio encoder is key the best performance, but that even the model trained completely from scratch about 10 hours of cartoon data performs substantially above chance.

2 Related work

3 Method

The main focus of this study is on the data and evaluation. We thus keep the components of our architecture simple, and follow established modeling practice whenever possible.

3.1 Dataset

The dataset consists of the set of videos of the English-language version of *Peppa Pig*. In addition to the raw videos we also use the annotation created by Papasarantopoulos and Cohen (2021).

These annotations feature written transcriptions aligned with the audio as well as segmentation into *dialog* and *narration*.[1] Dialogs are the parts spoken by the characters, while narrations are comments inserted by the narrator, which are more descriptive in nature. All the narration segments are uttered by the same voice actor. We use the dialogs for training the model, and set aside the narrations for evaluation purposes only. A small portion of the dialog data is also used for validation. Specifically, we use dialog from episodes 1–196 for training, and 197–209 for validation. We set aside narrations from episodes 1–104 for validation and 105–209 for testing.

[1] It should be noted that the quality of the alignment and segmentation in the original dataset is variable. In cases where exact alignment is needed, such as for word-level analyses, we re-align the transcriptions using `github.com/lowerquality/gentle`.

3.2 Preprocessing

Our model is trained to discriminate positive video-audio pairs from negative ones. The positive pairs are those that are temporally coincident in the original video file. In order to generate these training items we need to split the videos into fragments. For segmenting data for training, we *do not* use word or sentence-level subtitle alignment in order to make the setting naturalistic. Processing long segments of video and audio is not tractable on commodity GPU hardware, and we thus segment the data into brief snippets roughly comparable in length to the duration a short sentence or a phrase. We use the following two segmentation strategies:

Fixed Using this approach we simply split sections into fixed-length non-overlapping fragments of 2.3 second duration. This length is close to the mean duration of audio aligned to a single line of subtitles.

Jitter In this approach the mean duration of the segments is the same (2.3 seconds) but we randomly vary the length of the video, and independently, of the corresponding audio around this average duration. This means that (i) the segments can be partially overlapping and (ii) the video and the audio it is paired with are normally of different length. Specifically we sample the fragment duration d from the following distribution:

$$d \sim \min(6, \max(0.05, D + \mathcal{N}(\mu, \sigma^2))), \quad (1)$$

with $D = 2.3$ $\mu = 0$ and $\sigma^2 = 0.5$.

The video is subsampled to 10 frames per second, and to 180×100 resolution. The audio is converted to mono by averaging the two channels and the raw waveform is used as input.

For evaluation we have a number of different conditions and evaluation metrics described in detail in Section 3.3 and in some of these conditions we use the subtitles to guide segmentation. Table 1 shows the basic statistics of the training and validation splits.

GC: These numbers seem inconsistent between triplet/no triplet versions.

3.3 Evaluation

The most common approach to evaluation for visually grounded models trained on spoken image captions is caption-to-image retrieval (often combined with image-to-caption retrieval): in fact this technique has been carried over from text-based image-caption modeling. With the standard spoken caption dataset this approach is unproblematic since the content of the captions is not correlated with extra-linguistic clues in the speech signal, such as speaker identity (since speakers are randomly assigned to captions) or non-speech environmental sounds. In such an artificial setting, a retrieval metric measures the ability of the model to match spoken utterances to images based on their semantic content. This not the case for the *Peppa Pig* dataset: here we can expect that when a video segment depicts a particular character (e.g. George) then the audio in this segment is more likely to contain utterances spoken by the voice actor playing George. George has a favorite toy dinosaur: when this toy appears in a video segment we can likewise expect higher than random chance of George's voice in the audio. Due to these factors, in a naive retrieval setting, a model could obtain a high score by mostly capturing these non-linguistic correlations.

In order to control for these factors we leverage the narrator speech in the videos. These utterances are always spoken by the same actor, so speaker identity cannot be used as a clue for matching video and audio. Furthermore, the narration segments are akin to video captions in that they tend to describe what is happening in the video and thus their semantic content is more strongly correlated with the content of the video than in the case of the dialog, which is also a desirable feature for the purposes of system evaluation.

3.3.1 Retrieval

For the retrieval evaluation, as for training, we use the FIXED and JITTER segmentation strategies. We encode each audio clip in a candidate set sampled from the validation (or test) data using the speech encoder part of the model; we encode each video clip using the video encoder. We then measure cosine similarity between the audio clip and all the video clips. If the video clip corresponding to the audio is among the n most similar video clips, we count that as a success. The proportion of successes across all audio clips gives us the retrieval metric known as recall@n: specifically in this paper we focus on $n = 10$. We set the candidate set size to 100, and thus the random baseline for the recall@10 is 10%. In order to quantify uncertainty in this evaluation due to the test data we repeat this procedure 500 times with randomly sampled candidate sets and visualize the score distribution.

3.3.2 Triplets

Retrieval metrics such as recall@10 have some disadvantages. Firstly the absolute value of this metric may be hard to interpret as it depends on the size of the candidate set. Secondly, if we wanted to compare model performance with human performance, we could not feasibly ask human participants to provide the quadratic number of audio-video similarity judgments needed. For these reasons we evaluate model performance using the following simplified, controlled scenario: We extract clips aligned to a single subtitle line, group them by length, and for each pair of same-length video clips[2], we extract the audio from one of them (selected at random) – this is our *anchor*. The video clip from which the anchor was taken is the *positive* one, which the other video clip is the *negative* one. This triplet of stimuli form a single test item. We use the model's audio encoder to encode the anchor, and the video encoder to encode both video clips. We then check whether anchor is more similar to the positive or negative clips in terms of cosine similarity. More precisely, *triplet accuracy* is the mean over all triplets of the following quantity:

$$\frac{\mathrm{signum}(\mathrm{cosine}(A, P) - \mathrm{cosine}(A, N)) + 1}{2} \tag{2}$$

with A being the anchor, P positive and N negative. The triplet accracy metric is inspired by the ABX score of Schatz (2016). For triplet accuracy, regardless of the specific set of test items, we expect random-guessing performance to be at 0.5, and perfect performance to be 1.0. For this metric we also quantify uncertainty by resampling the triplets 500 times from the dataset, and display the score distribution.

3.3.3 Targeted Triplets

GC: Should we maybe use a more informative name such as

[2] To keep test items independent, the pairing of video clips is done such that each clip only occurs as a member of a single triplet.

Split	Type	Triplet	Size (h)	Items	Mean length (s)
train	dialog	No	9.83	11058	3.20
val	dialog	No	0.65	729	3.20
val	narration	No	0.80	897	3.20
test	narration	No	0.51	570	3.20
val	dialog	Yes	0.65	828	2.81
val	narration	Yes	0.92	1492	2.21
test	narration	Yes	0.71	1052	2.44

Table 1: Dataset statistics. For the triplet condition, videos are split such that each segment corresponds to a line of subtitles. For the non-triplet condition, videos are split into 3.2s segments.

Inspired by intermodal 2-alternative forced choice (2AFC) paradigms in child language acquisition (Hirsh-Pasek and Golinkoff, 1996), we design test trials that test the model's acquisition of grounded semantics under more controlled circumstances. The paradigm has been used in language acquisition research to evaluate children's early linguistic knowledge (e.g., Noble et al., 2011; Bergelson and Swingley, 2012), by testing whether they can distinguish a matching (target) visual referent from a foil (distractor) referent when prompted with a word or sentence.

GC: Maybe it's better to use this as motivation for both the general triplet and the targeted triplet setups.

In our case, the test can be seen as a special case of the triplets evaluation as described in the previous paragraph. Here, we aim to assess the model's acquisition of the semantics of commonly occurring words. Therefore, the *targeted* approach considers, in contrast to the general triplets evaluation, always pairs of triplets with *minimal differences* regarding one word in the transcripts of the anchor audios (e.g., *Peppa loves jumping* and *George loves jumping* can be used to test whether the model can discriminate the target word *Peppa* from the distractor word *George*).

We search the transcripts of the validation data for phrases with minimal differences with respect to the most commonly occurring nouns, verbs, and adjectives. Details on this search procedure can be found in Appendix A.2. Based on each pair of phrases, we create two counter-balanced test trials, an example and a corresponding counter-example as depicted in Figure 1. Here, the anchor A_x of the example triplet is the audio of *Peppa loves jumping* (a_p), the positive video P_x is the corresponding video (v_p) and the negative video N_x is the video corresponding to *George loves jumping* (v_g): $(A_x, P_x, N_x) = (a_p, v_p, v_g)$. In the counter-example triplet, the anchor A_y is the audio of *George loves jumping*, and the positive and negative video are flipped: $(A_y, P_y, N_y) = (a_g, v_g, v_p)$. In this way, we control the evaluation for linguistic biases in the dataset and ensure that a single-modality model that only considers the audio performs at chance (see also Nikolaus and Fourtassi, 2021).

We can't use the appendix for the TACL submission.

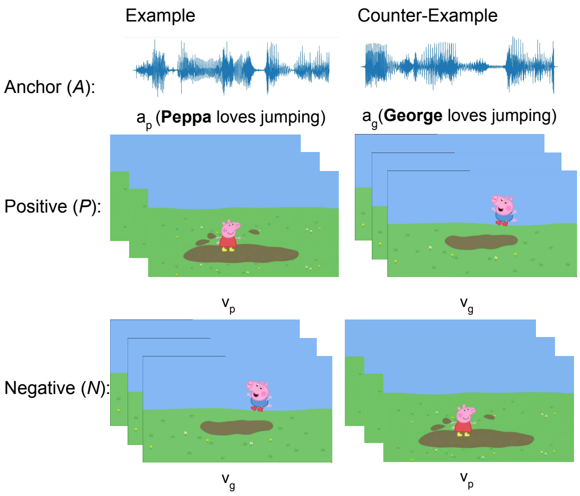

Figure 1: Targeted Triplets Evaluation

GC: I find the notation here a bit confusing, especially the overloaded one-letter names. Maybe simplify.

We measure overall targeted triplet accuracy using Equation (2). Additionally, we report per-word accuracy by calculating the triplet accuracy for all triplets that contain a given word (e.g. *Peppa*) either as target or distractor word, i.e. cases in which the model needs to succeed in either choosing a video containing the given word (the example triplet in Figure 1) or rejecting a video containing the given word (the counter-example triplet in Figure 1). We report accuracy for all words (nouns and verbs) for which we found at least 100 example-counterexample pairs of triplets. There were not enough examples for adjectives in the dataset to perform a targeted triplets evaluation on them.

3.4 Model

We adapt the high-level modeling approach from work on spoken image-caption data (Harwath et al., 2016; Chrupała et al., 2017): our objective function is based on a triplet-loss with margin which encourages the matching audio and video clip to be projected nearby in the embedding space, and mis-matching audio and video clips to be far away:

$$\ell = \sum_{av} \left[\sum_{a'} \max(0, S_{a'v} - S_{av} + \alpha) + \sum_{v'} \max(0, S_{av'} - S_{av} + \alpha) \right] \tag{3}$$

where α is a margin, S_{av} is a similarity score between a matching audio-video clip pair, and $S_{a'v}$ and $S_{av'}$ denote similarity scores between mismatched pairs, i.e. negative examples from the current batch. Our heuristic to generate positive and negative examples is very simple: namely we consider the example positive if the audio is exactly aligned with a video clip in our data. All other pairs of audio-video clips are considered negative.

The audio encoder portion of the model consists of a `small wav2vec2` model (Baevski et al., 2020) pretrained in an self-supervised fashion, with supervised fine tuning.[3] During training, we keep the feature extractor and the bottom $K = 3$ transformer layers of this encoder frozen. Its output is pooled across time using an attention mechanism with dimensionwise weights (Merkx et al., 2019):

$$\begin{aligned} \mathbf{A} &= \mathrm{softmax}_t \left(\mathrm{MLP}(\mathbf{X}) \right) \\ \mathbf{z} &= \sum_t \left(\mathbf{A}_t \odot \mathbf{X}_t \right), \end{aligned} \tag{4}$$

where $\mathbf{X}$ is the tensor with the encoder output vectors for each time-step: an MLP followed by a time-wise softmax is used to compute an attention weight for each time step and for each dimension. The pooling is followed by a linear projection and L_2 normalization.

As a video encoder we use the 18-layer ResNet (2+1)D architecture (Tran et al., 2018) pretrained on the action recognition dataset Kinetics-400 (Kay et al., 2017). The pretrained model is available via Pytorch.[4] The output of this module is aggregated using the attention mechanism with the same architecture as for the audio module, linearly projected to the same dimensionality as the audio (512) and L_2 normalized.

[3] Available from `https://dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_small.pt`.

[4] See `https://pytorch.org/vision/0.8/models.html#resnet-3d`.

4 Results

Performance metrics In the case of the narration data this scores is not confounded by speaker-based clues, which is an indication that the model possibly learned to detect some aspects of utterance meaning. We investigate this hypothesis further using multiple representational similarity analysis.

Targeted Triplets As a first baseline, we evaluate a model that has been pretrained but not fine-tuned on our dataset. The resulting performance is, as expected, close to chance level: 0.538. Additionally, we evaluate a model that is trained using static (image) data instead of video. The average accuracy is 0.705 . Finally, the best performing model according to the performance metrics (ID 68, audio and video pretraining) achieves an average targeted triplets accuracy of 0.745.

MN: add baselines: model that is completely untrained, and model where only the attention pooling layers are fine-tuned

Figure 4 and 5 show per-word accuracy for nouns and verbs, respectively. We perform boostrapping (n_resampling = 100) to estimate mean and standard deviation for each accuracy score.

We further compute correlations between the per-word accuracy and two possible predictors of age of acquisition: frequency and concreteness. The resulting correlations are presented in Appendix A.2. We do not find any significant correlation between the model's per-word accuracy and word concreteness or input frequency of a word in the training data.

5 Conclusion

References

Afra Alishahi, Marie Barking, and Grzegorz Chrupała. 2017. Encoding of phonology in a recurrent neural model of grounded speech. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 368–378, Vancouver, Canada. Association for Computational Linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

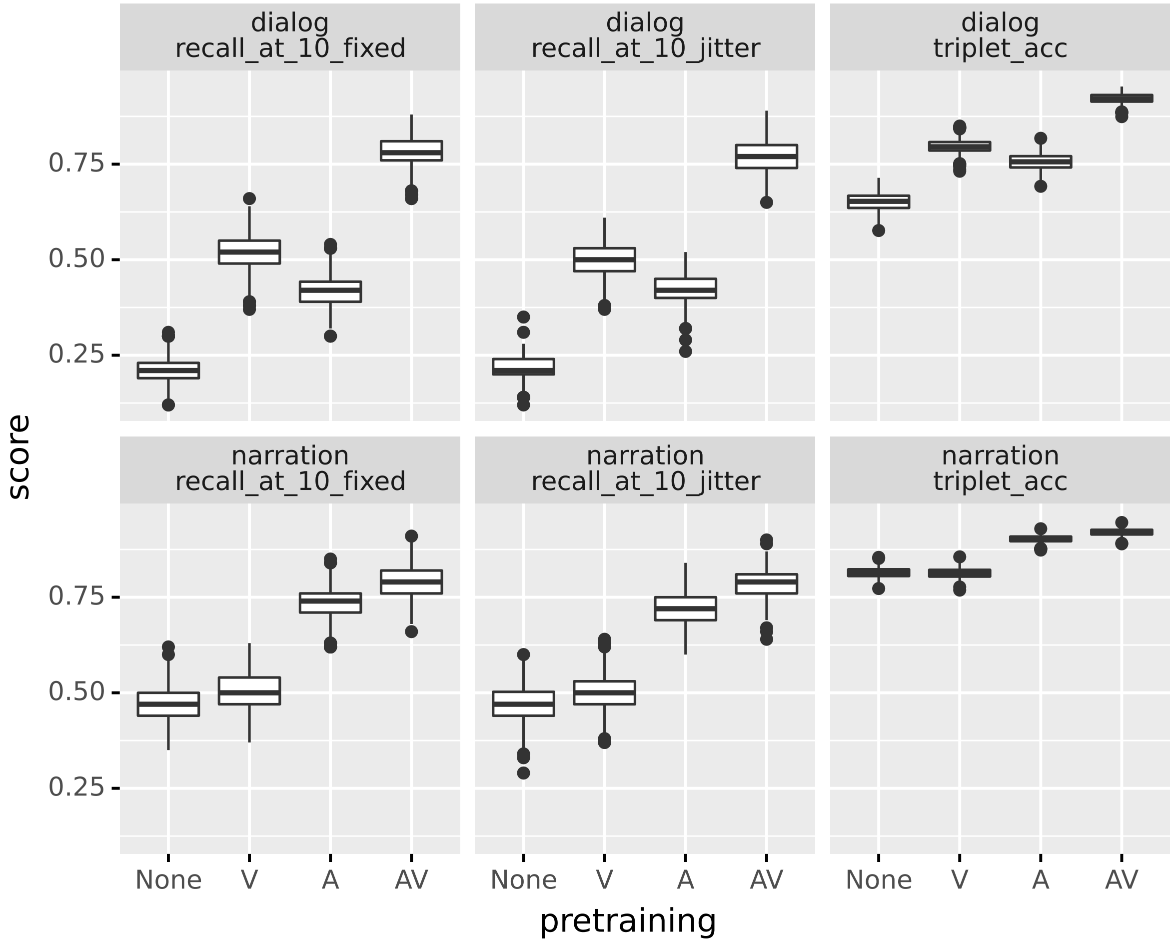

Figure 2: Effect of pre-training.

Elika Bergelson and Daniel Swingley. 2012. At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9):3253–3258.

Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. 2017. Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 613–622, Vancouver, Canada. Association for Computational Linguistics.

David Harwath and James Glass. 2015. Deep multimodal semantic embeddings for speech and images. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 237–244. IEEE.

David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. 2018. Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of the European conference on computer vision (ECCV)*, pages 649–665.

David F. Harwath, Antonio Torralba, and James R. Glass. 2016. Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1858–1866.

William N. Havard, Jean-Pierre Chevrot, and Laurent Besacier. 2019. Models of visually grounded speech signal pay attention to nouns: A bilingual experiment on english and japanese. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 8618–8622. IEEE.

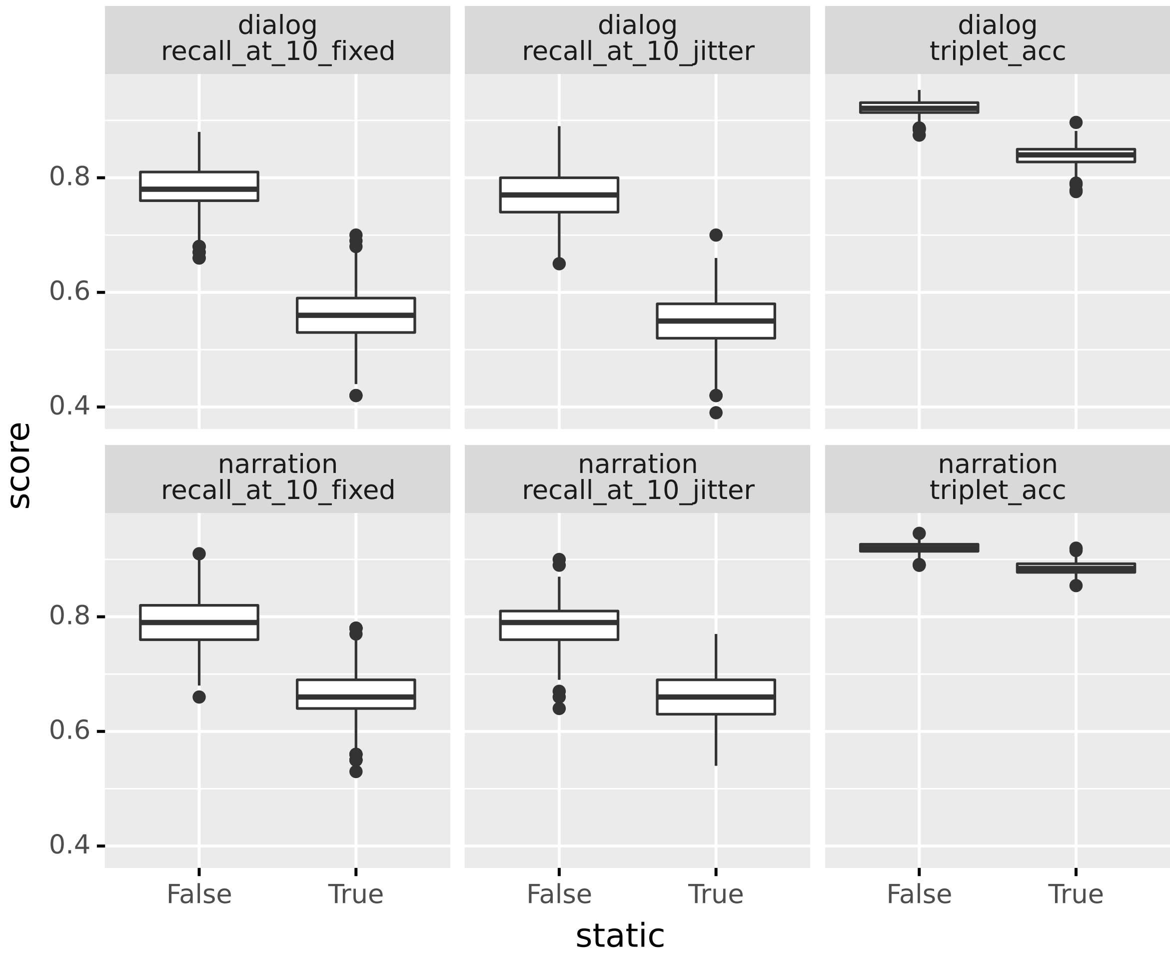

Figure 3: Effect of temporal information.

Kathy Hirsh-Pasek and Roberta Michnick Golinkoff. 1996. The intermodal preferential looking paradigm: A window onto emerging language comprehension.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. *Zenodo*.

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset. *CoRR*, abs/1705.06950.

Khazar Khorrami and Okko Räsänen. 2021. Can phones, syllables, and words emerge as side-products of cross-situational audiovisual learning? - a computational investigation. Preprint `psyarxiv.com/37zn`.

Danny Merkx, Stefan L. Frank, and Mirjam Ernestus. 2019. Language Learning Using Speech to Image Retrieval. In *Proc. Interspeech 2019*, pages 1841–1845.

Mitja Nikolaus and Abdellah Fourtassi. 2021. Evaluating the acquisition of semantic knowledge from cross-situational learning in artificial neural networks. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 200–210, Online. Association for Computational Linguistics.

Claire H Noble, Caroline F Rowland, and Julian M Pine. 2011. Comprehension of argument structure and semantic roles: Evidence from english-learning children and the forced-choice pointing paradigm. *Cognitive science*, 35(5):963–982.

Nikos Papasarantopoulos and Shay B. Cohen. 2021. Narration generation for cartoon

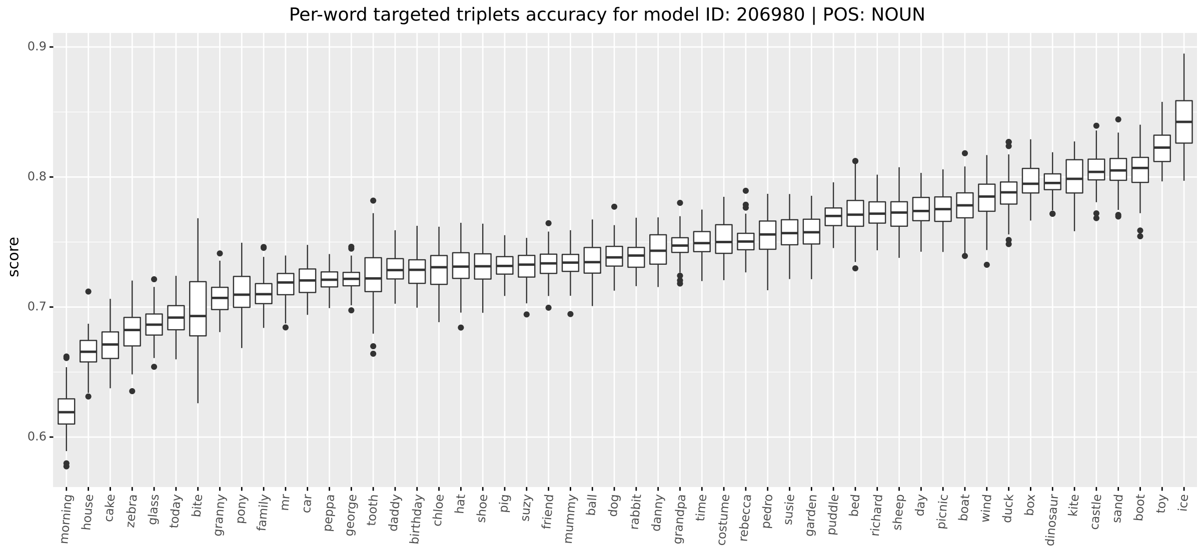

Figure 4: Per-word targeted triplets accuracy for nouns.

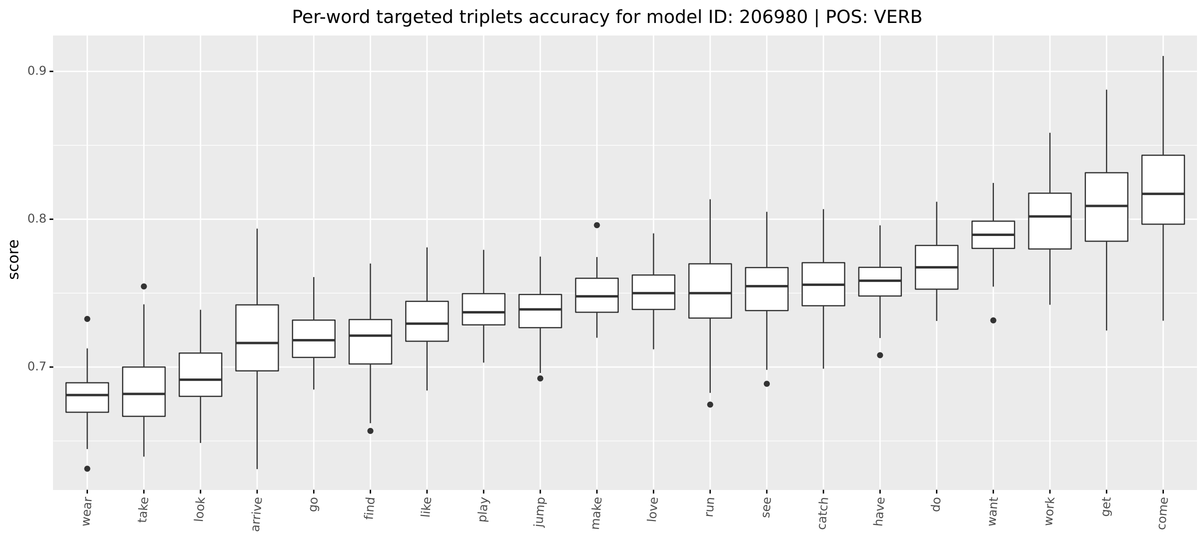

Figure 5: Per-word targeted triplets accuracy for verbs.

videos. Preprint: `https://arxiv.org/abs/2101.06803`.

Puyuan Peng and David Harwath. 2021. Fast-slow transformer for visually grounding speech.

Andrew Rouditchenko, Angie Boggust, David Harwath, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, et al. 2020. Avlnet: Learning audio-visual language representations from instructional videos.

Deb K Roy and Alex P Pentland. 2002. Learning words from sights and sounds: A computational model. *Cognitive science*, 26(1):113–146.

Thomas Schatz. 2016. *ABX-discriminability measures and applications*. Ph.D. thesis, Université Paris 6 (UPMC).

Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux. 2014. Learning words from images and speech. In *NIPS Workshop on Learning Semantics*.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459.

A Supplementary material

A.1 Retrieval and triplet accuracy

Table 2 and Table 3 show the performance of several model configurations on the retrieval and triplet tasks on the dialog and narration datasets respectively.

GC: Tables out of date

A.2 Targeted Triplets Evaluation Sets

To find commonly occurring nouns, adjectives, and verbs, we lemmatize and POS-tag all words in the transcripts of the validation dataset using spacy (Honnibal et al., 2020). Afterwards, we identify sets of all nouns $\{n_1, ..., n_n\}$, verbs $\{v_1, ..., v_o\}$ and adjectives $\{a_1, ..., a_p\}$ that occur at least 10 times in the validation data. Given these sets, we create sets of tuples $\{(n_1, n_2), (n_1, n_3), ..., (n_1, n_n), ..., (n_{n-1}, n_n)\}$ for all combinations of nouns and verbs, respectively. For each of these tuples, we search the validation data for pairs of phrases $(p_k = [w_1, ..., w_x], p_l = [w_1, ..., w_y])$ with same length $(x = y)$ and minimal difference regarding the tuple. That is, $n_1 \in p_1$, $n_2 \in p_2$, and if we replace n_1 with n_2 in p_1, it is equal to p_2.

For example, if n_1 = "peppa" and n_2 = "george", the phrases p_1 = ["peppa", "loves", "jumping"] and p_2 = ["george", "loves", "jumping"] are phrases with minimal differences. A phrase can also be a single word.

We set the minimum phrase duration to 0.3 seconds (for shorter sequences, we do not expect that the video data contains enough semantic information for a model to distinguish between target and distractor). For each phrase p_1 we look for the *longest* possible phrase p_2. Figure 6 shows the distribution of samples per duration.

Based on each minimal pair, we construct two counter-balanced test triplets as described in the main text.

Figures 7, and 8 show the number of samples for each noun and verb for which at least 100 sets of test triplets were available (for no adjective there were enough samples found).

A.3 Targeted Triplets Correlations

Figure 9 shows the correlation between per-word accuracy and frequency of this word in the training data. Figure 10 shows the correlation between per-word accuracy and concreteness scores.

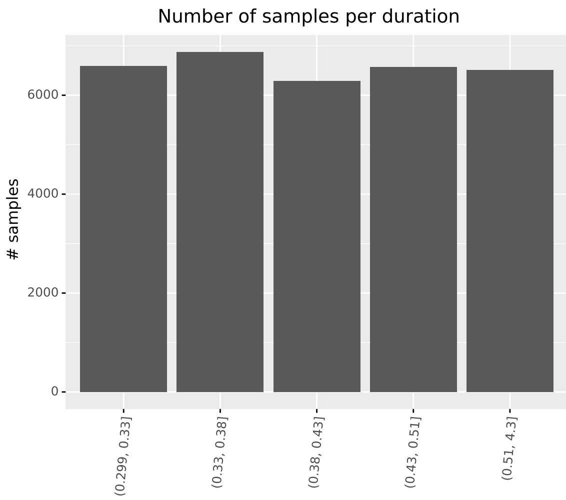

Figure 6: Number of samples per duration

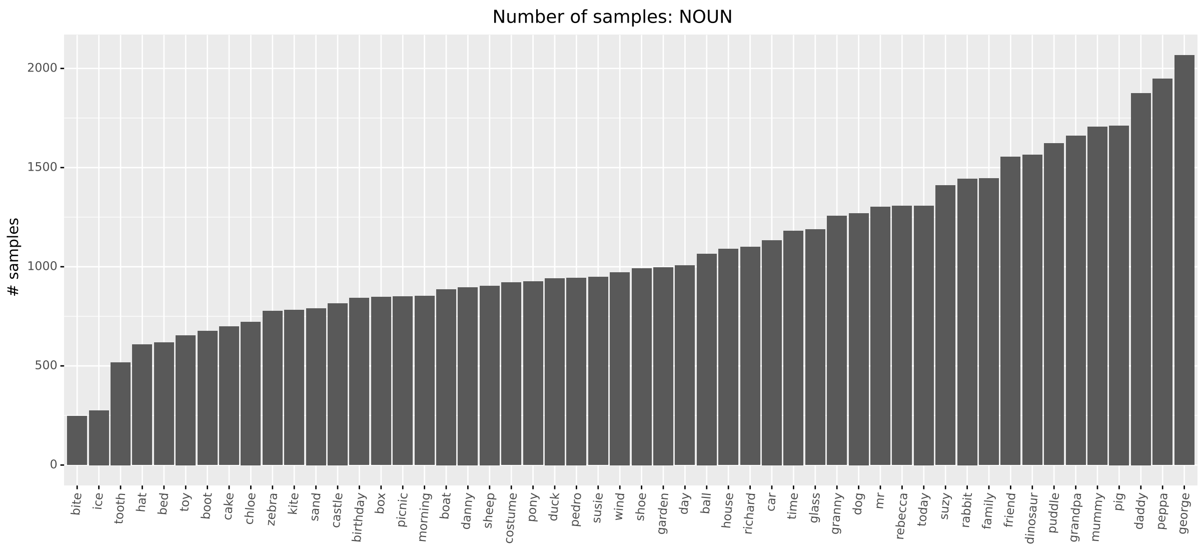

Figure 7: Number of samples: nouns

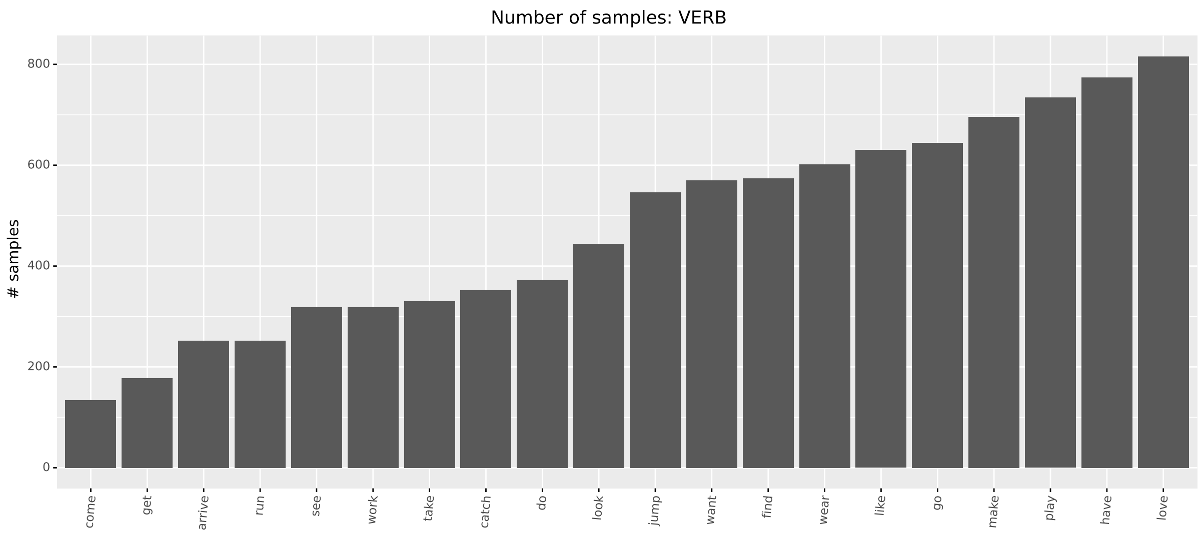

Figure 8: Number of samples: verbs

ID	Static	Jitter	Pretraining	Resolution	R@10 (fixed)	R@10 (jitter)	Triplet Acc
68			AV	180x100	0.561	0.533	0.834
206974		Yes	AV	180x100	0.557	0.539	0.830
206964		Yes	AV	360x200	0.594	0.589	0.852
206975		Yes	A	180x100	0.391	0.393	0.744
206976		Yes	V	180x100	0.220	0.205	0.630
206977		Yes	None	180x100	0.194	0.195	0.612
206978	Yes	Yes	AV	180x100	0.442	0.430	0.781

Table 2: Retrieval and triplet scores on dialog validation data.

ID	Static	Jitter	Pretraining	Resolution	R@10 (fixed)	R@10 (jitter)	Triplet Acc
68			AV	180x100	0.644	0.626	0.881
206974		Yes	AV	180x100	0.645	0.623	0.882
206964		Yes	AV	360x200	0.691	0.685	0.904
206975		Yes	A	180x100	0.585	0.574	0.869
206976		Yes	V	180x100	0.316	0.318	0.774
206977		Yes	None	180x100	0.362	0.346	0.770
206978	Yes	Yes	AV	180x100	0.556	0.555	0.850

Table 3: Retrieval and triplet scores on narration validation data.

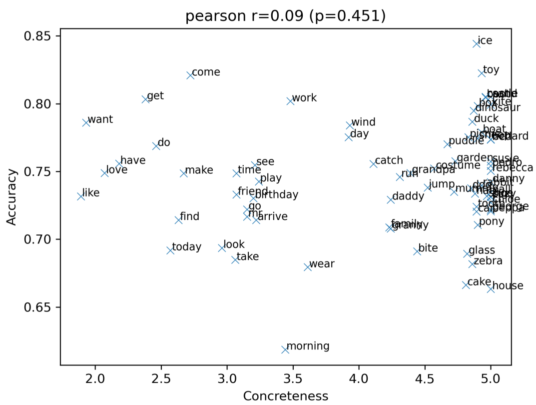

Figure 9: Correlation between accuracy and log frequency.

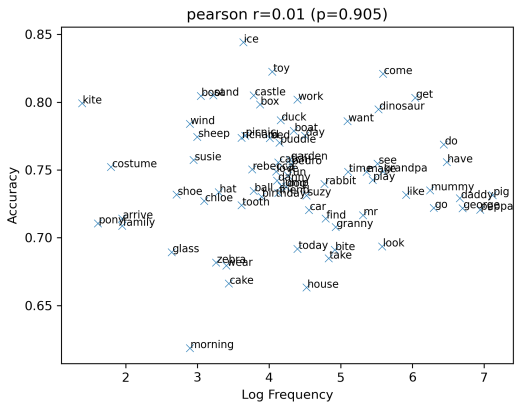

Figure 10: Correlation between accuracy and concreteness.