

Learning English with Peppa Pig

Abstract

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 2015. As noted by Chrupala (2021), most current approaches work well enough from an applied point of view but leave much to be desired as regards ecological validity. Most datasets consist of static images paired with their descriptions. Existing video datasets contain spoken descriptions of what happens in the video. 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 and environmental noise. These non-semantic correlations make it harder for the learner to zoom in on those aspects of the audio signal most relevant to learning meanings of linguistic units.

In the current study we make a first step towards simulating such a naturalistic grounding scenario: we use the well-know children’s cartoon *Peppa Pig* as a case study. Compared to commonly used video datasets, this data features 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. This choice of data thus allows us to directly address the ecological limitations of the current approaches.

We implement a model which learns to project visual scenes and spoken utterances into a join vector space, and train it on snippets of video containing dialog from the Peppa Pig cartoon, and carry out an in-depth evaluating of the nature of the learned representations using a variety of approaches.

Split	Type	Triplet	Size (h)	Items	Mean length (s)
Training	Dialog	No	9.83	11,058	3.2
Validation	Dialog	No	0.33	375	3.2
Validation	Narration	No	0.80	897	3.2
Validation	Dialog	Yes	0.16	202	2.8
Validation	Narration	Yes	0.45	726	2.2

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.

2 Related work

3 Method

3.1 Dataset

The dataset consists of the complete 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 of the audio as well as segmentation into *dialog* and *narration*. 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 actor. We use the dialogs for training the model, and set aside the narrations for evaluation purposes only.

Specifically, we use dialog from episodes 1–196 for training, 197–202 for validation and 203–209 for testing. We set aside narrations from episodes 1–104 for validation and 105–209 for testing.

3.2 Preprocessing

For training, we do not use word or sentence level segmentation in order to make the setting more naturalistic. Instead we split the dialog sections into 3.2 second non-overlapping fragments. The video is subsampled to 10 frames per second, and to 180x100 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.

Add the
test split.

3.3 Evaluation

4 Results

5 Conclusion

References

- Chrupała, G. (2021). Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques. Preprint: <https://arxiv.org/abs/2104.13225>.
- Papasarantopoulos, N. and Cohen, S. B. (2021). Narration generation for cartoon videos. Preprint: <https://arxiv.org/abs/2101.06803>.
- Roy, D. K. and Pentland, A. P. (2002). Learning words from sights and sounds: A computational model. *Cognitive science*, 26(1):113–146.