Dynamic visual grounding via video

LOT School 2024

Grzegorz Chrupała

Most work on visual grounding of spoken language has used static images, due both to the wider availability of curated captioned image datasets, as well as the inherent complexity of modeling two weakly synchronized modalities in the temporal domain.

However, for humans visual perception is inherently extended in time, and visual grounding of language based on still images is arguably suboptimal also from an engineering point ofview. Many aspects of language are related to processes or actions evolving over time rather than static states and may be more naturally grounded in video.

Instructional cooking videos collected from YouTube have been used in several papers.

A typical video captures the person carrying out the various actions involved in cooking, while they are talking explaining what is being done.

Clip from http://youcook2.eecs.umich.edu/

Another dataset, HowTo100M (https://www.di.ens.fr/willow/research/howto100m/), consists of other types of instructional videos featuring narrative descriptions.

Another dataset, Spoken Moments in Time (http://moments.csail.mit.edu/spoken.html), feature short video clips paired with spoken captions describing what's happening.

This dataset is very close in nature to the Spoken COCO dataset, but instead of static images it contains clips.

Both the cooking videos and the captioned clips are somewhat special as they contain someone describing an event or series of event in some detail.

AVLnet: Learning Audio-Visual Language Representations from Instructional Videos

Andrew Rouditchenko[1*] Angie Boggust[1*] David Harwath[2] Brian Chen[3]
Dhiraj Joshi[4] Samuel Thomas[4] Kartik Audhkhasi[5] Hilde Kuehne[4] Rameswar Panda[4]
Rogerio Feris[4] Brian Kingsbury[4] Michael Picheny[6] Antonio Torralba[1] James Glass[1]
[1]MIT CSAIL, [2]UT Austin, [3]Columbia University, [4]IBM Research AI, [5]Google, [6]NYU
roudi@mit.edu

2021

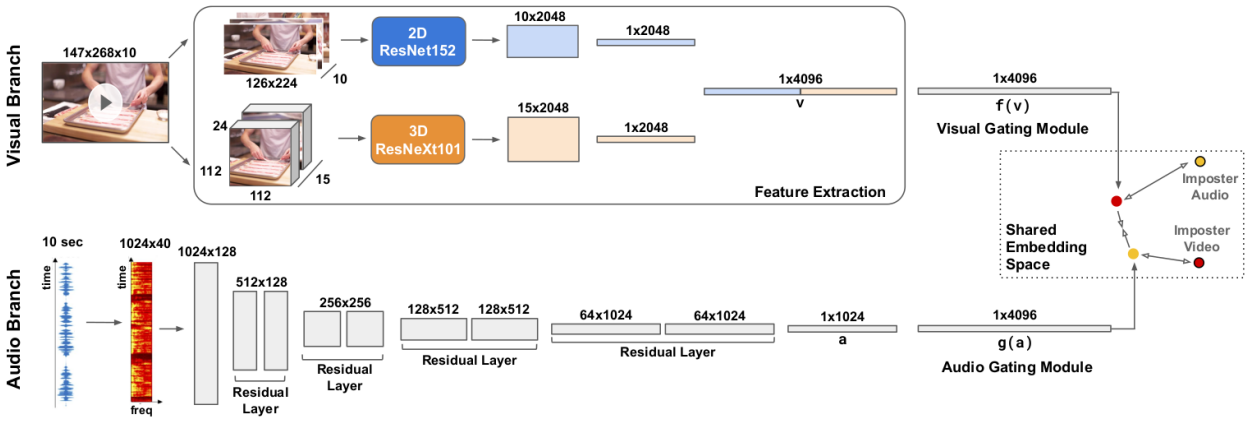


Figure from https://doi.org/10.48550/arXiv.2006.09199

AVLNet the network consists of an audio encoder ResNet-based, a video encoder which combines 3D and 2D modeling (also ResNet-based). ResNet is a type of convolutional architecture.

2D convolutions apply to spatial dimensions. 3D Resnet also convolves over the time dimensions. The video encoder then pools each video segment (10s long) into a single vector.

The audio is pooled and projected to the shared embedding space, and the loss function is similar to what we saw with models grounded via stil images.

The video encoder is pretrained on video classification, while the audio encoder is trained from scratch.

Method	Speech-241		Sounds-241	
	A→V	V→A	A→V	V→A
AVLnet zero-shot	88.0	88.0	32.4	33.6
AVLnet fine-tuned	**92.5**	**91.7**	44.0	46.8

“Figure 3: Video (top) and audio retrieval (bottom) results from AVLnet fine-tuned on YouCook2. Video clips are represented as their center frame, and audio clips are represented as their waveform and ASR transcript. The correct match is highlighted.”

The table (bottom) compares retrieval performance on clips which contain speech (Speech-241) with those which only contain non-speech sounds (Sounds-241).

A: They are fighting over bread.
C: What?
A: They are fighting over bread.
C: Yeah.

An brief real-world verbal interaction between a child and an adult in a city park.

The dialog is in Persian, and has been rendered in English above.

How does a video clip of natural interaction compare to a typical instructional video, or a captioned clip?

What the implications for ecological validity of instructional videos, or captions?

Image/video descriptions

Strong correlations between visual scene and the meaning of the utterances.

Few/weak correlations between scene and non-semantic features of the audio.

Real world interactions

Weaker correlations between visual scene and meaning of the utterances.

Correlations between scene and non-semantic features of the audio are common.

Learning English with Peppa Pig

Mitja Nikolaus
Aix-Marseille University
mitja.nikolaus@univ-amu.fr

Afra Alishahi
Tilburg University
a.alishahi@uvt.nl

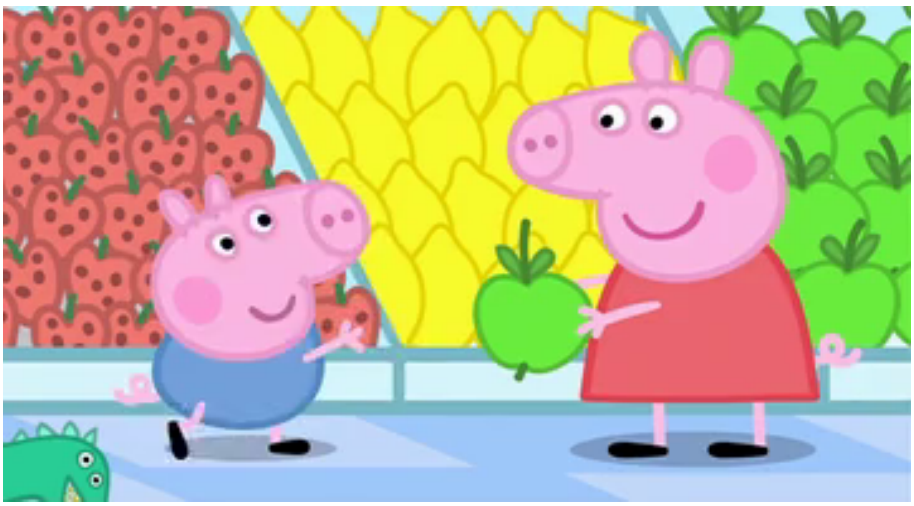
Grzegorz Chrupała
Tilburg University
grzegorz@chrupala.me

Motivated by considerations of ecological validity we used a children's animated series, Peppa Pig, as a dataset to model language acquisition via visual grounding.

These cartoons are aimed at young children, feature simple graphics and simple but realistic spoken language.

Correlations between speech and visuals tend to be present but are less consistent than in image captions or video descriptions.

Thre are also counfounding correlations between speakers and content of speech, as well as non-speech sounds which can distract the learner from the semantic content of speech.



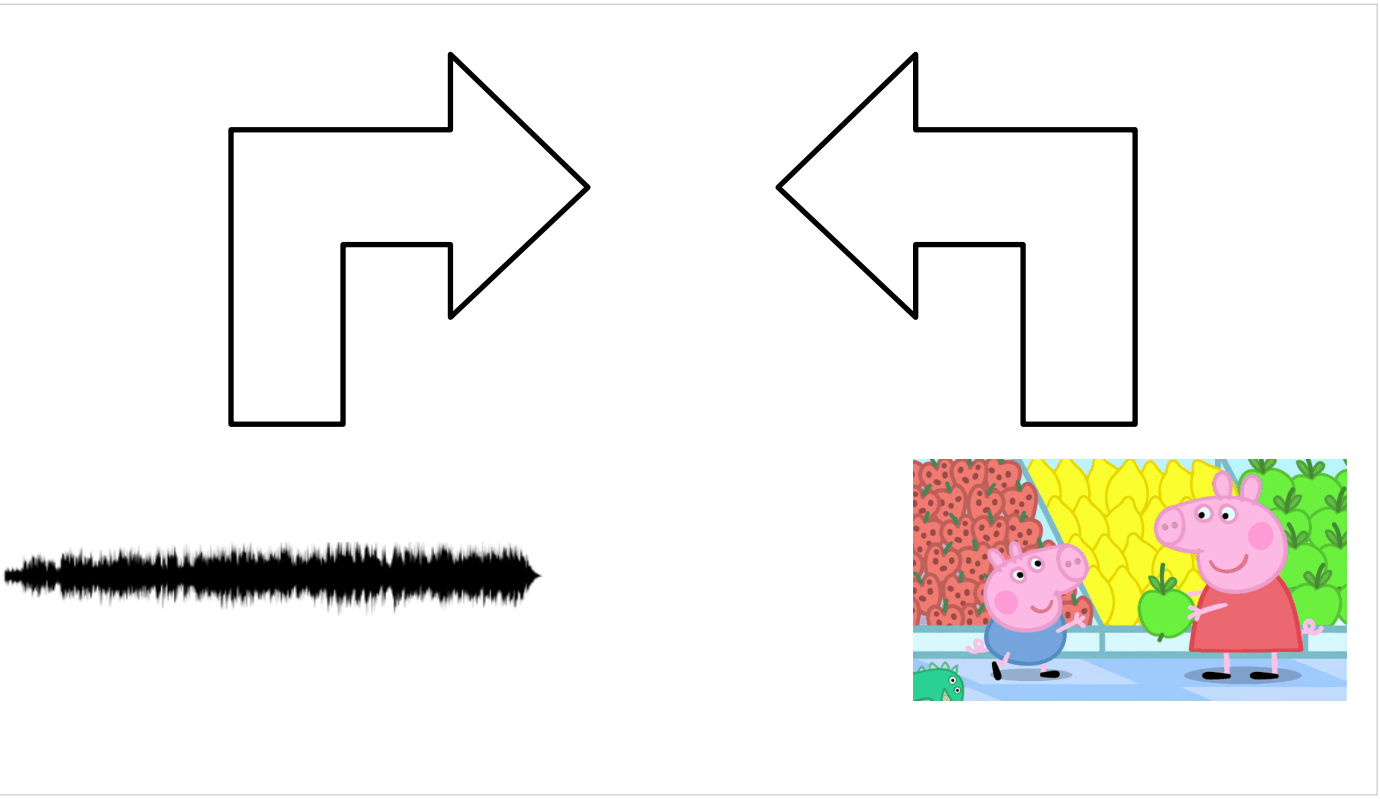
Can you think of a way to use the two different types of speech (dialogs and narrations) in the cartoons for different purposes?

We use dialogs between characters to train the model. The narrations are excluded from the training data and held out to be used only for evaluation.

The narrations are descriptive and caption-like in nature. They narrator voice is also not present in the training data.

Thus, if the model trained on dialogs generalizes to narrations, this is evidence that it's using semantic features of speech, and doesn't rely on confounds.

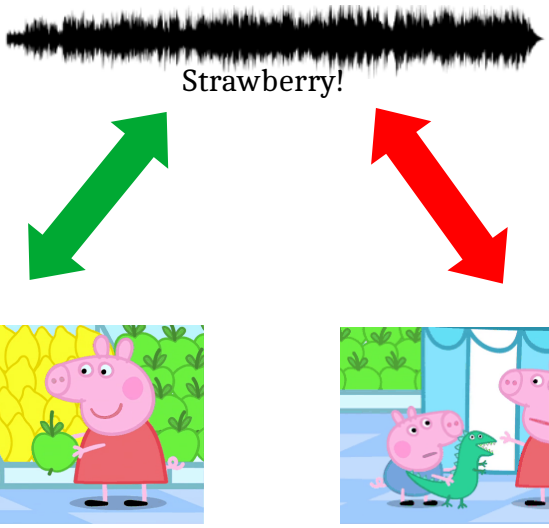
Split	Type	Size (h)	# Clips
train	dialog	10.01	15666
val	dialog	0.66	1026
val	narration	0.94	1467
test	narration	0.64	1006



We use a simple architecture with a wav2vec2-based audio encoder, and a ResNet(2+3)D video encoder.

We evaluate both pretrained encoders as well as training from scratch.

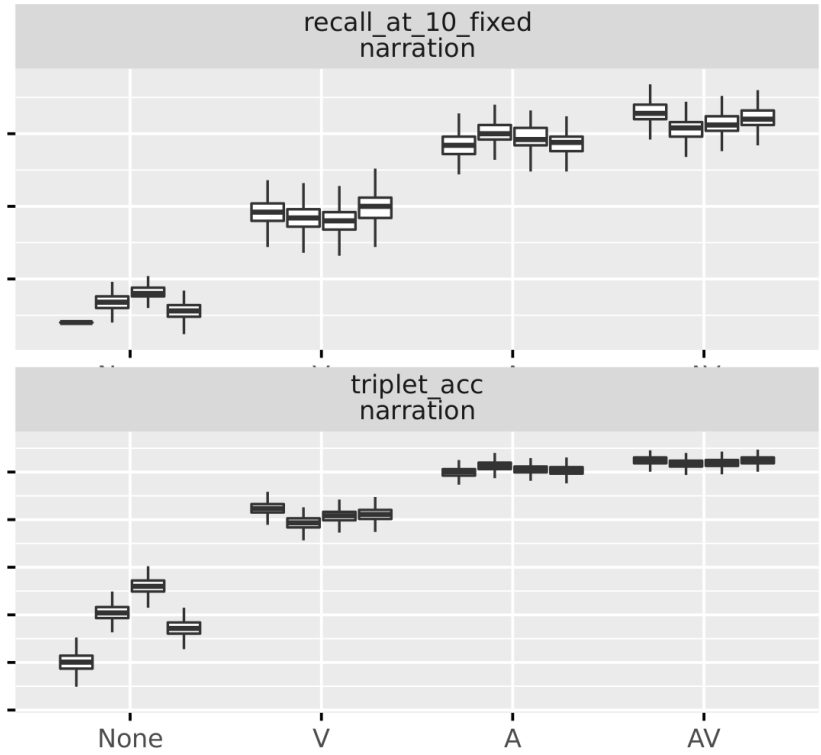
Our objective is a contrastive triplet-like loss.



We evaluate using the usual video clip retrieval setting.

Additionally we also use an alternative metric with simpler stimuli:

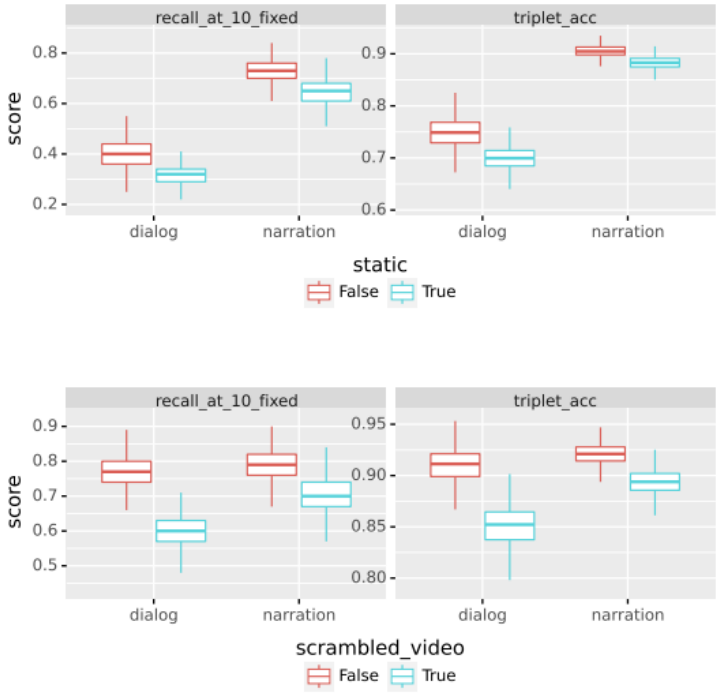
- Extract video/audio clips aligned with subtitle lines
- Repeat 500 times
  - Sample triplets of same length
  - Compute success rate: anchor close to positive than negative



We do a number of ablations.

Regarding the effect of pretraining, we see that pretraining always helps, and is crucial for the video encoder, and less so for the audio encoder.

How would you check to what extent the temporal dimension of the video clips was important for the model?

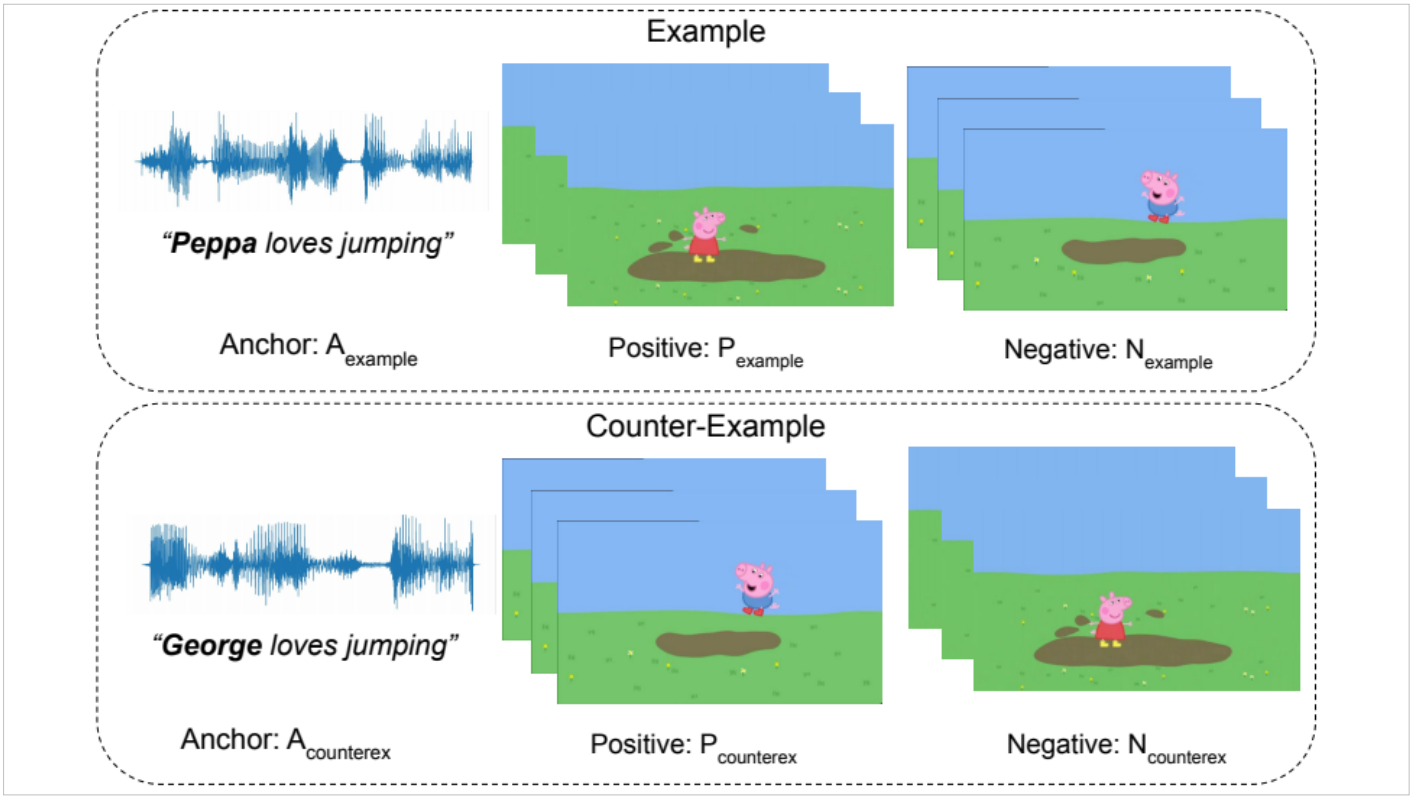


We check the effect of ablating the temporal dimension in two different ways:

1) Video encoder based on a 2D ResNet which encodes each video frame separately, and then pools the encodes frames.

2) Scrambling the video frames before feeding them to the video encoder.

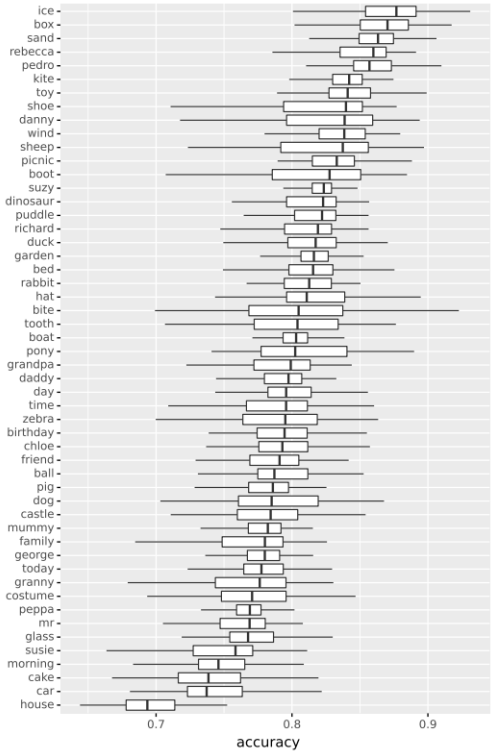
In both cases we see that removing the temporal signal substantially lowers the evaluation scores of the model.



A targeted, controlled triplet evaluation:

Triplets are matched to differ by a single word (via aligned transcriptions).

This setting allows us to partition results by word, grammatical category.



What we saw were mostly patterns specific to Peppa cartoons.

Character names (Rebecca, Pedro) were easy. Some surprises included low scores for *house* or *car:* house is used in varying visual contexts (house entrance, whole house, inside the house) and in displaced speech (going to somebody's house).

The pattern of results on verbs vs nouns was not illuminating.

These results highlight the limitations of the paper: while ecological validity was addressed to some extent, the data is too small scale for the findings generalize beyond the Peppa cartoon setting.

Learning language in a realistic grounded setting is still largely an open problem.