Speech2Action: Cross-modal Supervision for Action Recognition

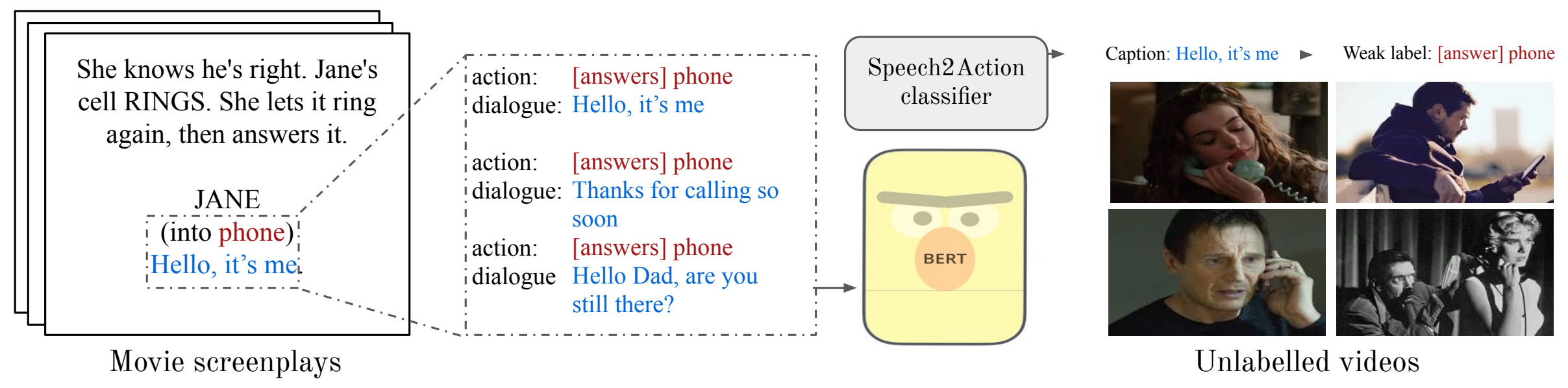
Arsha Nagrani[1,2] Chen Sun[2] David Ross[2] Rahul Sukthankar[2] Cordelia Schmid[2] Andrew Zisserman[1,3]

[1]VGG, Oxford [2]Google Research [3]DeepMind

CVPR SEATTLE
JUNE 14-19 2020 WASHINGTON

Problem Definition and Contribution

Goal: Action Recognition in Movies and TV shows using only the speech as supervision



She knows he's right. Jane's cell RINGS. She lets it ring again, then answers it.

JANE
(into phone)
Hello, it's me.

action: [answers] phone
dialogue: Hello, it's me

action: [answers] phone
dialogue: Thanks for calling so soon

action: [answers] phone
dialogue Hello Dad, are you still there?

Caption: Hello, it's me ► Weak label: [answer] phone

Movie screenplays

Unlabelled videos

Motivation:

- Manual annotation of human actions is expensive and not scalable
- The audiotrack is usually freely available for large video corpuses.

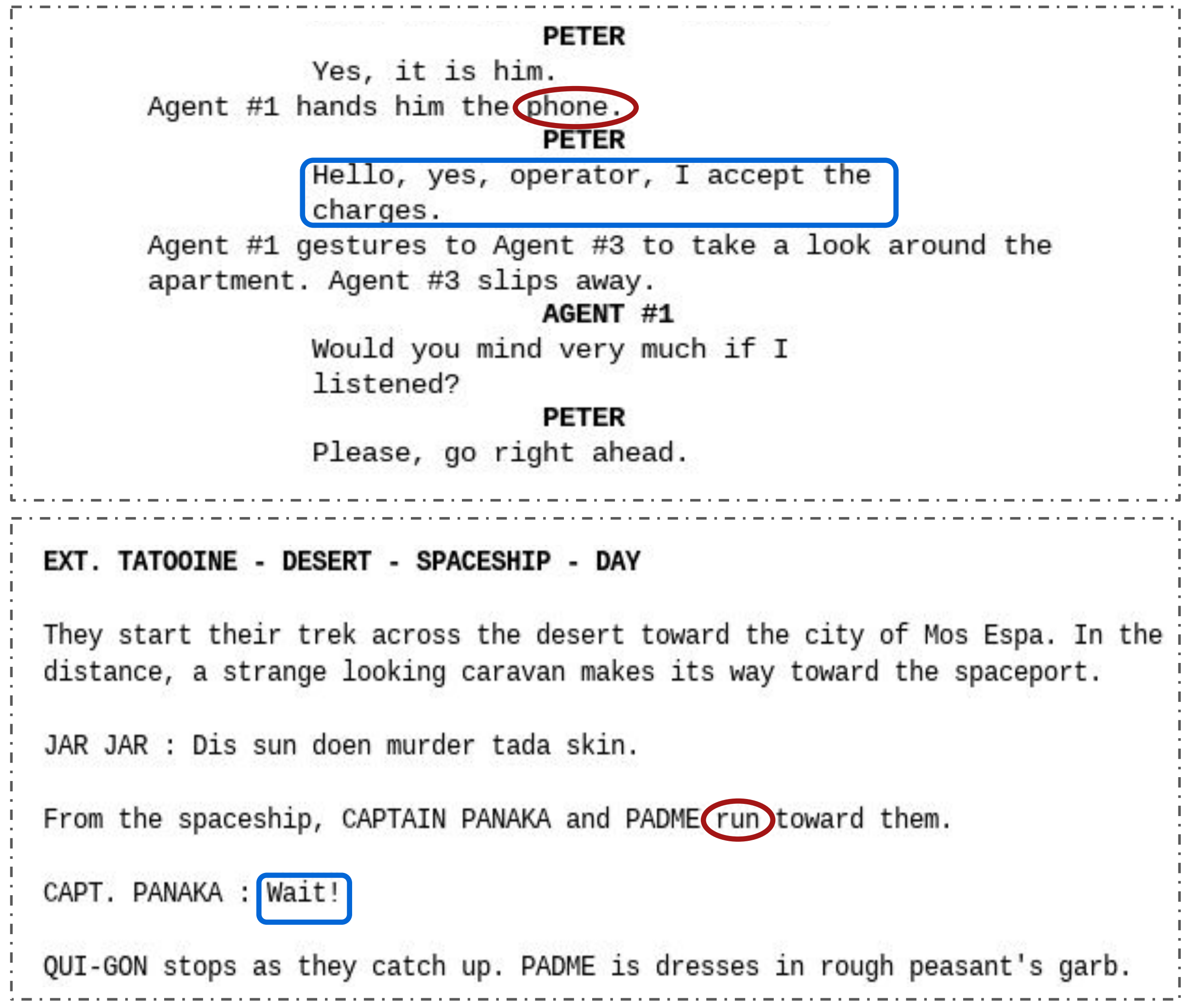
Key Contributions:

- A `Speech2Action` model trained from literary screenplays that predicts actions from transcribed speech *alone*
- By applying this `Speech2Action` model to a large unlabelled corpus of videos, we obtain obtain weak action labels for over 800K video clips
- An action classifier trained on these clips with *no* finetuning beats fully supervised performance on the AVA dataset.
- With finetuning, the classifier achieves state of the art results on HMDB51

IMSDb Dataset

- Download 1,080 movie scripts from `www.IMSDb.com` spanning 22 genres
- Separate out scene descriptions (which contain mention of actions) and speech segments.
- Create a text dataset of speech paired with action labels, using proximity in the movie scripts as a cue.

Examples of Movie Scripts



```
                              PETER
                    Yes, it is him.
       Agent #1 hands him the phone.
                              PETER
                    Hello, yes, operator, I accept the
                    charges.
       Agent #1 gestures to Agent #3 to take a look around the
       apartment. Agent #3 slips away.
                              AGENT #1
                    Would you mind very much if I
                    listened?
                              PETER
                    Please, go right ahead.
```

```
EXT. TATOOINE - DESERT - SPACESHIP - DAY

They start their trek across the desert toward the city of Mos Espa. In the
distance, a strange looking caravan makes its way toward the spaceport.

JAR JAR : Dis sun doen murder tada skin.

From the spaceship, CAPTAIN PANAKA and PADME run toward them.

CAPT. PANAKA : Wait!

QUI-GON stops as they catch up. PADME is dresses in rough peasant's garb.
```

Mining with the Speech2Action Model

Main idea: We train a text-based model to predict actions from transcribed speech alone. This model is trained on movie scripts downloaded from IMSDb. This can be applied to the transcribed speech from unlabelled videos to automatically get labels for video clips.

Speech2Action Model

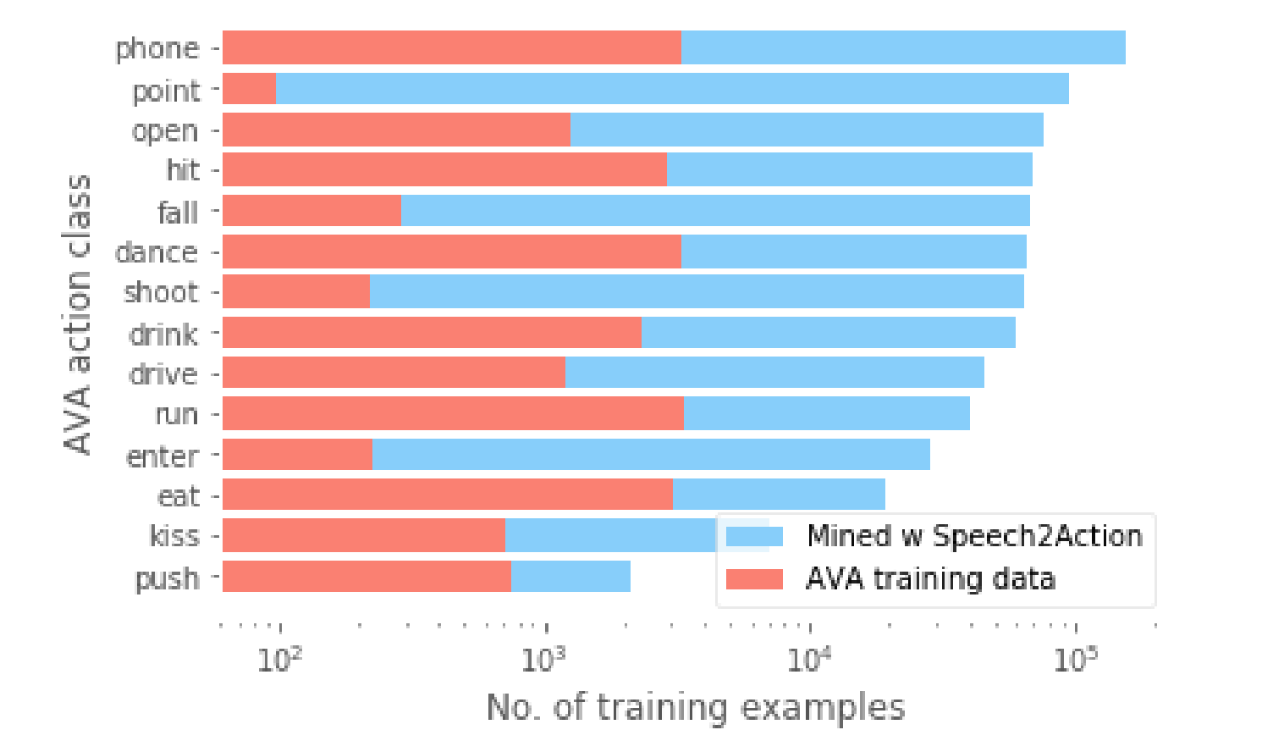
- We obtain speech-action paired data for 18 action classes from the IMSDb data
- We finetune a BERT model pretrained on English Wikipedia and the BooksCorpus

PHONE	KISS
Hello, it's me. May I have the number for Dr George Honey I asked you not to call unless hey, it's me Hello, it's me. Hello?	One more kiss Give me a kiss Good night my darling I love you my darling Noone had ever kissed me there before Goodnight angel my sweet boy

Mining Clips Automatically:

- We apply the Speech2Action model to the subtitles of unlabelled movies and TV shows.
- We then assign the label for highly confident predictions of the model to the accompanying video clip.
- In this manner we mined over 800K video clips and assign them with action labels based on the speech alone.

We mine two orders of magnitude more data than AVA automatically



More details at: `https://www.robots.ox.ac.uk/~vgg/research/speech2action/`
Contact: `arsha@robots.ox.ac.uk`

Results on Visual Action Recognition

Examples of clips mined using Speech2Action:

EAT

this chicken is very tasty

have you ever had szechwan cuisine before?

this food is so good.

are ronnie and nancy on the cover your menu?

DRINK

these drinks are strong.

ah, i am the one sipping the champagne now.

that's why i'm sitting here day drinking in the

after two belvedere martinis straight up with twists.

RUN

go, go, go!

run faster, baby!

don't move hey!

chase, chase!

POINT

see that, up there?

look at that right there.

and that one there.

is that it over there?

PHONE

please leave a message after the tone.

yes, i need jeff on his secure line.

i am trying brother from other phone.

pick up, oleg.

DANCE

why are you green and dancing?

and nandita, let's see your sita dance.

how can you not dance?

you dance, yank?

Examples of abstract actions mined using Speech2Action:

COUNT

two quarters, three dimes, one nickel, two pennies.

twenty four thousand four hundred.

20 dollar, four cents

thirty six thousand four hundred, five hundred,

FOLLOW

after you

stay close behind me now

just follow my lead

follow me quick!

Results on 14 AVA mid and tail classes

Data	Per-Class AP													
	drive	phone	kiss	dance	eat	drink	run	point	open	hit	shoot	push	hug	enter
AVA (fully supervised)	0.63	0.54	0.22	0.46	0.67	0.27	0.66	0.02	0.49	0.62	0.08	0.09	0.29	0.14
`S2A-mined` (zero-shot)	0.83	0.79	0.13	0.55	0.68	0.30	0.63	0.04	0.52	0.54	0.18	0.04	0.07	0.04
`S2A-mined` + AVA	**0.86**	**0.89**	**0.34**	**0.58**	**0.78**	**0.42**	**0.75**	0.03	**0.65**	**0.72**	**0.26**	**0.13**	**0.36**	**0.16**

Action Recognition Model

- We train an S3D-G model for 18-way classification on video clips labelled with the Speech2Action model
- We evaluate on AVA with NO finetuning, on mid and tail classes. These actions occur *rarely*, and are hard to get manual supervision for. For 8 classes, we exceed fully supervised performance without a single manually labelled training example.
- On HMDB51, we obtain a 17% improvement over training from scratch and also outperform previous self-supervised and weakly supervised works.

Results on HMDB51

Method	Architecture	Pre-training	Acc.
Shuffle&Learn	S3D-G (RGB)	UCF101	35.8
OPN	VGG-M-2048	UCF101	23.8
ClipOrder	R(2+1)D	UCF101	30.9
3DRotNet	S3D-G (RGB)	Kinetics	40.0
DPC	3DResNet18	Kinetics	35.7
CBT	S3D-G (RGB)	Kinetics	44.6
DisInit (RGB) 2019	R(2+1)D-18	Kinetics∗∗	54.8
Korbar et al. 2018	I3D (RGB)	Kinetics	53.0
-	S3D-G (RGB)	Scratch	41.2
Ours	S3D-G (RGB)	`S2A-mined`	**58.1**

Acknowledgments

Arsha is supported by a Google PhD Fellowship. We are grateful to Carl Vondrick for early discussions.