Speech2Action: Cross-modal Supervision for Action Recognition

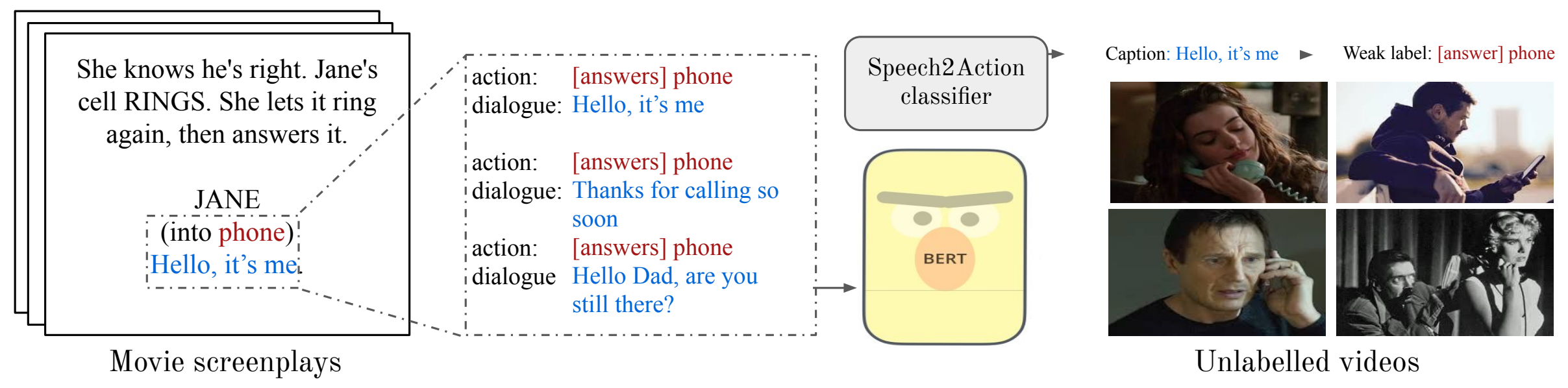
Arsha Nagrani[1,2] Chen Sun[2] David Ross[2] Rahul Sukthankar[2] Cordelia Schmid[2] Andrew Zisserman[1,3]

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

CVPR SEATTLE JUNE 14-19 2020 WASHINGTON

Problem Definition and Contribution

Goal: Action recognition in video using only transcribed speech as supervision



Motivation:

- Manual annotation of actions is expensive, not scalable
- Audiotrack is often freely available!

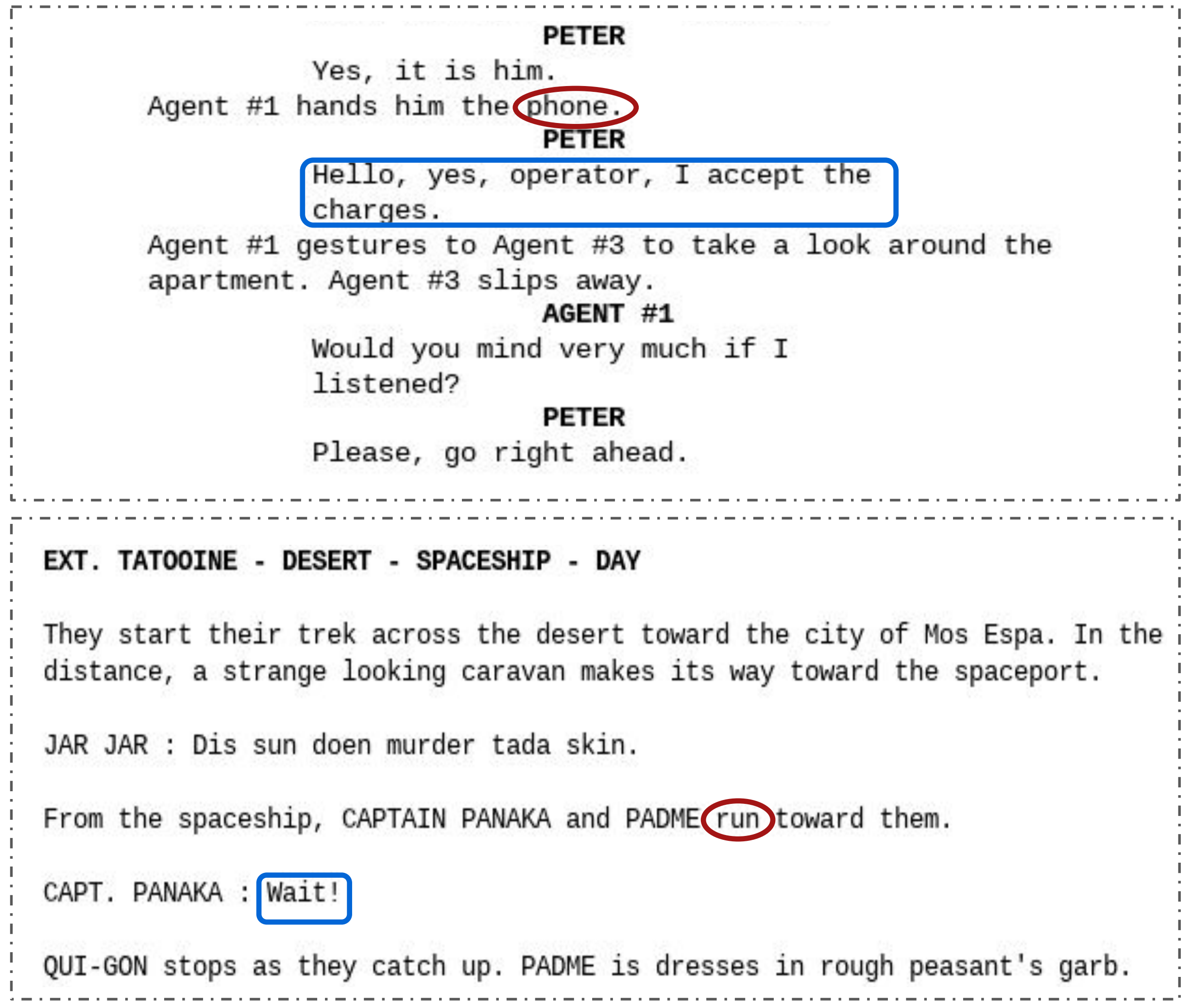
Key Contributions:

- A `Speech2Action` model trained from screenplays that predicts actions from transcribed speech *alone*
- Applying `Speech2Action` to a unlabelled video, we obtain weak action labels for > 800K video clips
- An visual action classifier trained on these clips with *no* fine-tuning gets SOTA performance

IMSDb Dataset

- 1,080 movie scripts from `www.IMSDb.com`, 22 genres
- Create a text dataset of **speech** paired with **action** labels from the scene directions, based on proximity in the movie script

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.

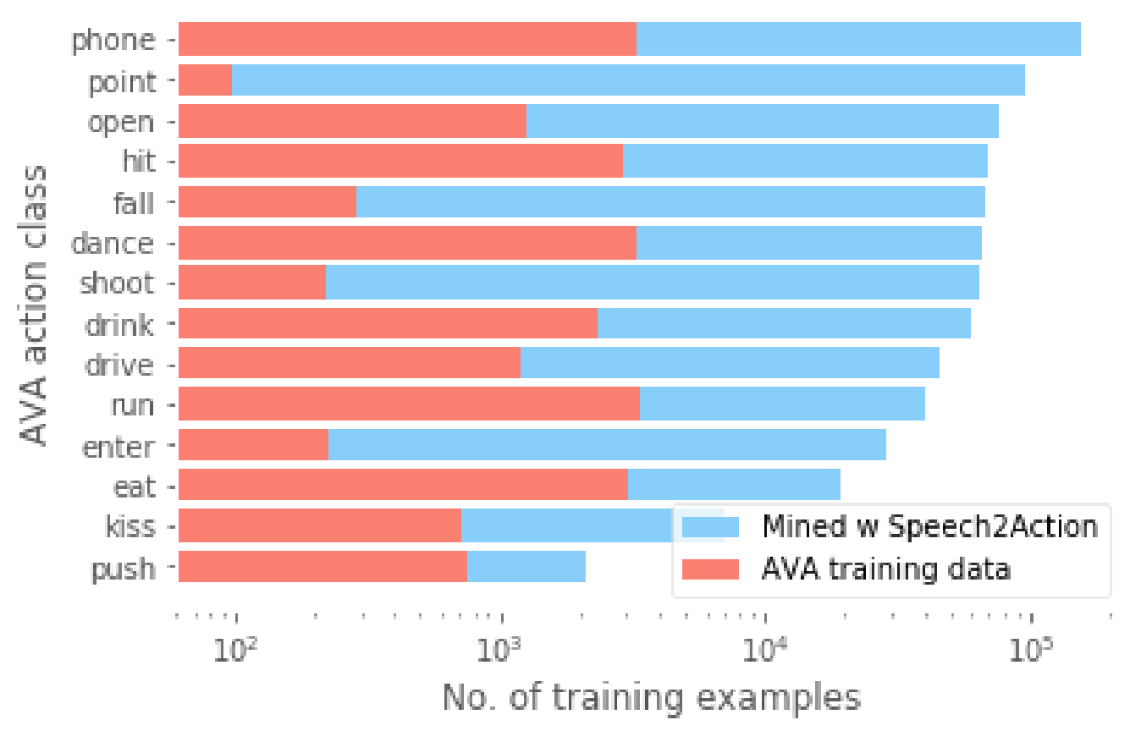
Mining with the Speech2Action Model

Main idea: Train a text-based model on movie scripts to predict actions from transcribed speech alone. Apply this to the transcribed speech from unlabelled videos to automatically get weak labels for video.

Speech2Action Model

- Finetune a pretrained BERT model on speech-action pairs in moviescripts

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



Mining Clips Automatically:

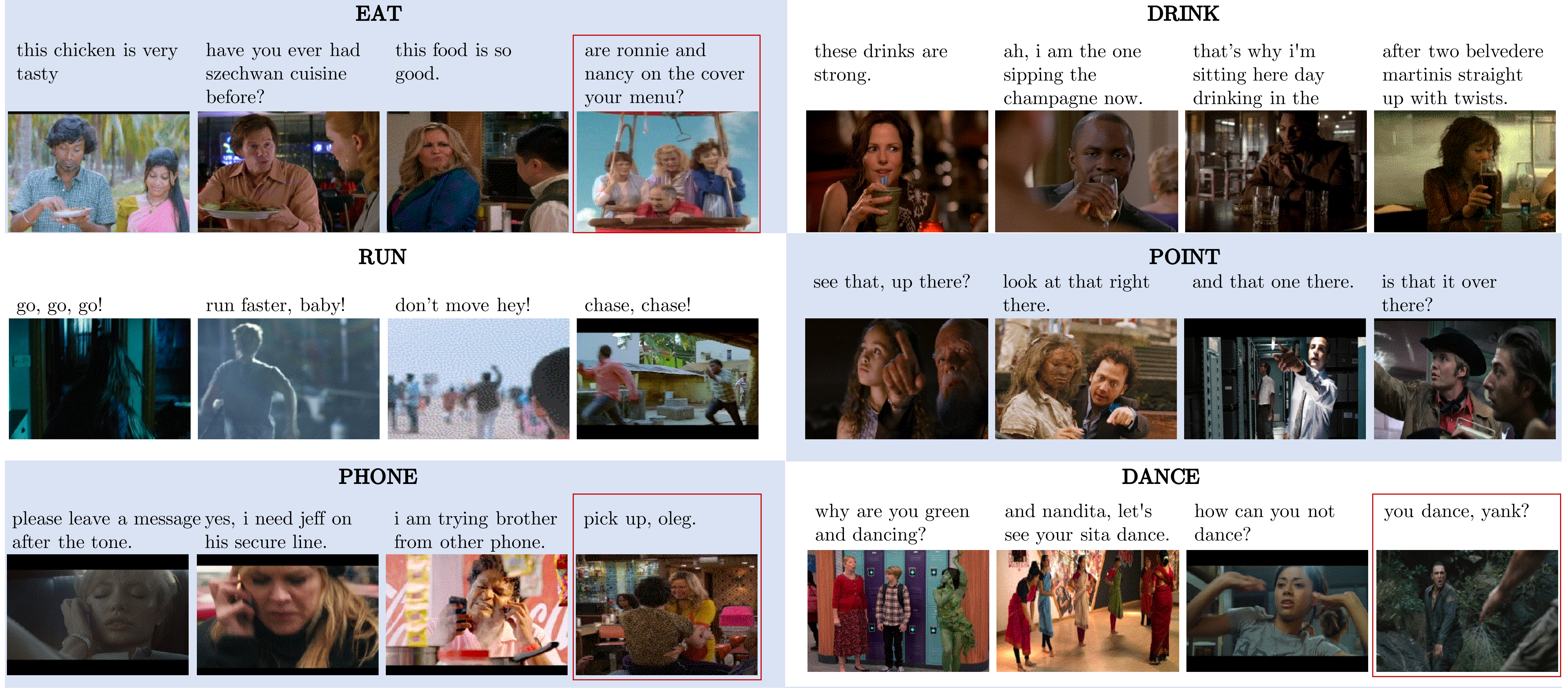
- Apply Speech2Action to subtitles of unlabelled movies.
- Assign label for highly confident preds to accompanying video.
- Mine over 800K video clips and assign them with action labels based on the speech alone.

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**

Visual Action Recognition

- Train S3D-G with an 18-way softmax
- Evaluate on AVA with NO finetuning, on mid and tail classes. These actions occur *rarely*. For 8 classes, exceed fully supervised performance without any supervision.
- On HMDB51, obtain 17% improvement over training from scratch and outperform SOTA.
- Even able to label abstract actions like count, follow etc!

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