Improving Visual and Speech Recognition on Out-Domain Data

Liangliang Cao

Google Inc. and Umass

http://llcao.net

Acknowledgement

UMass

- Aruni RoyChowdhury
- Prithvijit Chakrabarty
- Ashish Singh
- SouYoung Jin
- Huizhu Jiang
- Eric Learned-Miller

All of the pictures on visual recognition are from Aruni's CVPR19 with author's consent.

Google

- Chung-Cheng Chiu
- Arun Narayanan
- Wei Han
- Rohit Prabhavalkar
- Yu Zhang
- Navdeep Jaitly
- Ruoming Pang
- Tara N. Sainath
- Patrick Nguyen
- Yonghui Wu

Problem

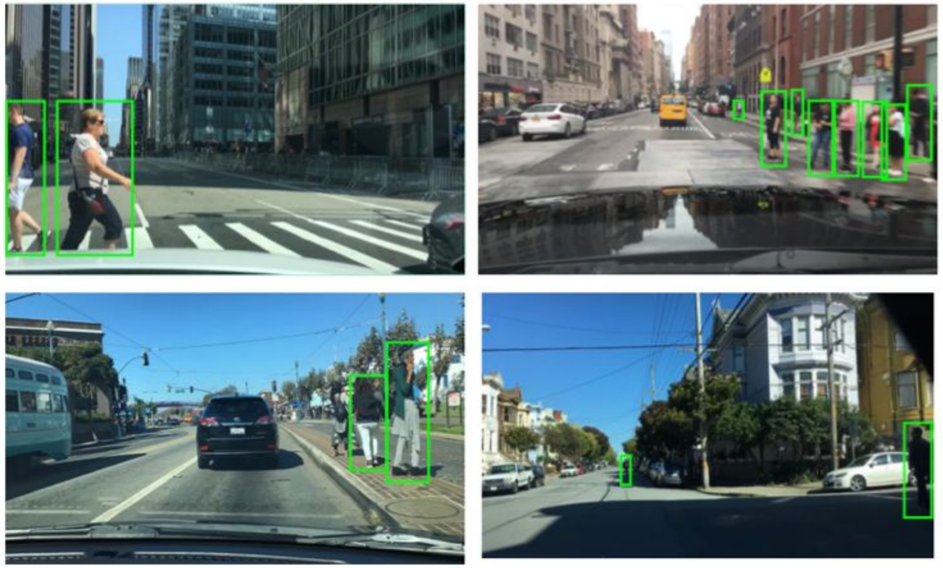
The popular machine learning paradigm:

1. collect a dataset
2. split it into training and test sets (of the same distribution)
3. train a model and report the performance on testing set

Problem: Few discuss the performance on out of domain data.

The accuracy may degenerate significantly!

Example of out-domain recognition



Daytime, sunny

How about fog?

Two applications on out-domain recognition

Challenges from out-domain Data

- Not label or limited labels for out-domain
- We prefer ONE model instead of MANY models

In this talk, we’d go through two studies under this topic:

- visual detection
- speech recognition

Study I

Improving visual object detection with unlabeled videos

Automatic adaptation of object detectors to new domains using self-training, CVPR 2019 (with Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huizhu Jiang, Eric Learned-Miller)

[Nam and Han 2016. MDNet tracker]

***Hard* positive:**
Missed by detector,
picked up by tracker

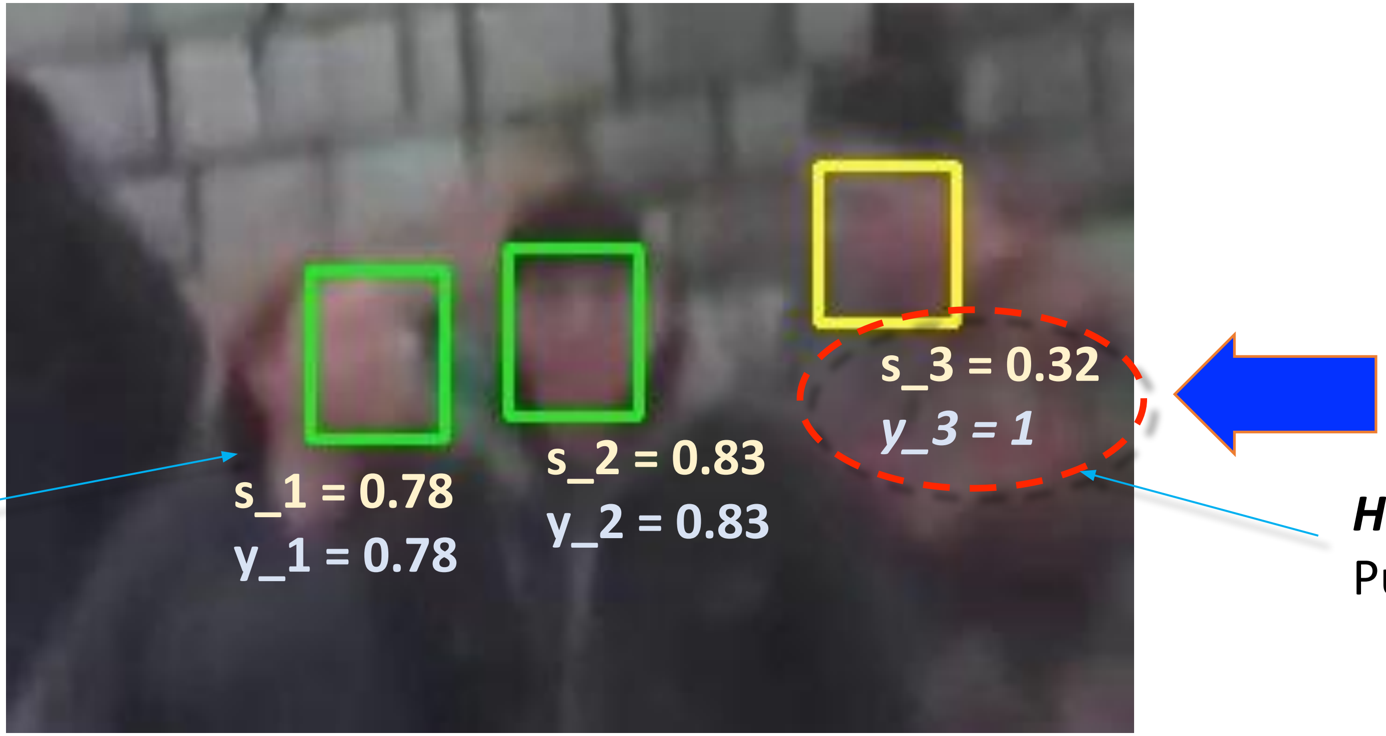
***Easy* positive:**
High-confidence
detector prediction

Images from [the open-sourced IJB–S dataset](#).

Distillation Training with soft labels and hard example

- ***Emphasize*** hard examples (label = 1)
- Enforce ***same prediction scores*** as baseline model for easy examples

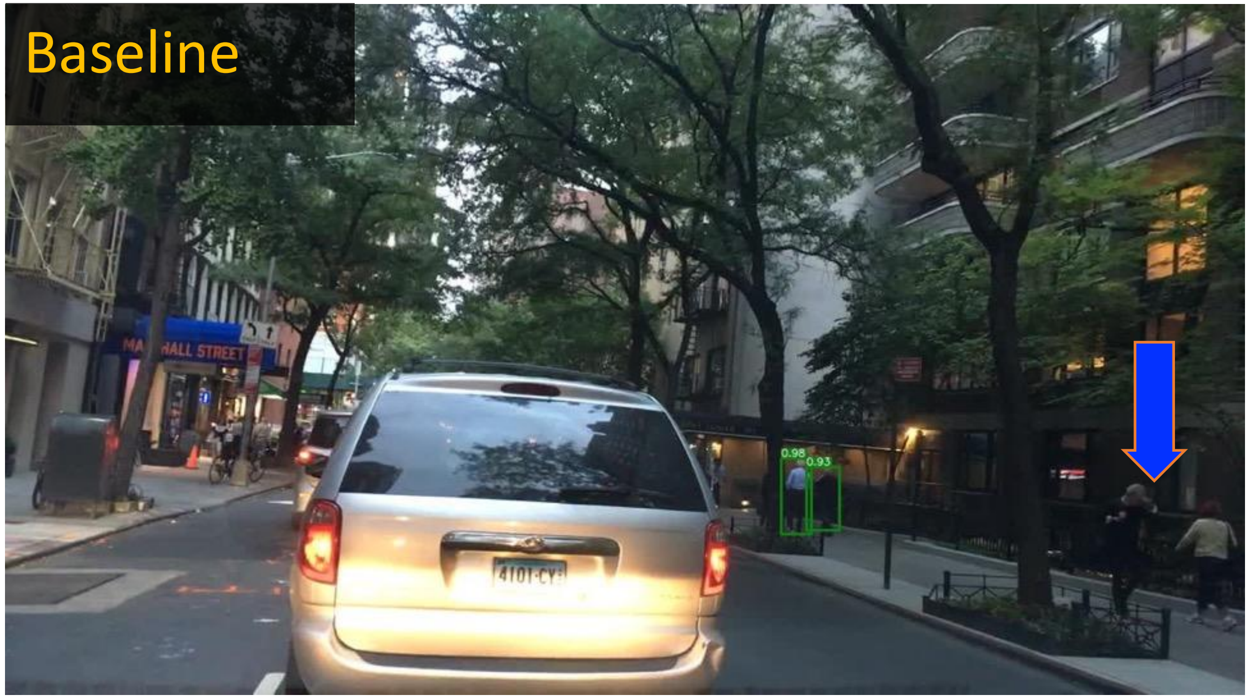
Easy positive:
Constrain to
remain same

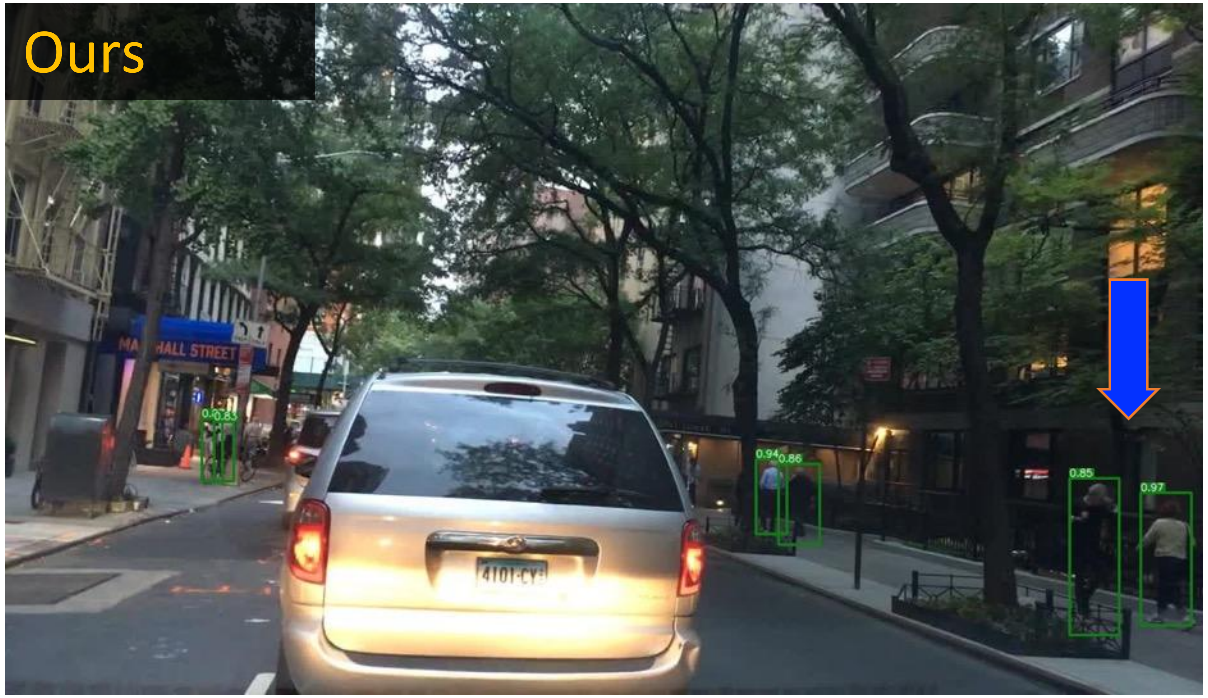


Hard positive:
Push up to “1”

Details in our CVPR paper.

Results on Berkeley deep drive (BDD 100k)





	Baseline: BDD(clear, daytime)	Our method (adapt with unsupervised videos)
AP on BDD (snowy, rainy, cloudy, dusk, evening)	15.21	28.43

Study II

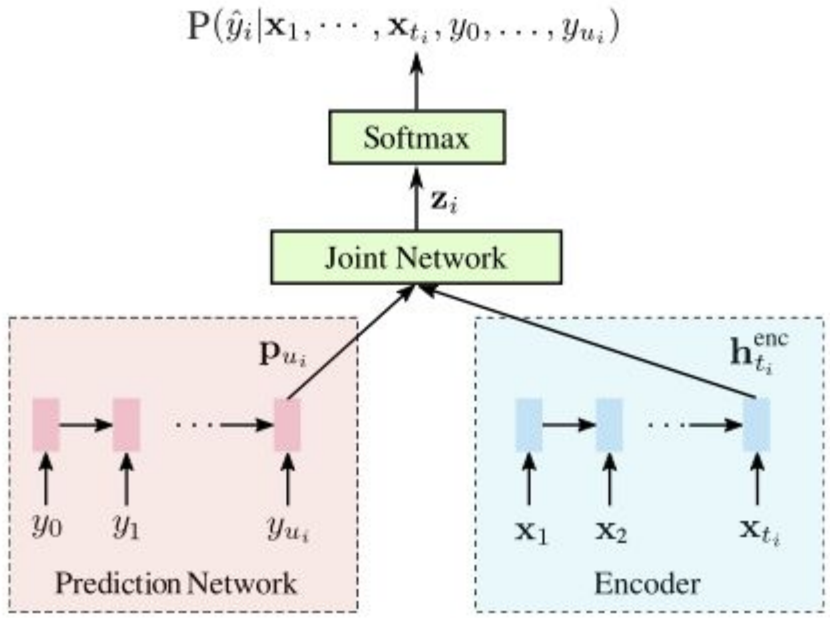
RNN-T models for out-domain speech recognition

RNN-T Models Fail to Generalize to Out-of-Domain Audio: Causes and Solutions, under submission, arXiv:2005.03271 (with Chung-Cheng Chiu, Arun Narayanan, Wei Han, Rohit Prabhavalkar, Yu Zhang, Navdeep Jaitly, Ruoming Pang, Tara N. Sainath, Patrick Nguyen, Yonghui Wu)

RNN-Transducer (RNN-T)

- Conventional speech recognition is composed of an acoustic model (AM) and a language model (LM)

- RNN-T is an end2end model that unifies AM and LM into one. It can be used for streaming purpose. Compared with conventional models, RNN-T is 10x smaller than the conventional AM+LM model.



Word error rate (WER) w/ deletion errors

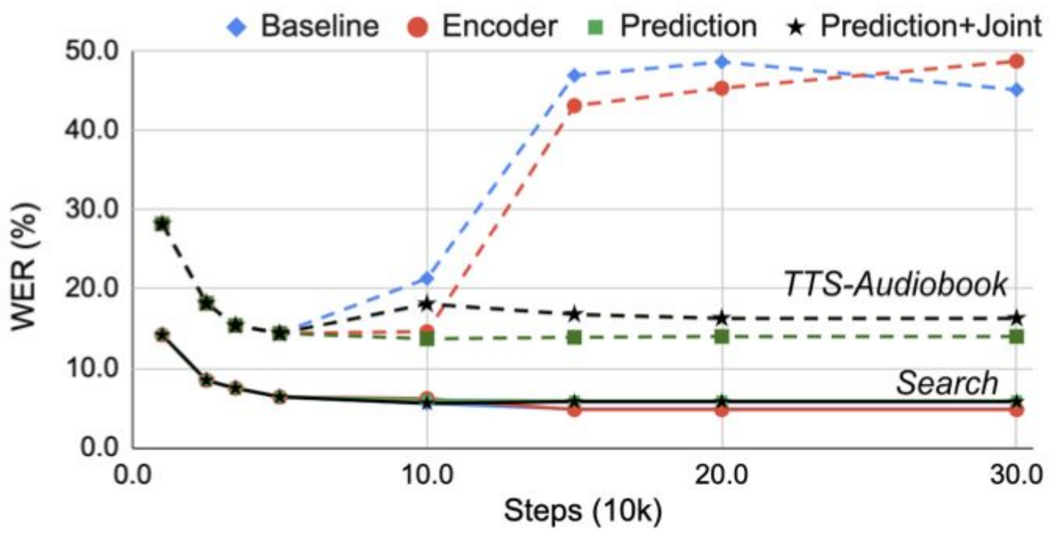
	Reg.
Librispeech test clean	3.2/0.2
Librispeech test other	7.8/0.7
YT-short	99.8/99.5

RNNT model trained from Librispeech get surprising deletion errors on YouTube audios.

The causes of high deletion errors

Since end-to-end models learn all components jointly, the effect is more pronounced than conventional models.

End- to-end models trained on one domain (e.g., short utterances) may not perform well for out-domain data (e.g., long utterances).



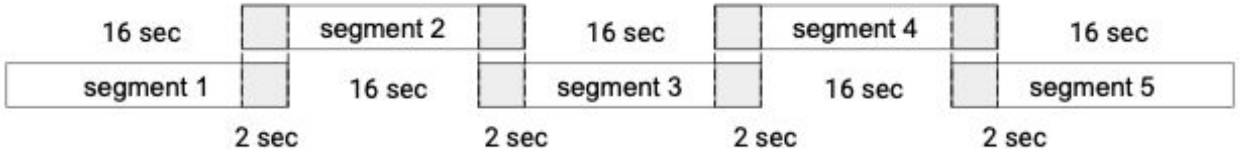
Solution

1. Cocktail of regularizating encoder during training
    - Variational noise
    - Random state sampling and random state passing
    - SpecAugment

Models	Search	TTS-Audiobook	YT-short
Baseline	4.9	48.6	67.0
VN	4.7	31.3	59.8
SpecAugment	**4.6**	16.5	52.9
+ RSP	5.1	**11.9**	27.3
+ RSP + VN	5.1	**11.9**	**25.3**

2. Dynamic overlapping inference (DOI)

	Reg.	DOI
Librispeech test clean	3.2/0.2	3.2/0.2
Librispeech test other	7.8/0.7	7.8/0.6
YT-short	99.8/99.5	33.0/3.6



Conclusion

- Out-of-domain data is often challenging for deep networks
- We may consider sequential data (videos and audios) and self-supervised learning for the out of domain problem.
- We have seen successful studies in

    - visual detection
    - speech recognition

The ultimate goal is to learn **one model which works in all scenarios.**