Domain-Invariant Speaker Vector Projection by Model-Agnostic Meta-Learning

Jiawen Kang

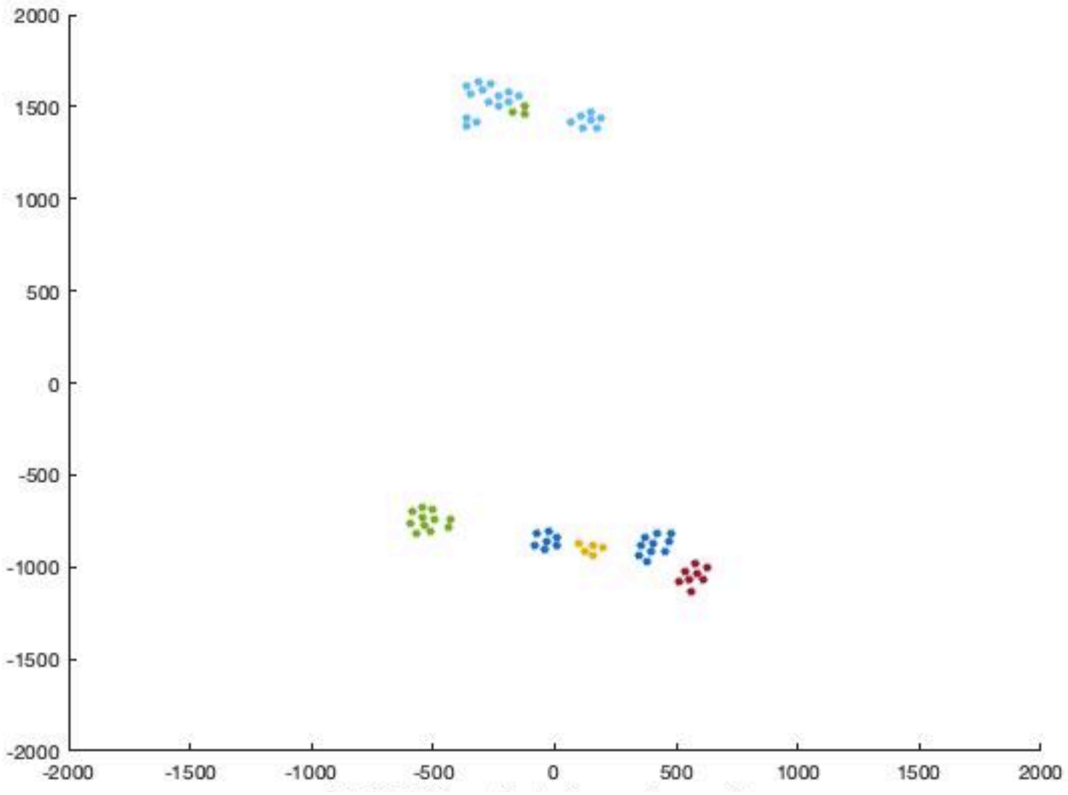
Ruiqi Liu

Problem statement

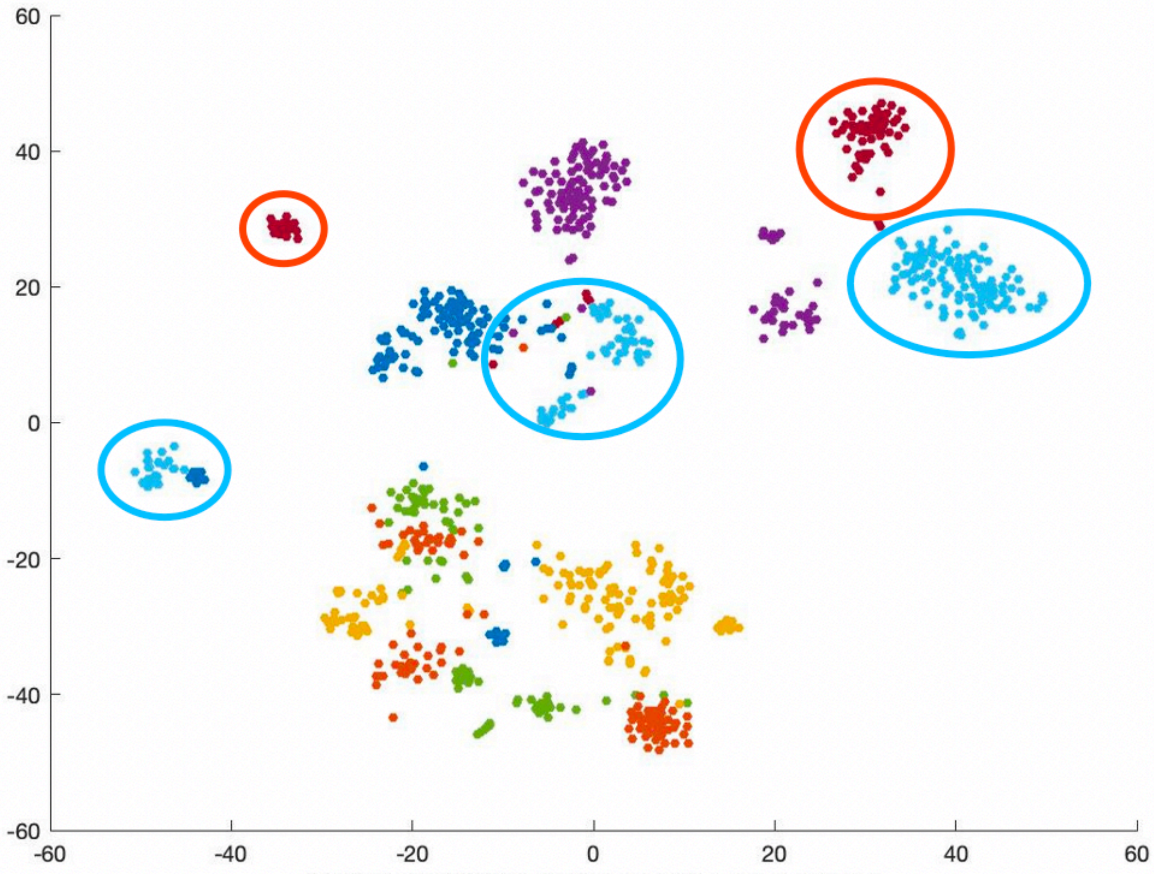
- CN-Celeb reveals the shortcoming of current VPR system:

  In unconstrained conditions, the performance of the current speaker recognition techniques might be much worse than it was thought.

VPR system performances

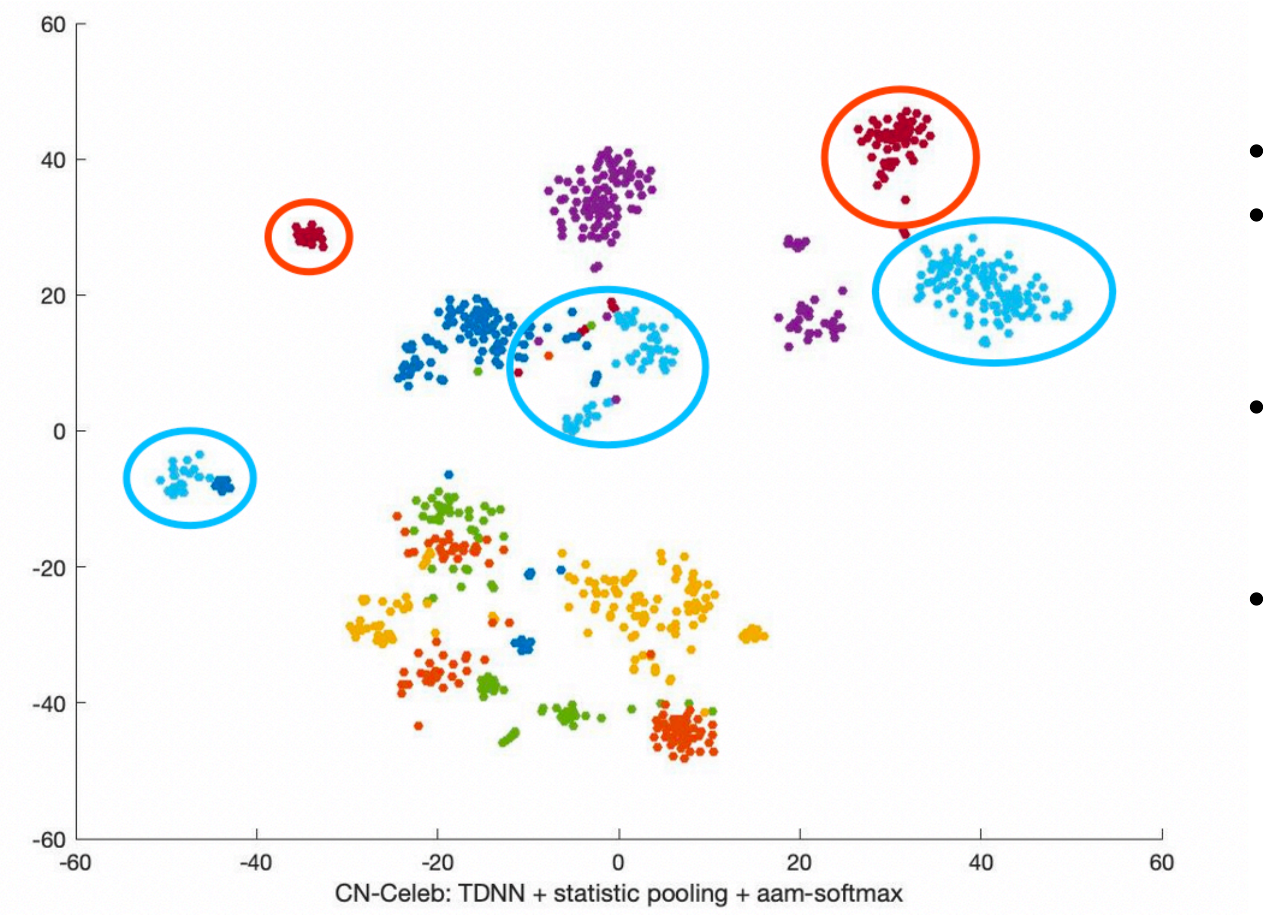


SITW



CN-Celeb

Domain shift problem



Domain shift

- Data from different genre have different domains.
- Speakers of a same domain tend to have the same domain.

- Traditional discriminative DNN method can only handle visible domain.

- We want to obtain a domain-Invariant speaker vector to enhance the robustness of speaker recognition system.

Solutions

- Direction: Meta-learning

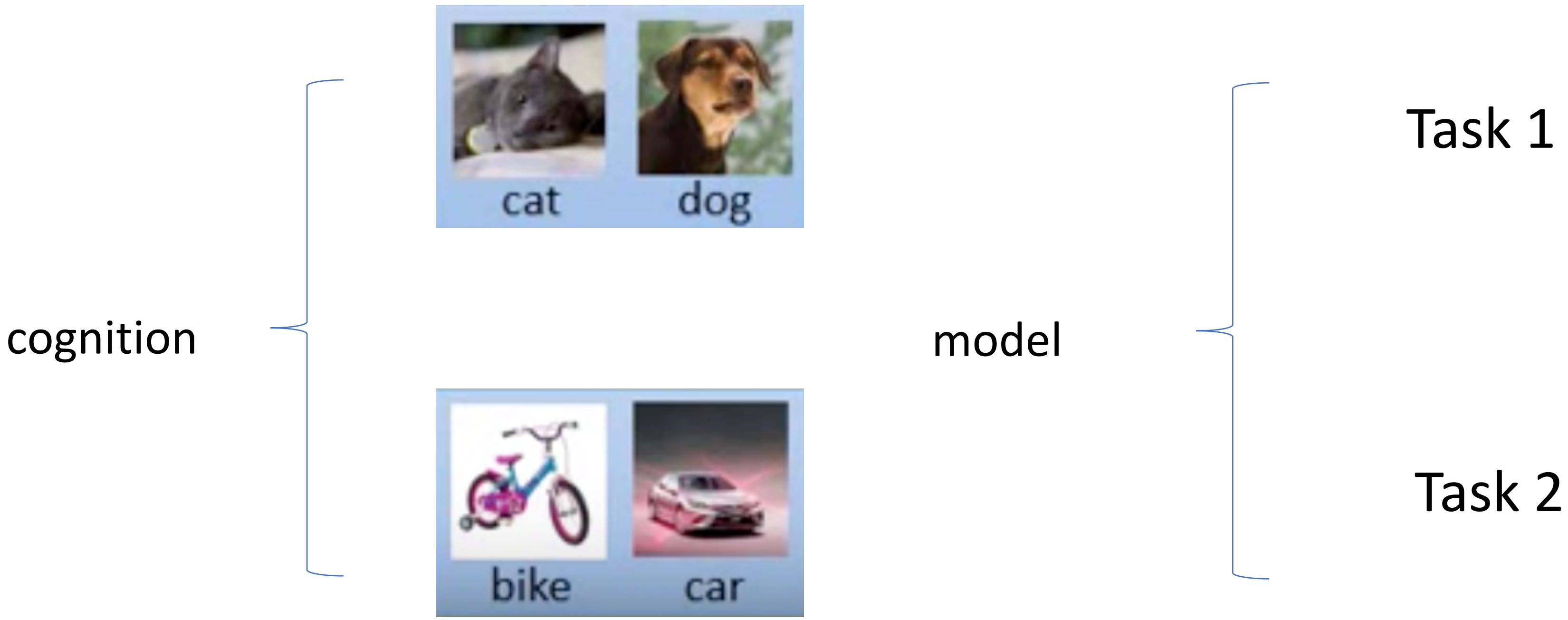
  Learning to learn: Fast adaptation trained models to new tasks.

- Approach:

  MAML: Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks

Meta-learning(元学习)

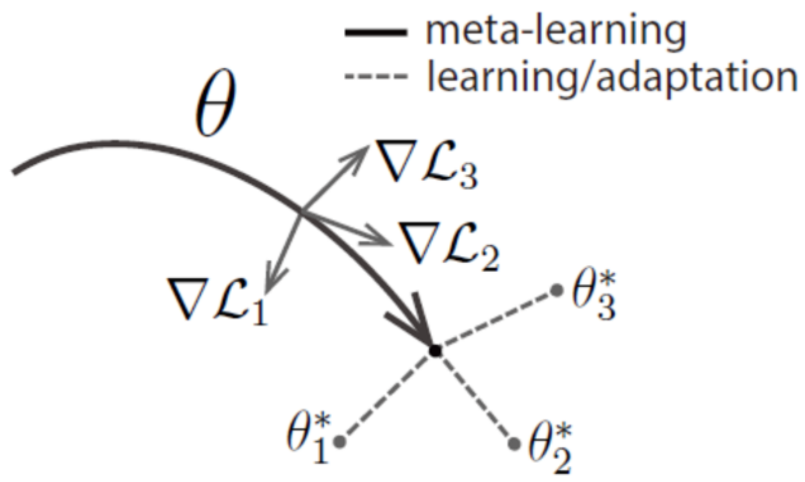
- Starting point：
  - After building cognition, human can learn things in a very fast speed, since human can learn more “**basic**” things.

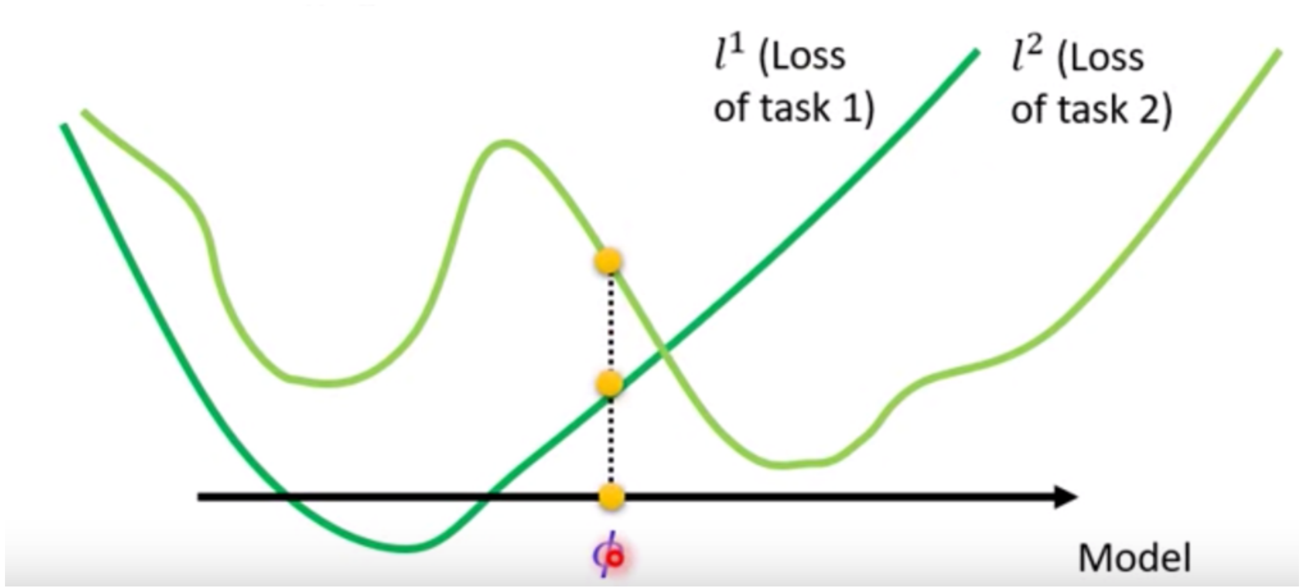


  - Meta leaning aims to build such a cognition to machine, make them learn how to learn.

Revisit MAML

- Finding an optimal initialization position θ (initial parameters), to fast reach the optimal position θ_n for task_n.



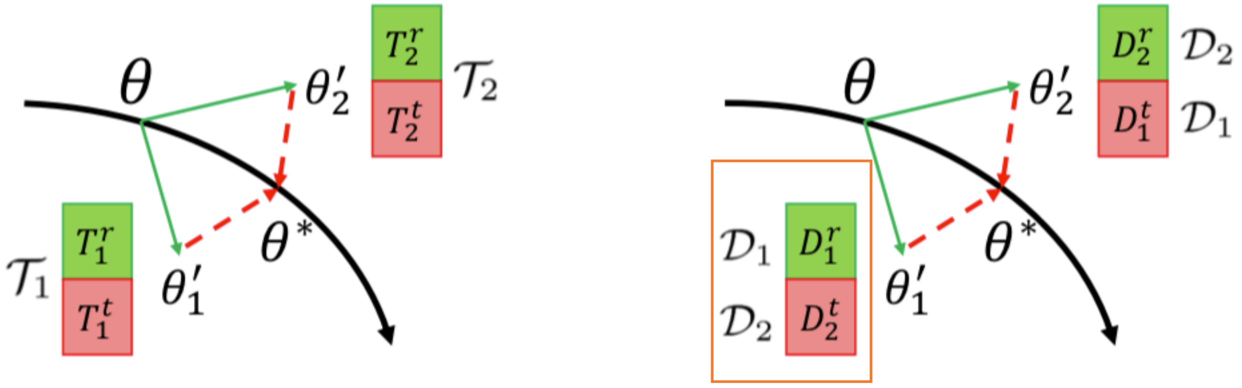


For our task

- Tasks -> domains: averaged optimal point for different domains, remain to be adaptation.

- Can be a generalization model:
    - Different domains follow a same task, the “initialization position θ” can be regarded as a domain-invariant optimal point, so the adaptation process is not necessary.

Robust MAML



(a) MAML (b) Robust MAML

$$\theta' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta; m_i^r)$$

$$\mathcal{L}_{\mathcal{T}_i}(f_{\theta'}; m_i^t) = \mathcal{L}_{\mathcal{T}_i}(f_{\theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta; m_i^r)}; m_i^t)$$

$$\theta \leftarrow \theta - \beta \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_{\theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta; m_i^r)}; m_i^t)$$

$$\theta' = \theta - \alpha \nabla_\theta \mathcal{L}(f_\theta; m_i^r)$$

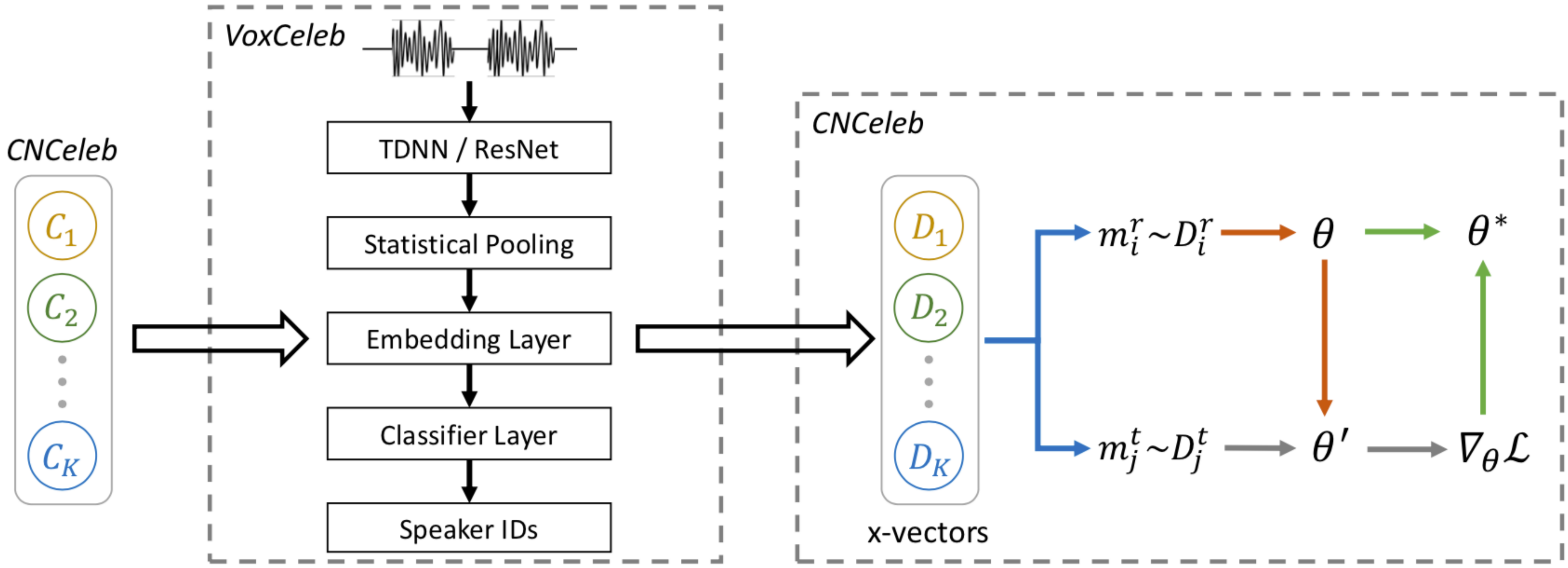
$$\mathcal{L}(f_{\theta'}; m_j^t) = \mathcal{L}(f_{\theta - \alpha \nabla_\theta \mathcal{L}(f_\theta; m_i^r)}; m_j^t)$$

$$\theta \leftarrow \theta - \beta \nabla_\theta \mathcal{L}(f_{\theta - \alpha \nabla_\theta \mathcal{L}(f_\theta; m_i^r)}; m_j^t)$$

Tricks:

- Single-speaker and multi-condition (SSMC) data is better.
- Balance the genre proportional is important. i.e. too much clear genre (like interview) will do harm to the meta training.

Domain-invariant projection net



(a) Deep speaker embedding

(b) Robust MAML

Experiments

name	Input x output
input	512
dense1	512 x 512 (embedding)
dense2	512 x 512
dense3	512 x 512
Softmax/Arcsoftmax	512 x 800

Input batch: Pair data from same speakers and different genre

Out-set genre: singing, movie, interview

Baseline

Table 1: *Performance (EER%) of the baseline systems.*

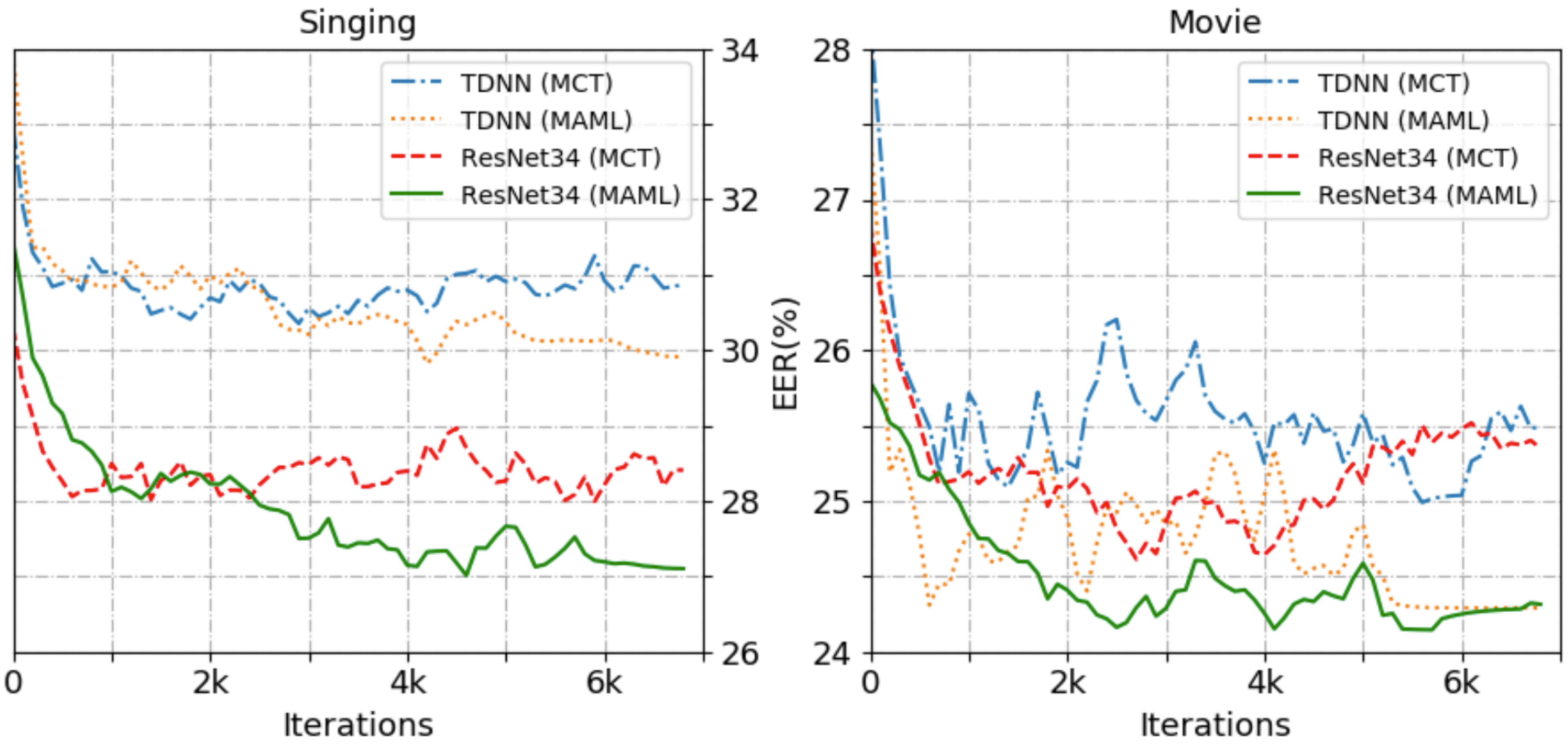
Test Set	TDNN		ResNet34	
	Cosine	LDA/PLDA	Cosine	LDA/PLDA
SITW.Eval.Core	5.139	2.433	3.226	1.968
CNC.Eval.Singing	29.95	26.88	28.47	27.18
CNC.Eval.Movie	26.09	20.24	25.19	21.29
CNC.Eval.Interview	19.68	15.97	19.23	15.47

Cosine Results

Table 2: *Performance (EER%) with cosine scoring.*

Cosine	TDNN			ResNet34		
Domain	Base	MCT	MAML	Base	MCT	MAML
Singing	29.95	30.85	29.86	28.47	28.40	**27.08**
Movie	26.09	25.46	24.27	25.19	24.92	**24.21**
Interview	19.68	17.51	**16.82**	19.23	16.92	16.87

EER Curves


Singing
Movie
TDNN (MCT)
TDNN (MAML)
ResNet34 (MCT)
ResNet34 (MAML)
EER(%)
Iterations
0
2k
4k
6k
26
28
30
32
34
24
25
26
27
28

PLDA Results

LDA/PLDA	TDNN			ResNet34		
Domain	Base	MCT	MAML	Base	MCT	MAML
Singing	25.67	25.50	25.35	23.83	23.66	**23.53**
Movie	19.63	18.74	18.85	18.19	**17.75**	**17.75**
Interview	13.63	13.47	13.58	12.05	**11.85**	**11.85**

Conclusion

- Robustness MAML is effective from the perspective of cosine EER.
- PLDA is partly helpful for domain shift problem.

Further discuss

- Ideally, Robust MAML doesn‘t require SSMC data.
  - Data volume
  - Domain number in each batch (“Meta batch”) is too small
- Metric-based meta learning.
  - The objective function can be more fit for back-end model.
- Light-weighted MAML-based adaptation
  - Adaptation to new domain.
  - Maybe enroll adaptation.