Karlsruhe Institute of Technology

Latent Space Knowledge Distillation under Heterogeneous Data

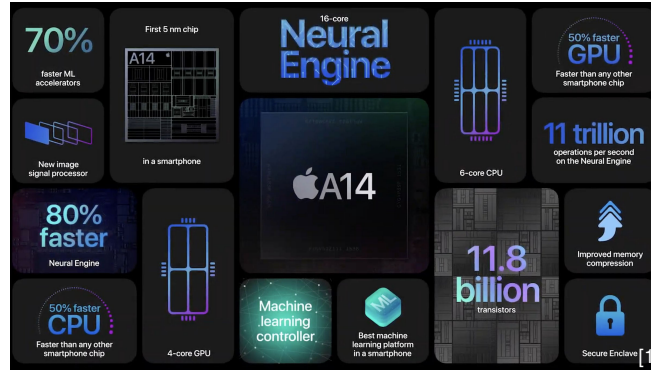
Master Thesis Presentation

Matthias Schmitt | 17.12.2021

Machine Learning at the Edge

- Neural co-processors in smartphones
- AI frameworks support edge devices
- Inference at the edge possible

- Big models with billions of parameter? Still out of reach?!
- What about training at the edge?



[1]

TensorFlow Lite [2]

PyTorch android [3]

Background: Knowledge Distillation

Hinton et al. "Distilling the Knowledge in a Neural Network" NIPS'14 DL workshop

- Compress large cumbersome model into new more efficient model
- Teacher guides the training of the student

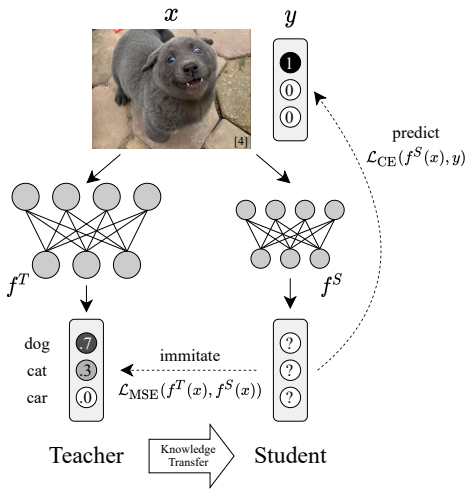
Background: Knowledge Distillation

Hinton et al. "Distilling the Knowledge in a Neural Network" NIPS'14 DL workshop

- Compress large cumbersome model into new more efficient model
- Teacher guides the training of the student
- Student model imitates teacher's output via mean squared error loss

$$(1-\alpha)\ \mathcal{L}_{\mathrm{CE}}(f^S(x), y) + \alpha\ \mathcal{L}_{\mathrm{MSE}}(f^T(x), f^S(x))$$

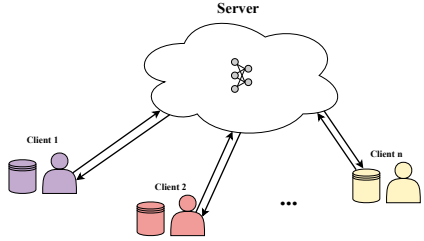
- Knowledge about the decision boundary is transferred



Background: Federated Learning

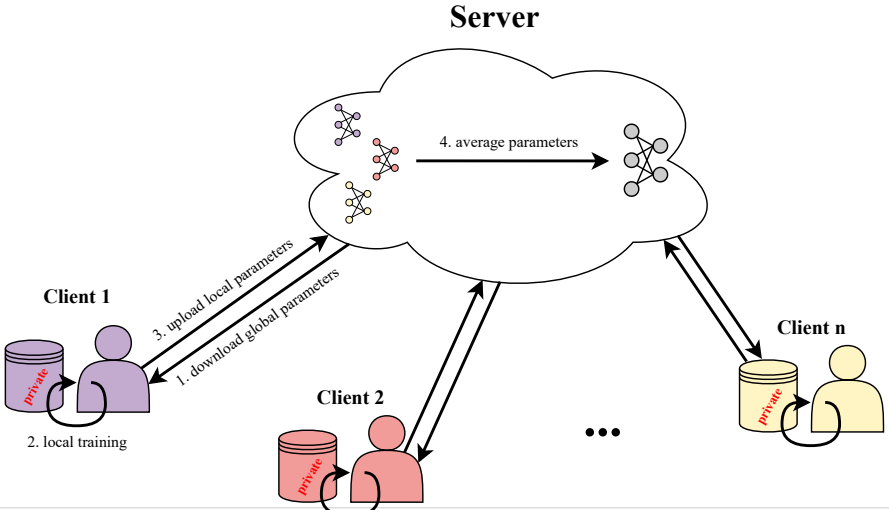
Training at the Edge

- Constrained clients collaboratively train a global model
- Combine computational resources and local data
- **But:** Private training data is not disclosed



Background: Federated Averaging

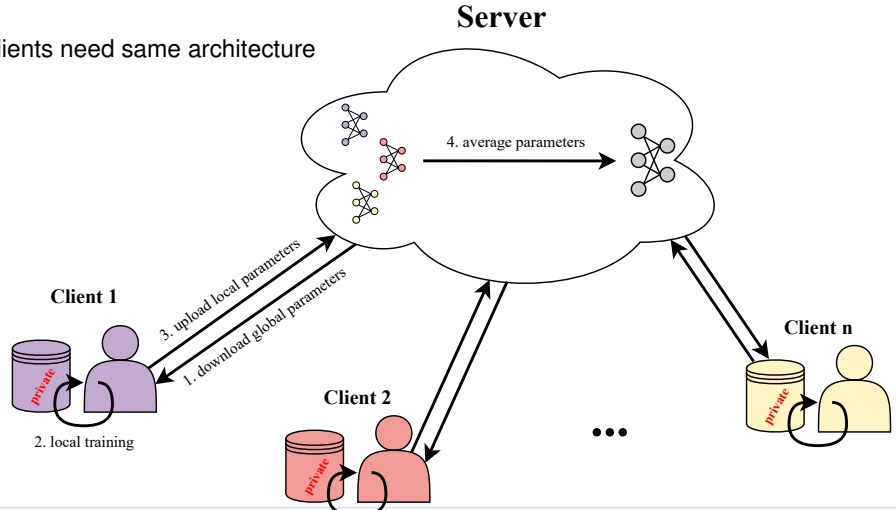
McMahan et al. "Communication-Efficient Learning of Deep Networks from Decentralized Data" AISTATS'17



Background: Federated Averaging

McMahan et al. "Communication-Efficient Learning of Deep Networks from Decentralized Data" AISTATS'17

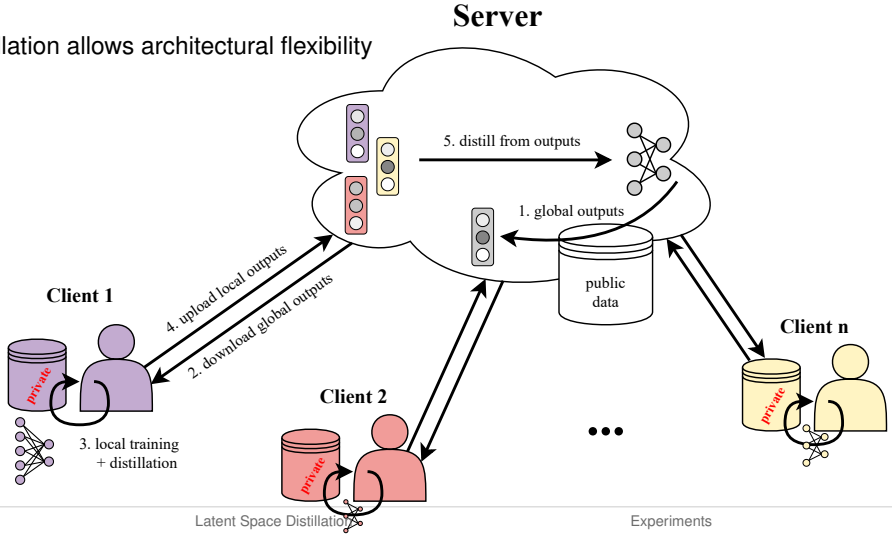
- All clients need same architecture



Background: Federated Distillation

Li et al. "FedMD: Heterogenous Federated Learning via Model Distillation" FL-NeurIPS'19

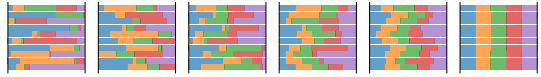
- Distillation allows architectural flexibility



Challenges for Federated Learning

Diverse participants with varying hardware in different environments

- Varying compute capabilities ⇒ Federated Distillation



- Data heterogeneity:
  not independent identically distributed (iid) over clients
- Data scarcity:
  clients have insufficient amount of local data

⇒ Distillation suffers, less incentive to collaborate

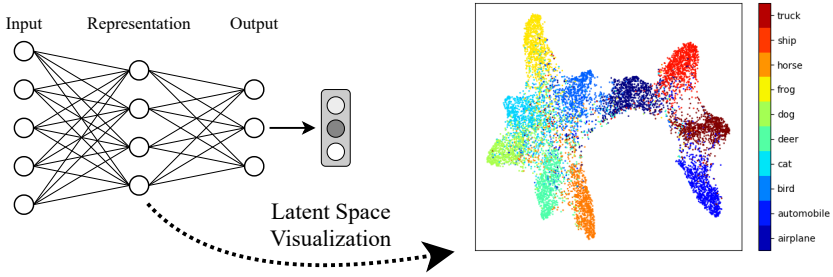
This work:

- Higher accuracy in non-iid cases by distilling *internal representations*

Latent Space Distillation

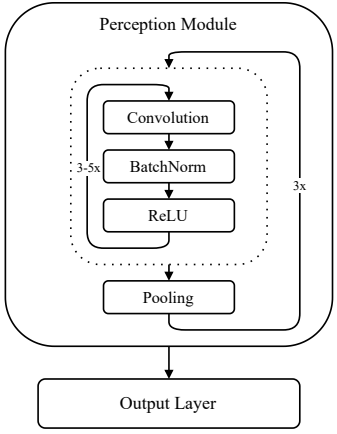
Latent Space
aka Embedding Space

- Lower-dimensional representation of high-dimensional data
- Deeper layers operate on more abstract interpretations
- *"The knowledge lies in the learned mapping from input vectors to output vectors."* — Hinton et al.



Architecture

- Perception modules differ in size

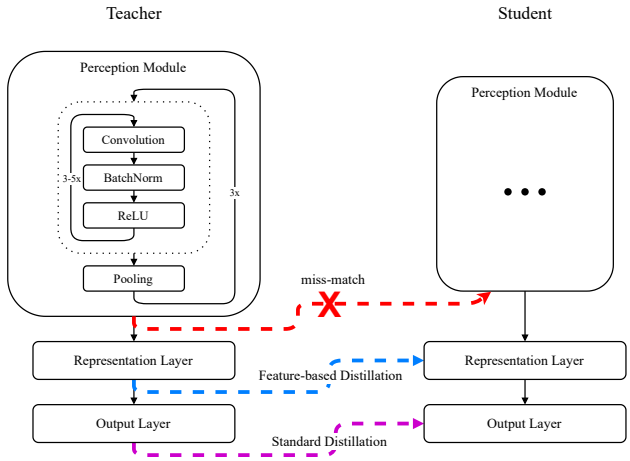


Architecture

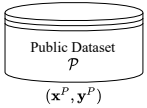
- Perception modules differ in size
  $\Rightarrow$ Latent space size miss-match

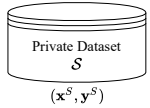
This work:

- Representation layer with output size l
  $\Rightarrow$ Feature-based distillation via MSE

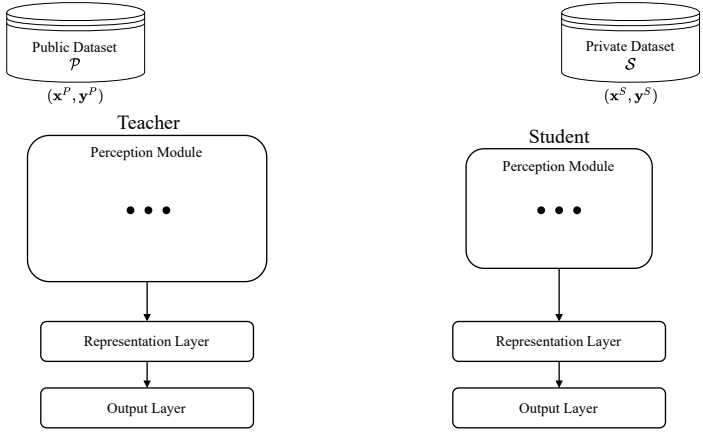


Latent Space Distillation

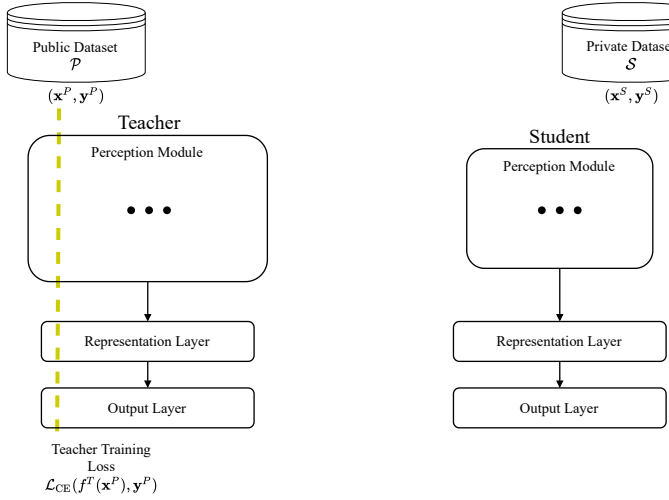




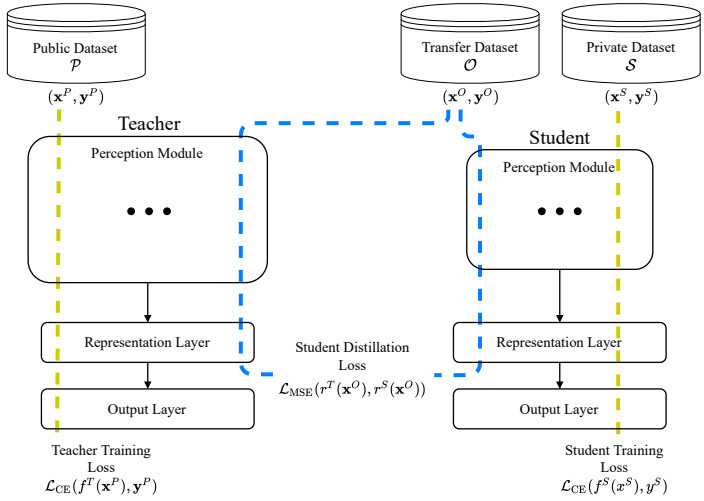
Latent Space Distillation



Latent Space Distillation



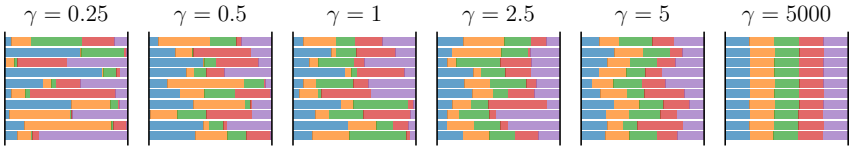
Latent Space Distillation



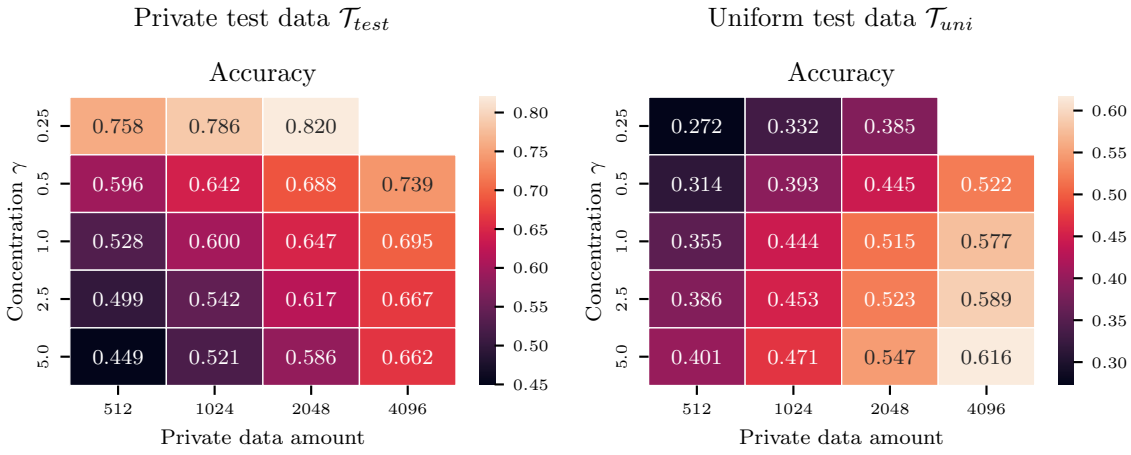
Experiments

Synthetic Heterogeneity

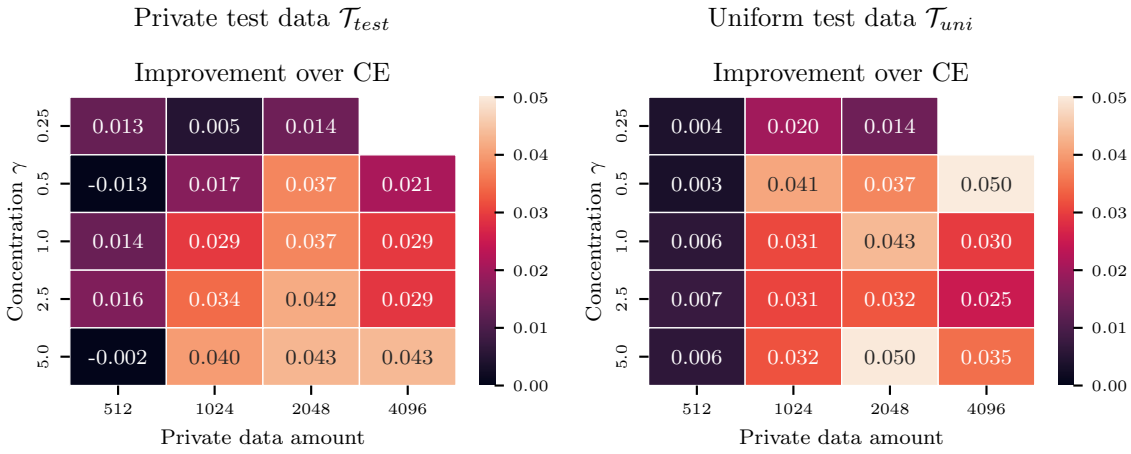
- Simulate realistic federated scenario
- Draw class distribution from Dirichlet distribution
  concentration parameter γ controls heterogeneity
- Sample from CIFAR-10 according to class distribution



Heterogeneity & Scarcity I

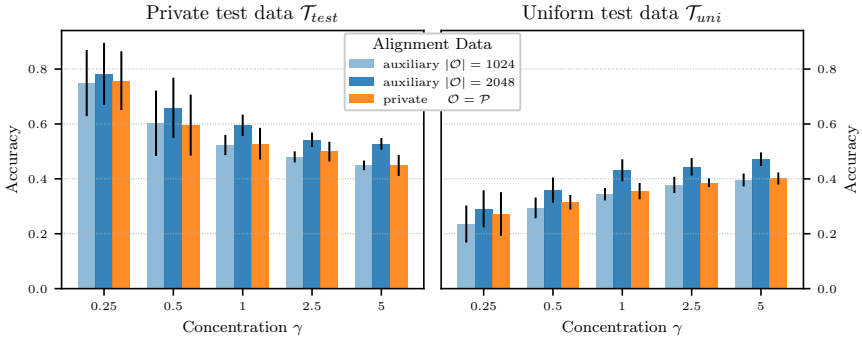


Heterogeneity & Scarcity II



Extrem Scarcity - Auxiliary Data

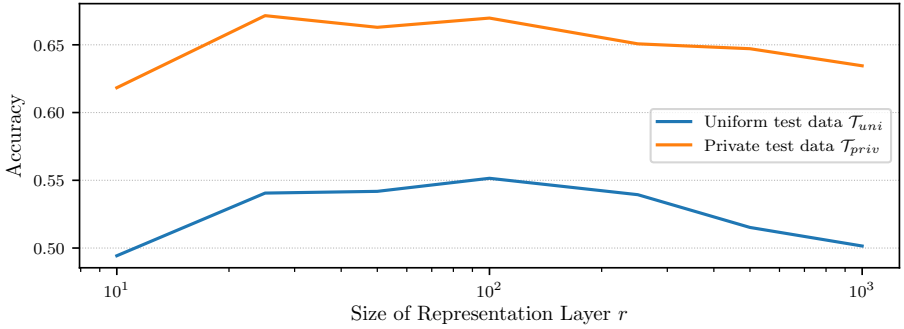
- 512 samples are not enough for distillation
- Common idea: distill on close auxiliary dataset
- Transfer set $\mathcal{O}$ subset of CIFAR-100



Latent Space Dimensionality

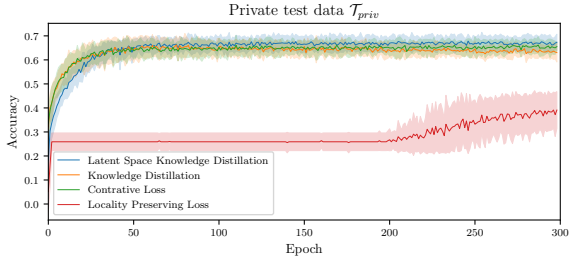
What size should the representation layer have? Is bigger always better?

- Best: similar size as teacher's perception module output
- Bigger does not compensate for the harder distillation
- Smaller loses expressiveness



Comparison with SOTA

Test data	CE	KD	CRD	LPD	**LKD**
$\mathcal{T}_{priv}$	0.634	0.631	0.655	0.391	**0.674**
$\mathcal{T}_{uni}$	0.492	**0.551**	0.530	0.183	0.544



- Contrastive Representation Distillation (CRD)
  Tian et al. “Contrastive Representation Distillation” ICLR’19
- Locality Preserving Distillation (LPD)
  Chen et al. “Learning Student Networks via Feature Embedding” TNNLS’21
- both do not require matching feature sizes

Main Findings & Conclusion

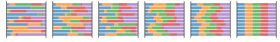
- Challenges in federated learning scenarios require robuster knowledge distillation
- Latent space distillation improves accuracy in heterogeneous data settings
- Insert representation layer size similar to teacher perception dimension
- Mitigate effects of extrem data scarcity with auxiliary data

Challenges for Federated Learning

Diverse participants with varying hardware in different environments

- Varying compute capabilities ⇒ Federated Distillation
- Data heterogeneity:
  not independent identically distributed (iid) over clients
- Data scarcity:
  clients have insufficient amount of local data

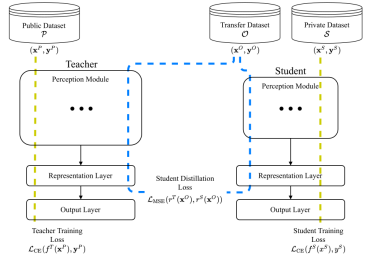
⇒ Distillation suffers, less incentive to collaborate



This work:

- Higher accuracy in non-iid cases by distilling *internal representations*

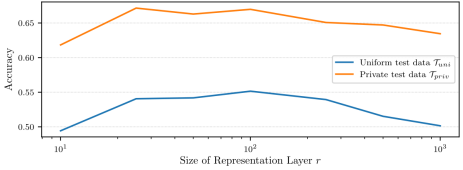
Latent Space Distillation



Latent Space Dimensionality

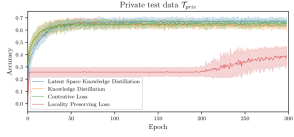
What size should the representation layer have? Is bigger always better?

- Best: similar size as teacher perception module output
- Richer abstractions does not compensate for the harder distillation
- Smaller layer loses expressiveness



Comparison

Test data	CE	KD	CRD	LPD	**LKD**
$\mathcal{T}_{priv}$	0.634	0.631	0.655	0.391	**0.674**
$\mathcal{T}_{uni}$	0.492	**0.551**	0.530	0.183	0.544



- Contrastive Representation Distillation (CRD)
  Tian et al. "Contrastive Representation Distillation" ICLR'19
- Locality Preserving Distillation (LPD)
  Chen et al. "Learning Student Networks via Feature Embedding" TNNLS'21
- both do not require matching feature sizes

Images Sources

[1|iphone a14] https://images.idgesg.net/images/article/2020/10/iphone-12-a14-100862081-orig.jpg
[2|TF lite] https://images.app.goo.gl/QZSHoDozCVNCEwMCA
[3|Pytoch mobile] https://towardsdatascience.com/object-detector-android-app-using-pytorch-mobile-neural-network-407c419b56cd
[4|dogcat] https://i.insider.com/5e600587a9f40c6f39701245