Warsaw University of Technology

Application of Domain Adaptation Techniques for Classifying Particles Basing On the Data From ALICE Experiments

Dawid Sitnik
Supervisor: Dr hab. inż. Tomasz Trzciński

Agenda

1. Problem Definition
2. Domain Adaptation
3. Models Evaluation
4. Identification of Unadapted Particles
5. Summary

Problem Definition

Aim

Classification of particles in Large Hadron Collider (LHC).

Data

Unlabeled data from detectors of LHC and labeled data with analogous attributes from simulations.

Problem

Different distributions of attributes from production and simulation datasets

Solution

Create a classifier which implements unsupervised domain adaptation techniques

Large Hadron Collider (LHC), photo 1.

Large Hadron Collider (LHC), photo 2.

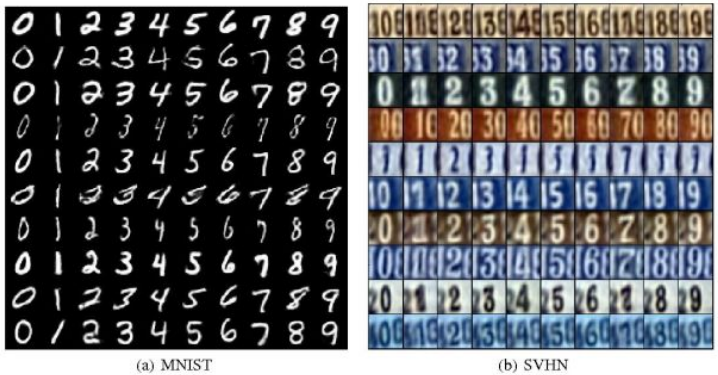
Unsupervised Domain Adaptation

Unsupervised domain adaptation is concerned with the problem of transferring the knowledge from a labeled source domain to unlabeled target domain, when both domains has different distributions of attributes.

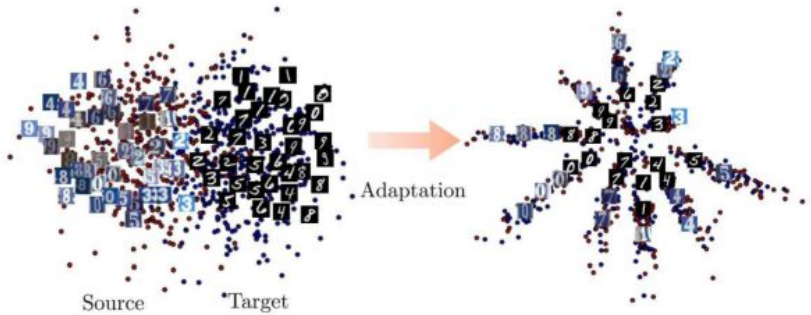
The aim of domain adaptation is to minimize a difference between distributions of both domains

Modern domain adaptation methods could be splitted into:

- *Adversarial*
- *Minimizing divergence metric*
- *Hybrid (adversarial + minimizing divergence metric)*



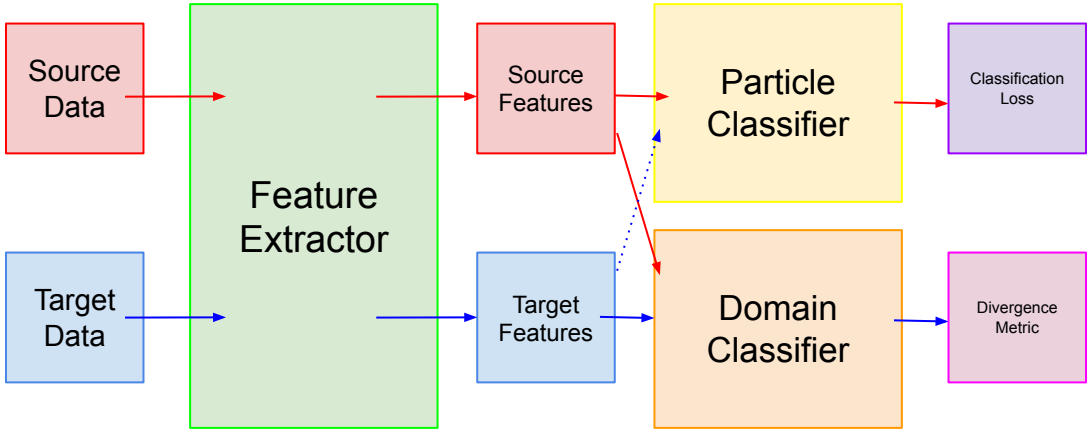
MNIST and SVHN datasets



Visualization of domain adaptation

Intuition Behind Models

Minimizing a Divergence Metric

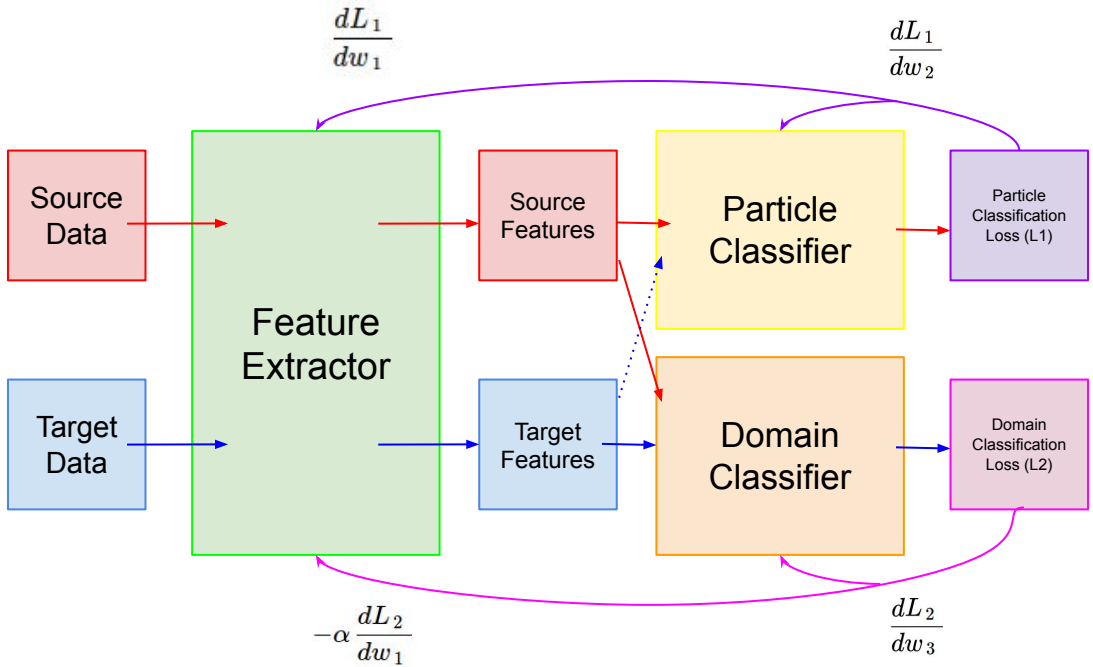


General scheme of model minimizing a divergence metric

loss =
classification loss + divergence metric

Intuition Behind Models

Adversarial Model



General scheme of adversarial model

Implemented Models

Adversarial:

- Conditional Adversarial Domain Adaptation Network (*CDAN*)
- Domain-Adversarial Network (*DANN*)

Minimizing divergence metric:

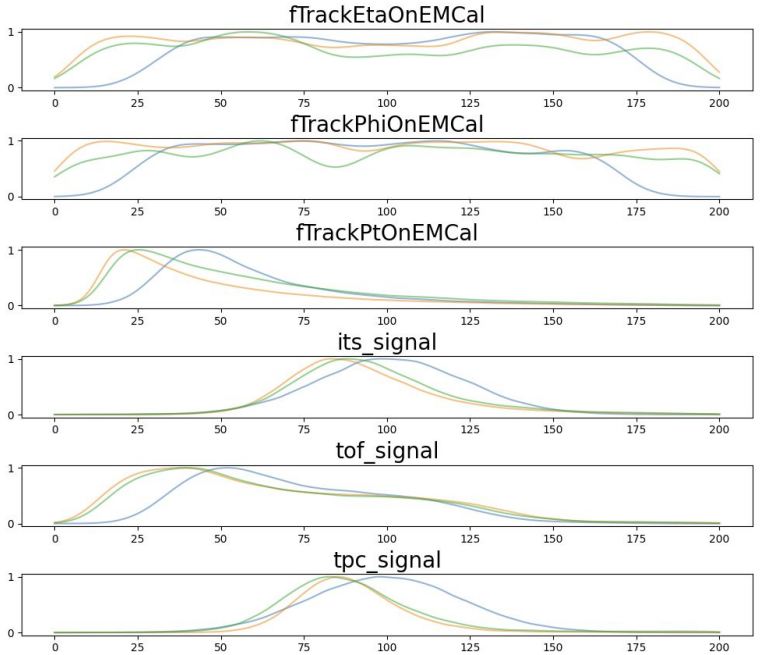
- The Multiple Kernel Maximum Mean Discrepancy Network (*MK-MMD*)
- Wasserstein Distance Guided Representation Learning (*WDGRL*)

Adversarial + minimizing divergence metric:

- Joint Adaptation Networks (*JAN*)
- The Margin Disparity Discrepancy Network (*MDD*)

Validation Method

- Production data from the ALICE detector does not contain labels, which disables direct evaluation of the models
- New dataset which is normalized data from simulations with an extra noise from a normal distribution N(0,0.005) is used for simulation



Distribution of model's input attributes. The green line marks distribution of production dataset, the yellow - simulation dataset and blue - perturbed dataset.

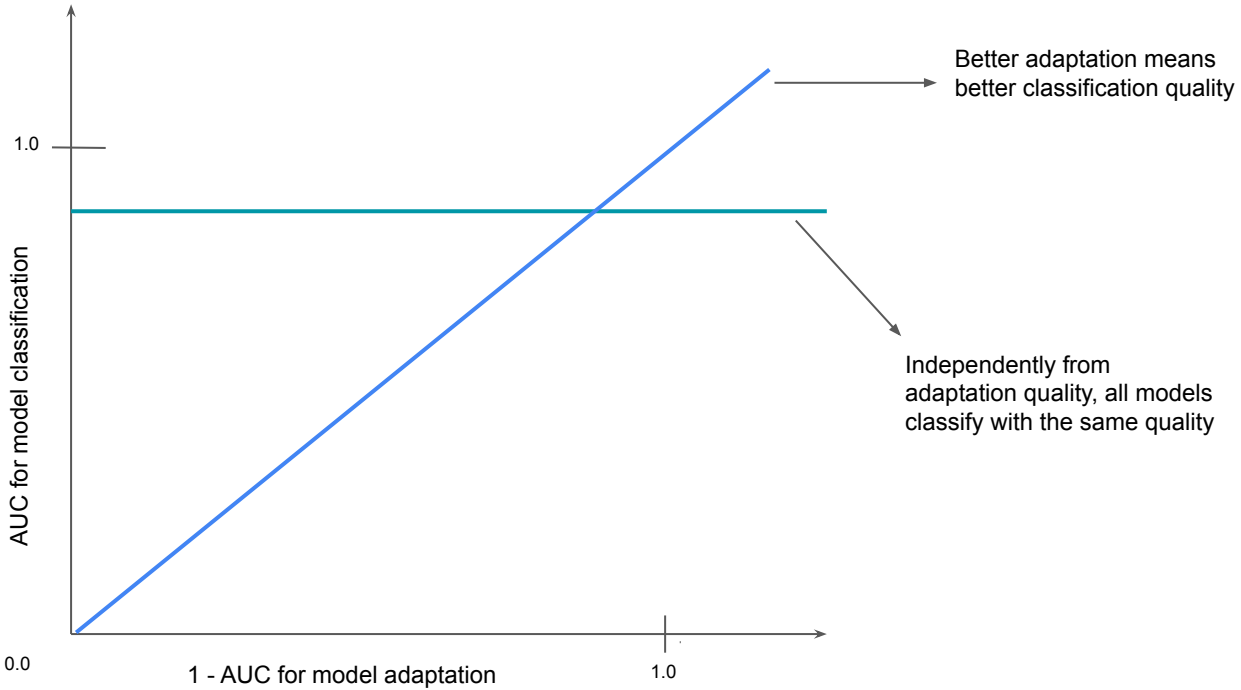
Validation Result

Model Name	Electrons	Pions	Protons	Kaons	Mean
Source	0.98	0.97	0.59	0.10	0.66
MDD	0.98	0.97	0.63	0.10	0.67
JAN	0.97	0.96	0.70	0.17	0.70
DANN	0.93	0.98	0.67	0.25	0.71
WDGRL	0.98	0.98	0.73	0.21	0.72
CDAN	0.96	0.99	0.69	0.28	0.73
DAN	0.98	0.99	0.77	0.32	0.77

Area Under Precision-Recall Scores for Each Model.

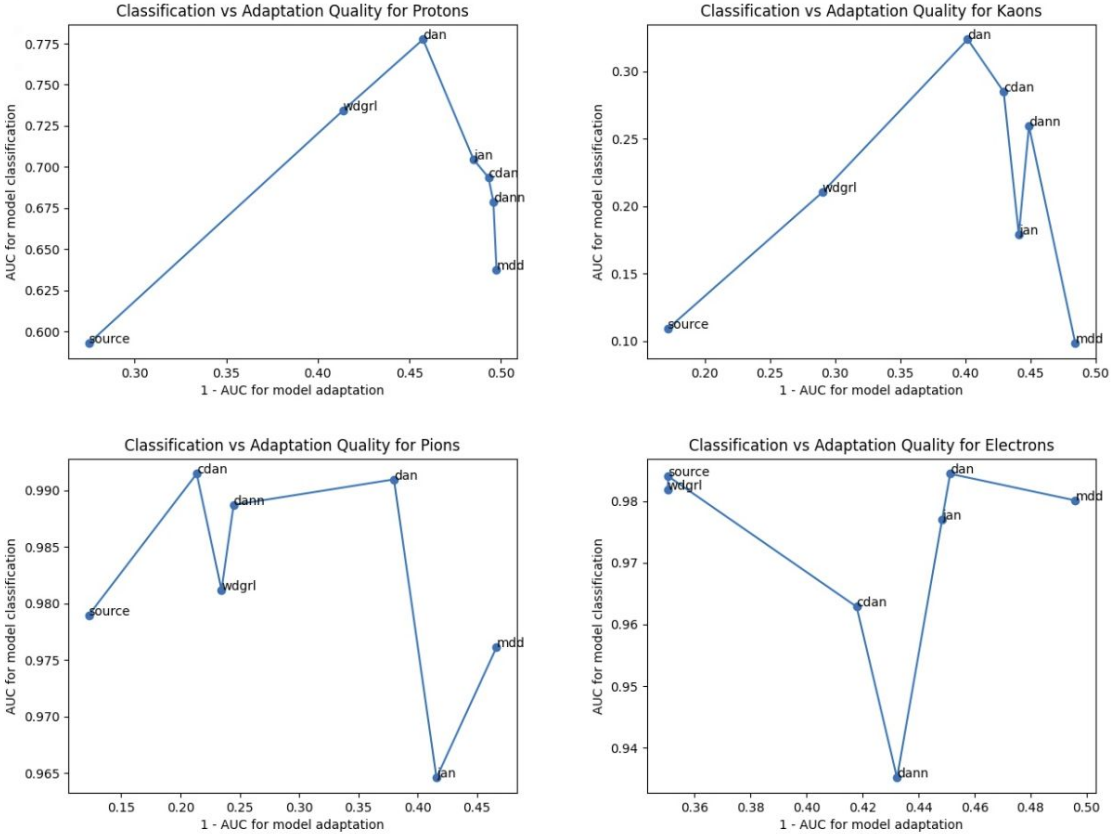
Classification vs Adaptation Quality

Intuition



Classification vs Adaptation Quality

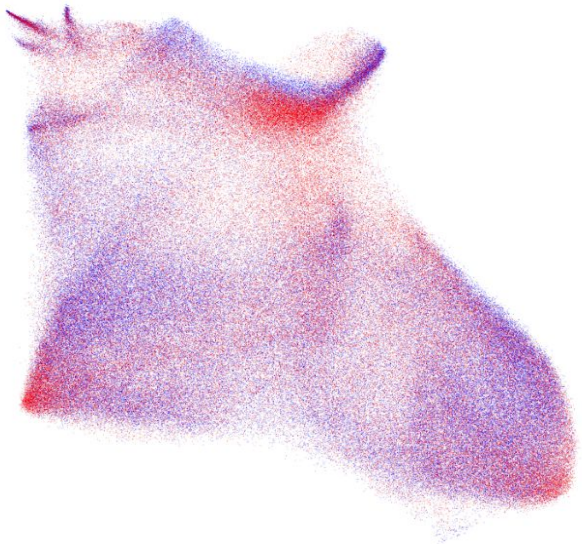
Results



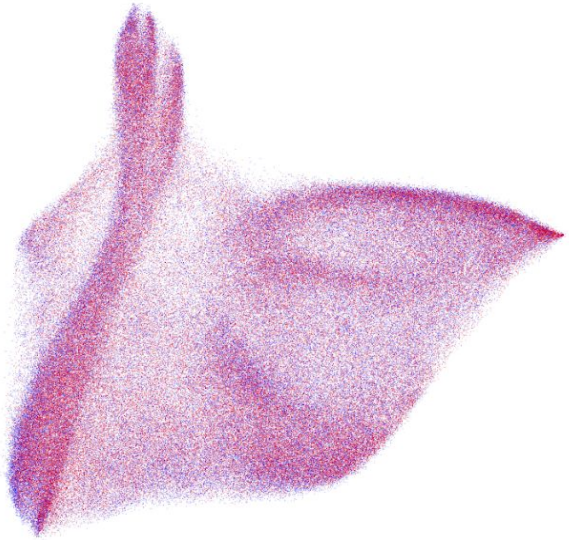
Comparison of classification and adaptation quality for each particle type.

Identification of Unadapted Particles

-



Visualization of 2-dimensional embedding from the feature extractor of the model without domain adaptation

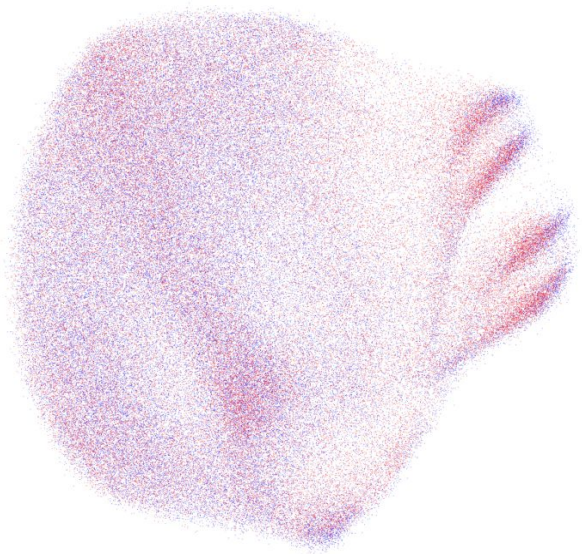


Visualization of 2-dimensional embedding from the feature extractor of the model with domain adaptation

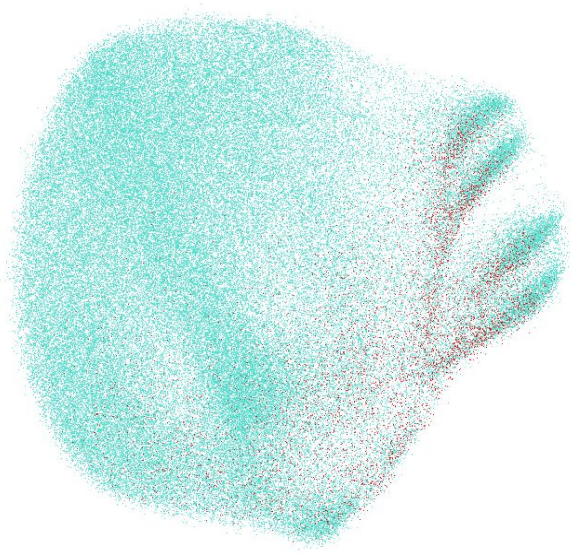
Identification of Unadapted Particles

1. Create a new dataset which is multidimensional (100 dimensions) representation of attributes used for classification (an output vector from the last layer of feature extractor of the model)
2. Grouping a new dataset into segments with similar characteristics
3. For each segment, calculating a metric which describes quality of a domain adaptation
4. Marking segments with the lowest value of that metric

Identification of Unadapted Particles



Visualization of particles from source (red) and target (blue) domains



Visualization of unadapted particles (red)

Summary

During the current work:

- It was proved that the usage of novel domain adaptation techniques can be applied in the field of high-energy physics.
- Six different domain adaptation models were implemented and compared.
- It was evidenced that using domain adaptation techniques does not always improves the classification and that a better quality of domain adaptation doesn't always yields better classification
- A new method based on machine learning algorithms that enable to mark unadapted particles was proposed.

Bibliography

- *Large Hadron Collider (LHC), photo 1.:*
  https://www.google.com/url?sa=i&url=https%3A%2F%2Fantyweb.pl%2Fcern-nastepaca-wielki-zderzacz-hadronow%2F&psig=AOvVaw3MpipLN160PE81p6Y82GVX&ust=1618345249199000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCKDIsv_D-e8CFQAAAAAdAAAAABAI
- *Large Hadron Collider (LHC), photo 2.:*
  https://www.google.com/url?sa=i&url=https%3A%2F%2Fhome.cern%2Fscience%2Faccelerators%2Flarge-hadron-collider&psig=AOvVaw3MpipLN160PE81p6Y82GVX&ust=1618345249199000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCKDIsv_D-e8CFQAAAAAdAAAAABAO
- *MNIST and SVHN datasets:*
  https://www.google.com/url?sa=i&url=http%3A%2F%2Fcookinglove.com%2Flr8u4rf%2Fsvhn-dataset.html&psig=AOvVaw01KanQXDuu4pv-GlYAyLaz&ust=1618345376923000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCPjm977E-e8CFQAAAAAdAAAAABAD
- *Visualization of domain adaptation:*
  https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcS6IZhThDWX_PcNR5uy6zaw7krs_oA84xvHpg&usqp=CAU

Questions