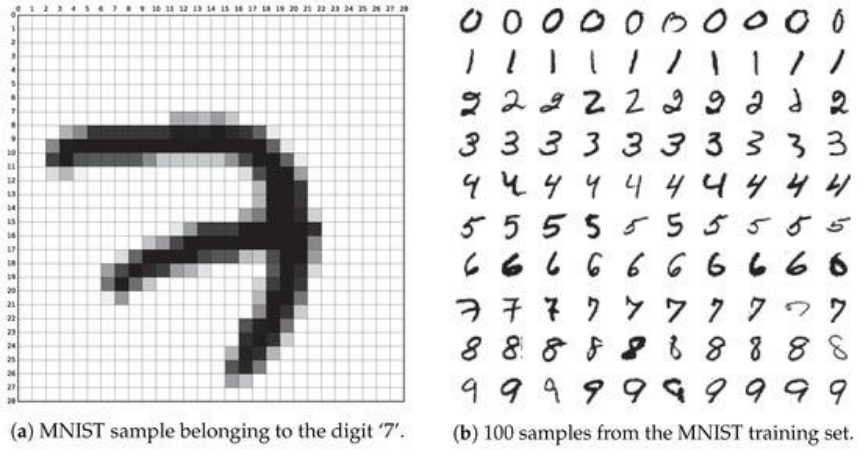
Application of the Lottery Ticket Hypothesis in NLP and Early Pruning

TECHNISCHE UNIVERSITÄT DARMSTADT

Intermission



(a) MNIST sample belonging to the digit '7'. (b) 100 samples from the MNIST training set.

Source: https://www.mdpi.com/applsci/applsci-09-03169/article_deploy/html/images/applsci-09-03169-g001-550.jpg

Source: https://www.bonaccorso.eu/wp-content/uploads/2016/07/28019400581_e1eb13ccc8_b.jpg

Tim Unverzagt

Prof. Dr.-Ing. Ralf Steinmetz
KOM - Multimedia Communications Lab

© author(s) of these slides including research results from the KOM research network and TU Darmstadt; otherwise it is specified at the respective slide

Structure

TECHNISCHE UNIVERSITÄT DARMSTADT

Introduction

Motivation

Background

Related Work

Task Definition

Progress

Outlook

Structure

TECHNISCHE UNIVERSITÄT DARMSTADT

Introduction

Motivation

Background

Related Work

Context of the thesis

Task Definition

Progress

Outlook

Content of the thesis

Initial Thoughts

- Many good reasons to initialize & train neural networks with many parameters
- Empirical evidence that many networks can be reduced after training while maintaining performance
  - aka. "Pruning"
  - Desirable due to bias towards small models (Ockham's razor)

Initial Thoughts

- Many good reasons to initialize & train neural networks with many parameters
- Empirical evidence that many networks can be reduced after training while maintaining performance
  - aka. "Pruning"
  - Desirable due to bias towards small models (Ockham's razor)
- Main Question:
  - "How important are the pruned weights during training?"

Lottery Ticket Hypothesis

- Due to the sheer number of "small" subnetworks in a "huge" network at least one such subnetwork has been initialized favorably for the given task
  - Name said subnetwork "lottery-ticket"

Lottery Ticket Hypothesis

- Due to the sheer number of "small" subnetworks in a "huge" network at least one such subnetwork has been initialized favorably for the given task
  - Name said subnetwork "lottery-ticket"
- The lottery-ticket can be trained as a standalone-network achieving a similar performance to the original

Lottery Ticket Hypothesis

- Due to the sheer number of "small" subnetworks in a "huge" network at least one such subnetwork has been initialized favorably for the given task
  - Name said subnetwork "lottery-ticket"
- The lottery-ticket can be trained as a standalone-network achieving a similar performance to the original
- A lottery-ticket can be identified by analysis of the fully trained original network
  - e.g. keeping only the weights of the largest magnitude finds a lottery ticket

Motivation

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

Mot

Back

Task

Pro

Out

Motivation

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

Mot

Back

Task

Pro

Out

Time & Memory

- Speedup during execution just as regular pruning
  - But remarkable compression rate: up to ~50x
- Decrease in memory usage during execution
- Possible speedup during development
  - There might be a way to identify lottery tickets early

Motivation

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

Mot

Back

Task

Pro

Out

Time & Memory

- Speedup during execution just as regular pruning
- Decrease in memory usage during execution
- Possible speedup during development
  - There might be a way to identify lottery tickets early

Interpretability

- Understanding lottery-tickets might enhance our knowledge of neural networks in general
  - Finding a way to identify a lottery-ticket might help understanding how exactly neural networks learn

Background – Unsupervised Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

Mot

Back

Task

Pro

Out

Background – Unsupervised Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

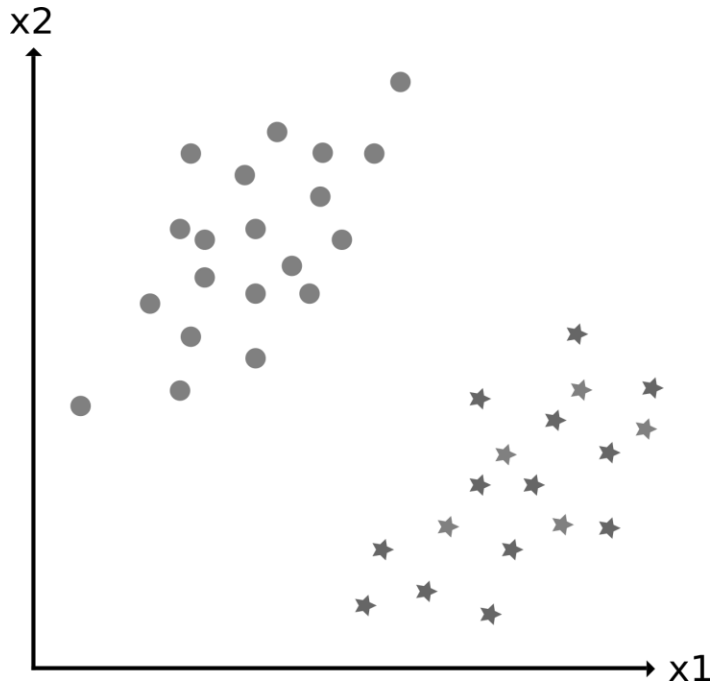
Mot

Back

Task

Pro

Out



Source: Produced by the author

Background – Unsupervised Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

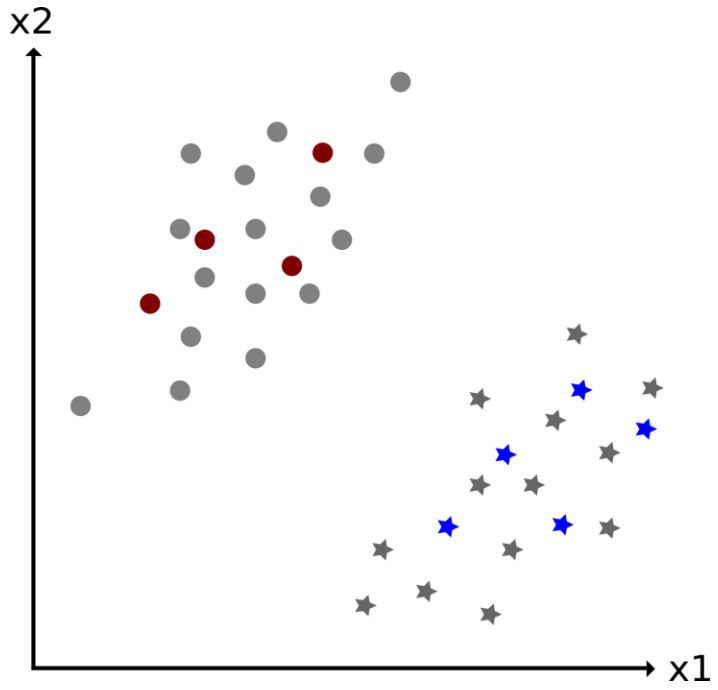
Mot

Back

Task

Pro

Out



Source: Produced by the author

Background – Unsupervised Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

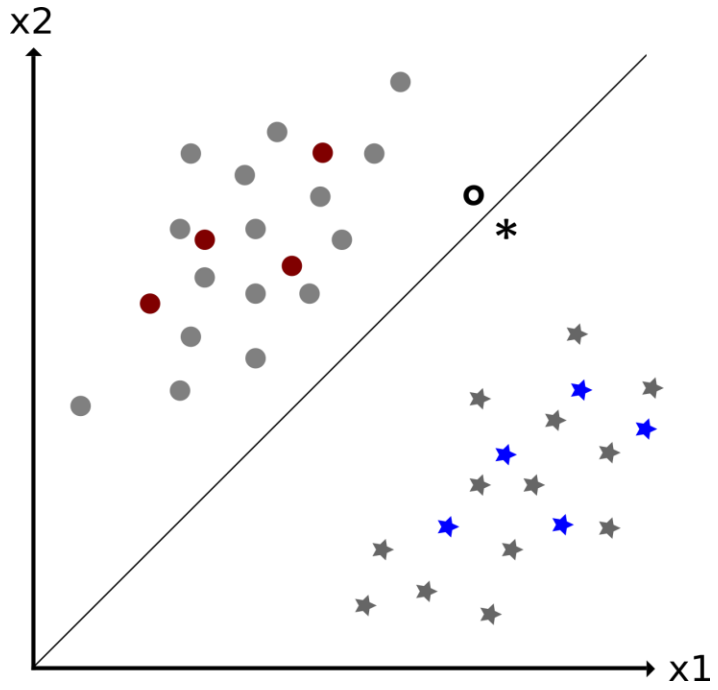
Mot

Back

Task

Pro

Out



Source: Produced by the author

Background – Unsupervised Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

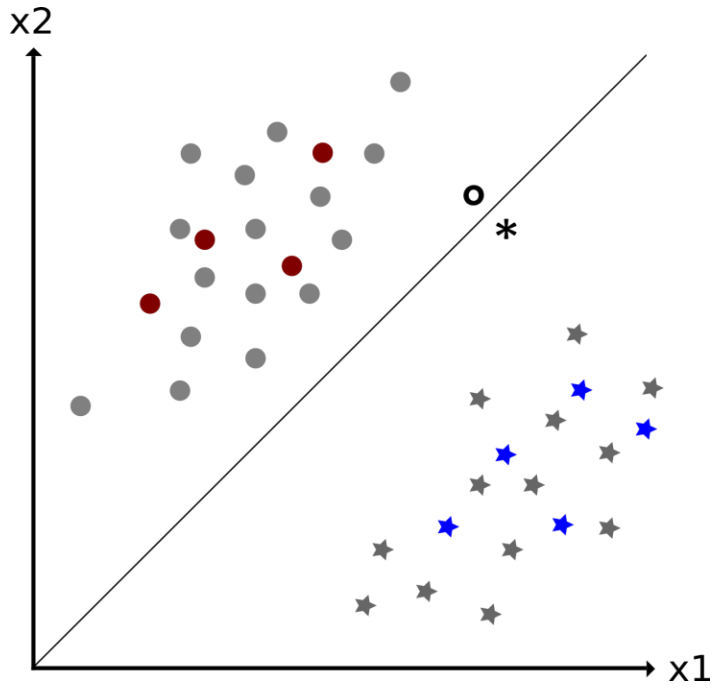
Mot

Back

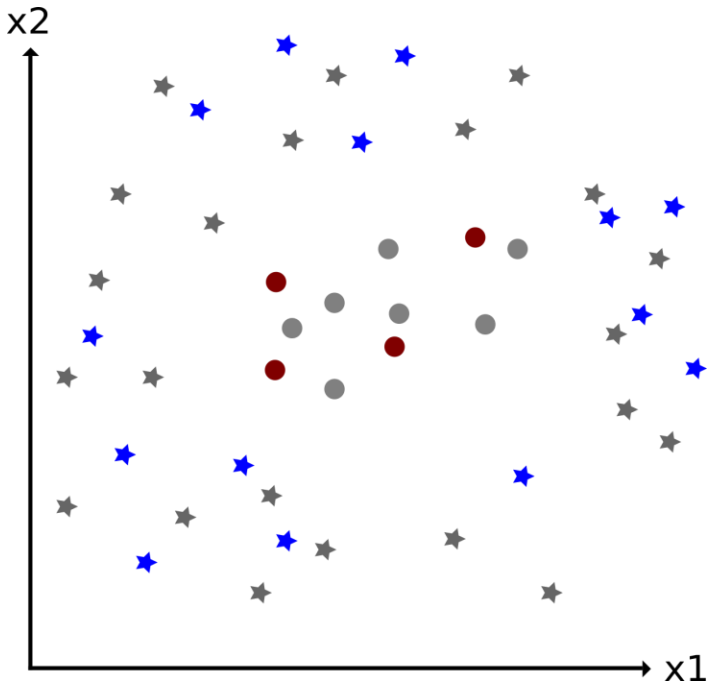
Task

Pro

Out



Source: Produced by the author



Source: Produced by the author

Background – Unsupervised Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

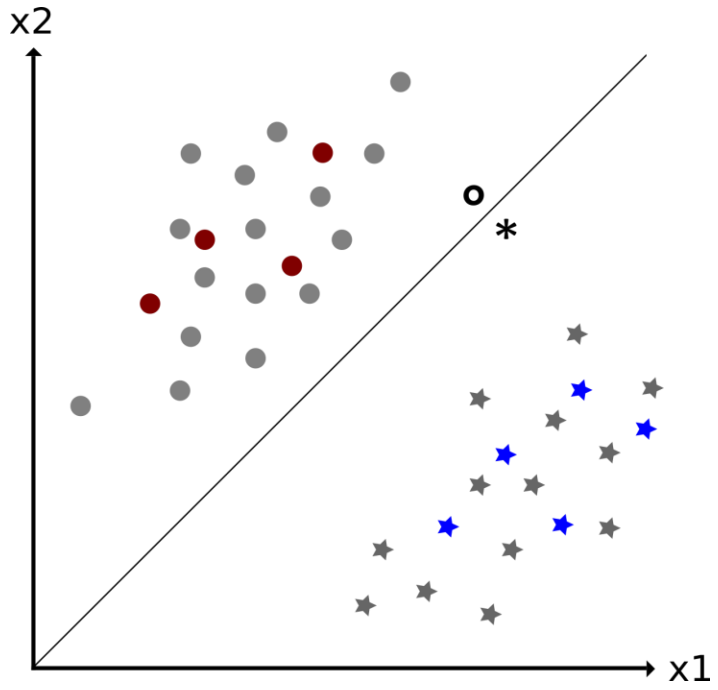
Mot

Back

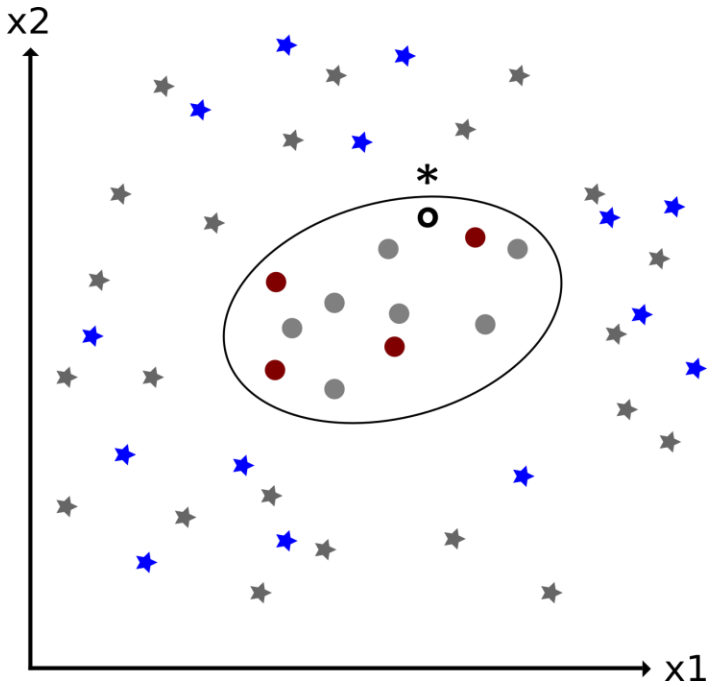
Task

Pro

Out



Source: Produced by the author



Source: Produced by the author

Background – Unsupervised Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

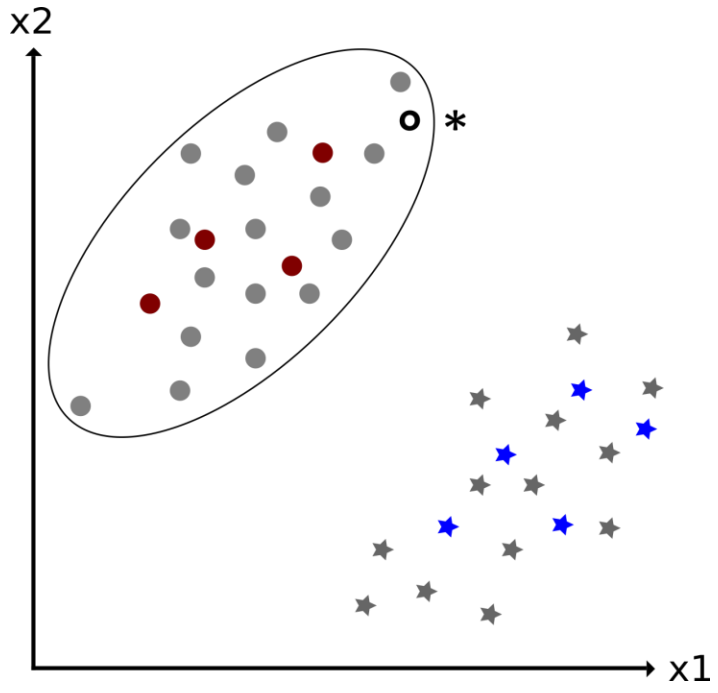
Mot

Back

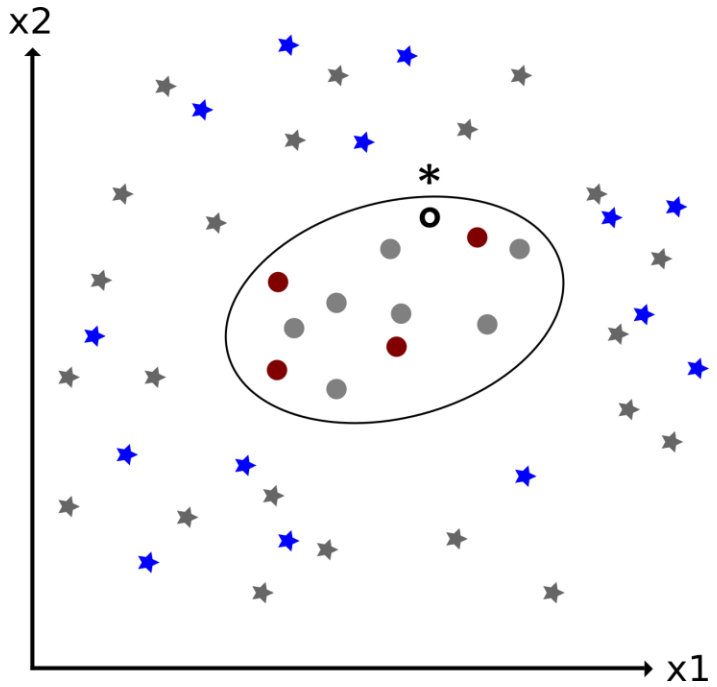
Task

Pro

Out

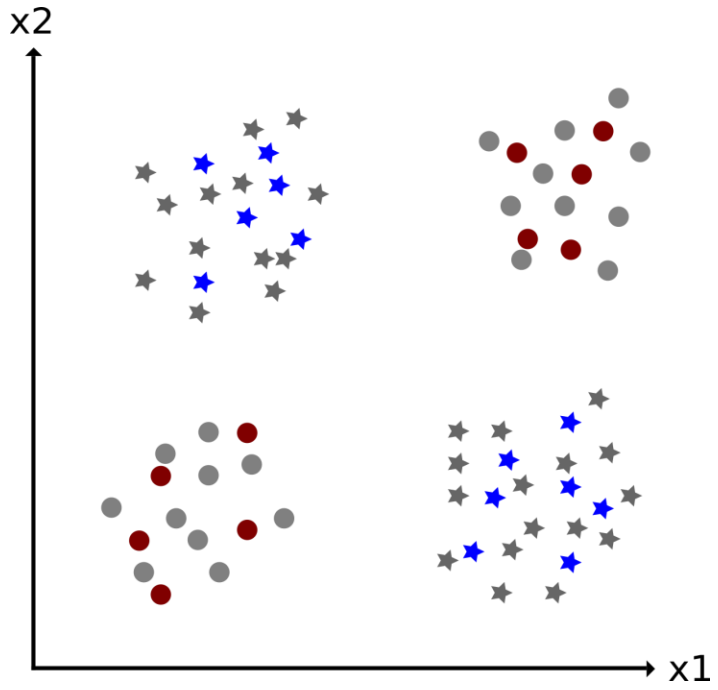


Source: Produced by the author



Source: Produced by the author

Background – Unsupervised Learning



Source: Produced by the author

Background – Unsupervised Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

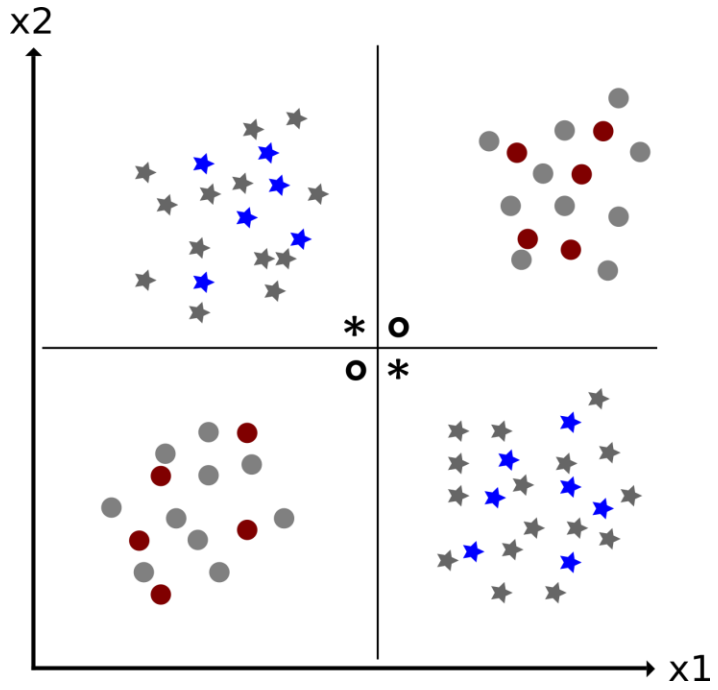
Mot

Back

Task

Pro

Out



Source: Produced by the author

Background – Unsupervised Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

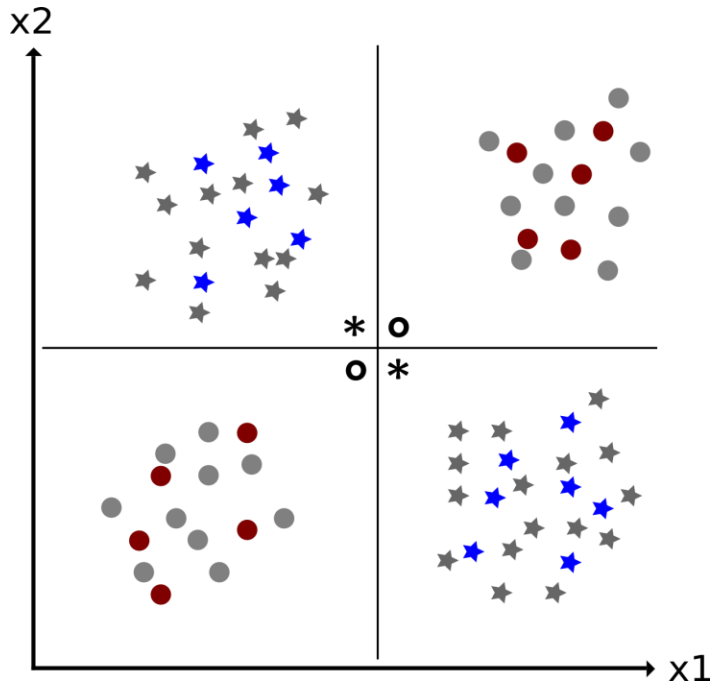
Mot

Back

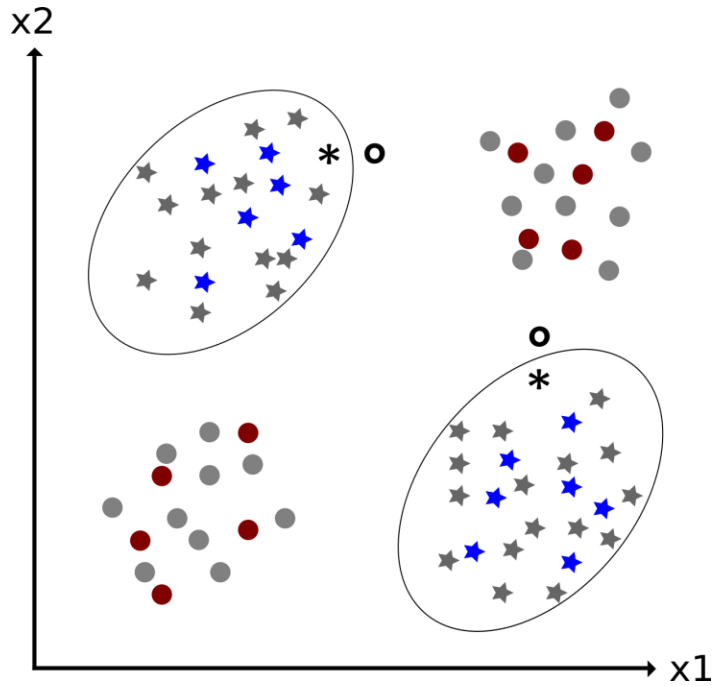
Task

Pro

Out



Source: Produced by the author



Source: Produced by the author

Background – Unsupervised Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

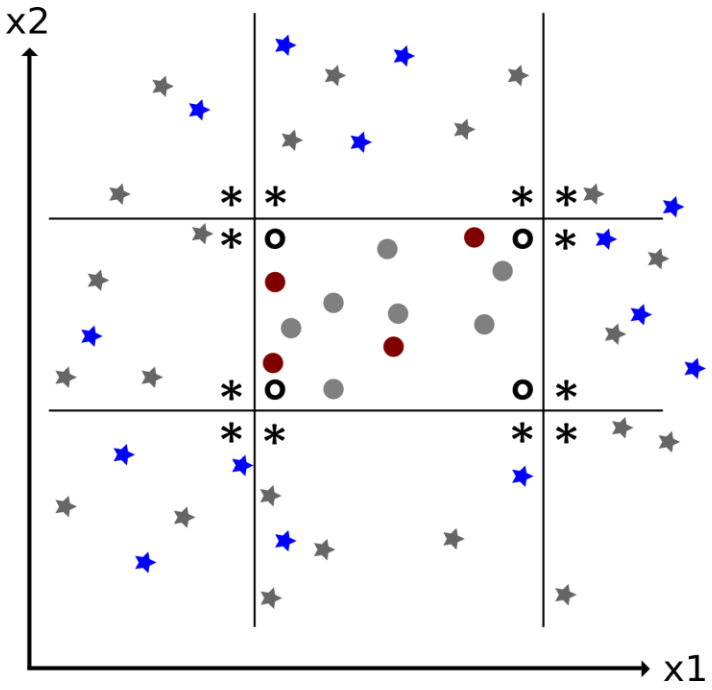
Mot

Back

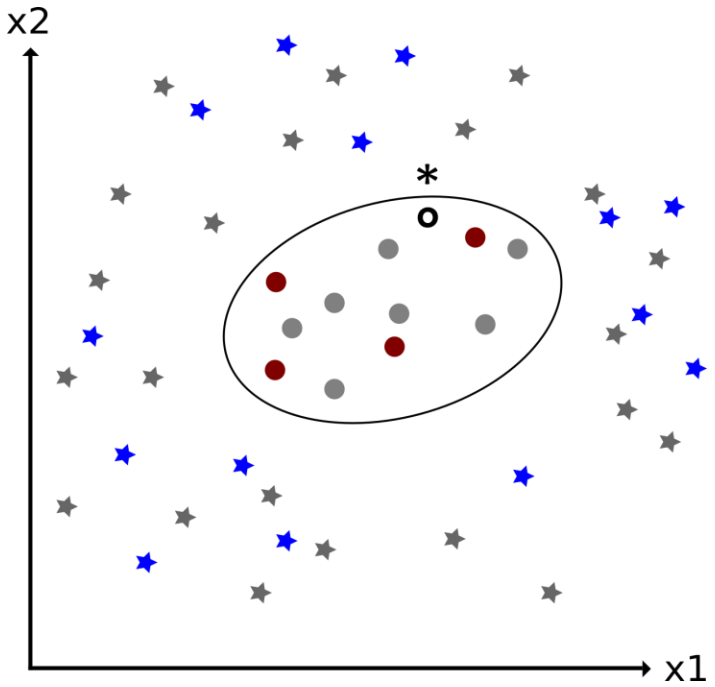
Task

Pro

Out



Source: Produced by the author



Source: Produced by the author

Background – Unsupervised Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

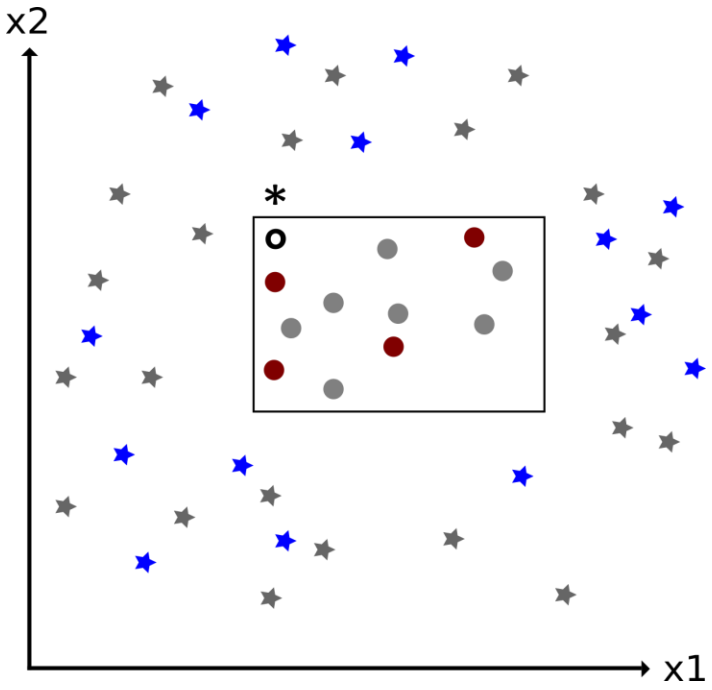
Mot

Back

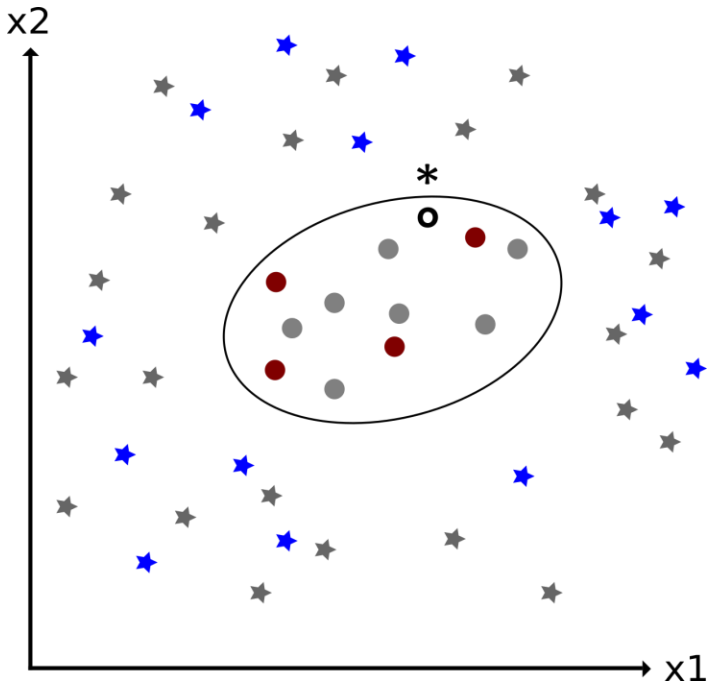
Task

Pro

Out



Source: Produced by the author



Source: Produced by the author

Background – Neural Networks Basics

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

Mot

Back

Task

Pro

Out

Source: "Biology 2e" OpenStax
Sec.35 .2

Background – Neural Networks Basics

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

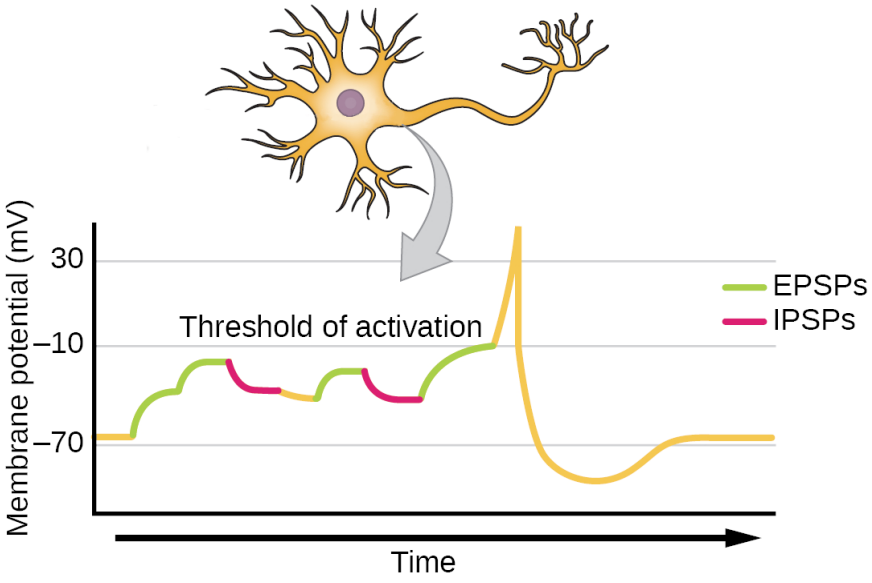
Mot

Back

Task

Pro

Out



Source: "Biology 2e" OpenStax Sec.35 .2

Background – Neural Networks

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

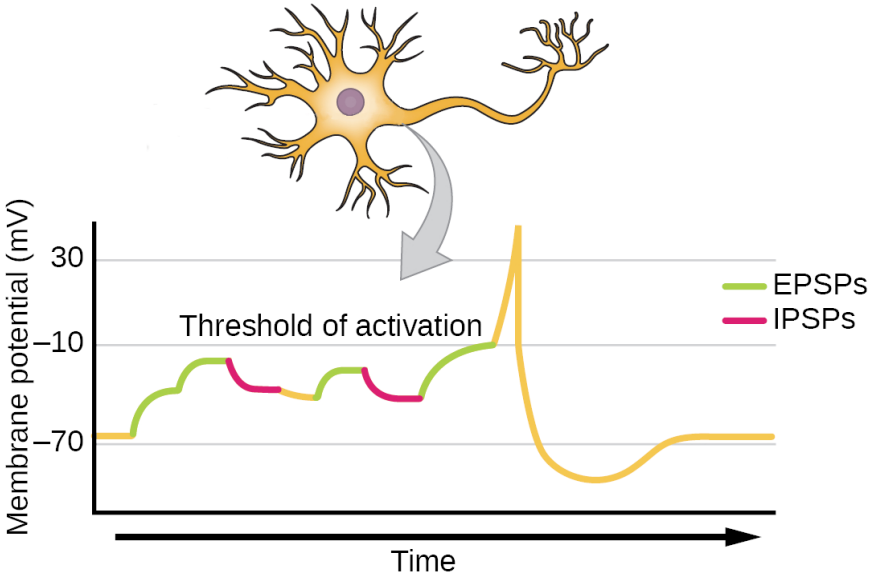
Mot

Back

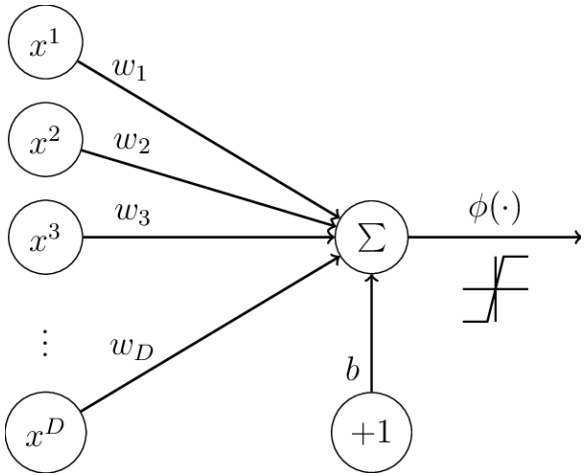
Task

Pro

Out



Source: "Biology 2e" OpenStax
Sec.35 .2



Source: "Neural Netw orks Regularization
Trough Representation Learning"
p. 17

Background – Neural Networks Basics

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

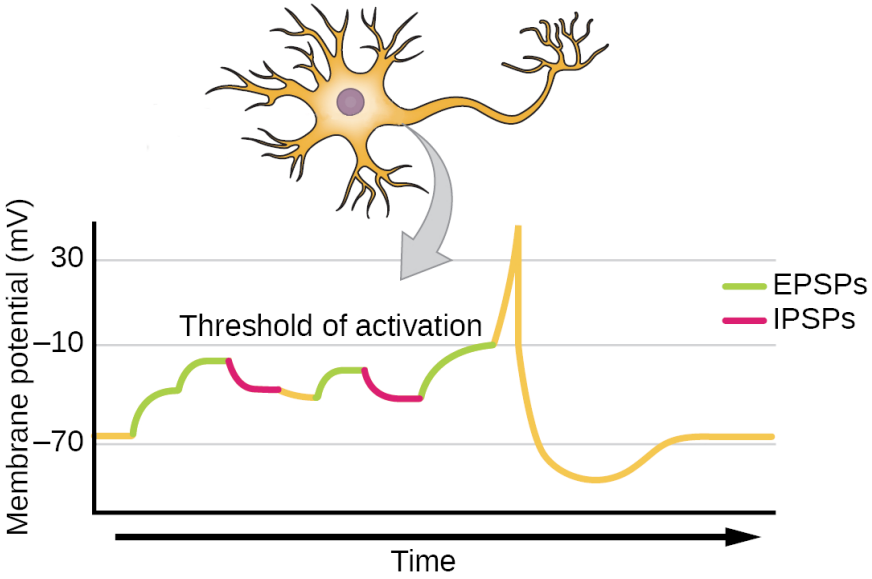
Mot

Back

Task

Pro

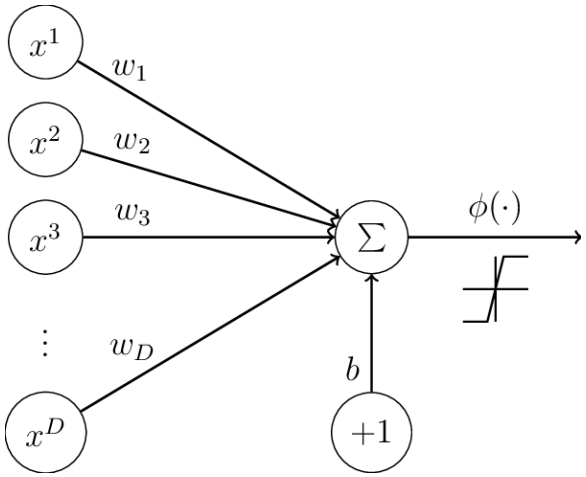
Out



Source:
https://openstax.org/resources/ee34862d28dde2838ca58b4ff73225a93afbee38

"Biology 2e" OpenStax
Sec.35 .2

$$\text{Out} = \phi((\Sigma_i x_i) + b)$$



Source:
https://www.researchgate.net/profile/Soufiane_Belharbi/publication/326439111/figure/fig2/AS:669487372181508@1536629636103/Perceptron-model-Notation-x-i-is-the-i-th-component-of-x-the-same-as-x-i-in

"Neural Networks Regularization Trough Representation Learning"
p. 17

Related Work

TECHNISCHE UNIVERSITÄT DARMSTADT

- Int
- Mot
- Rel
- Task
- Pro
- Out

Related Work – Background

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

Mot

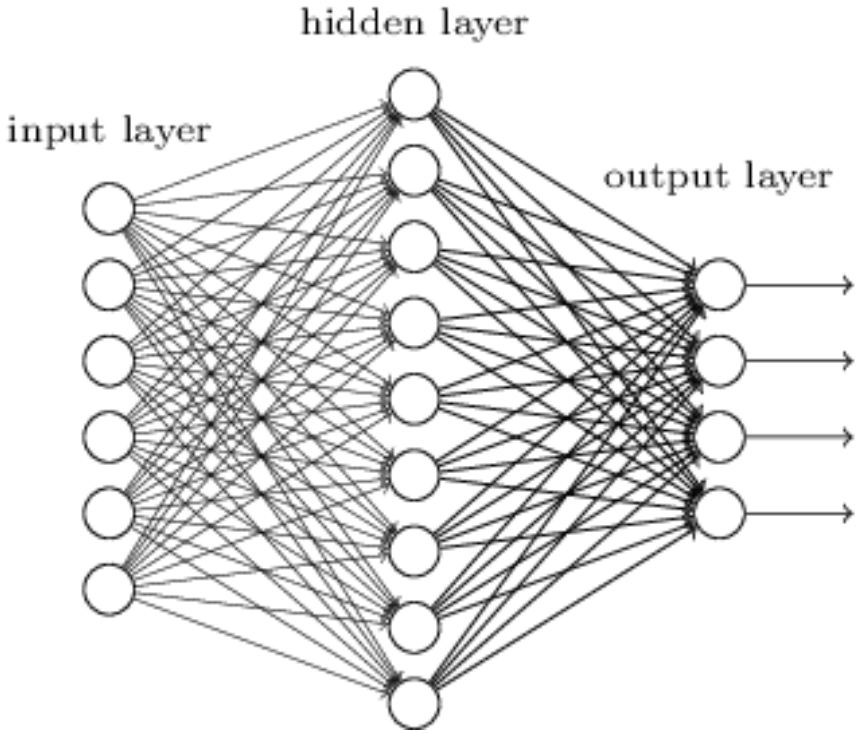
Rel

Task

Pro

Out

Fully Connected Neural Network



Source:
https://hackernoon.com/hn-images/1*Kdnux0Kw1yQ4D8dq__mYCA.png

Related Work – Background

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

Mot

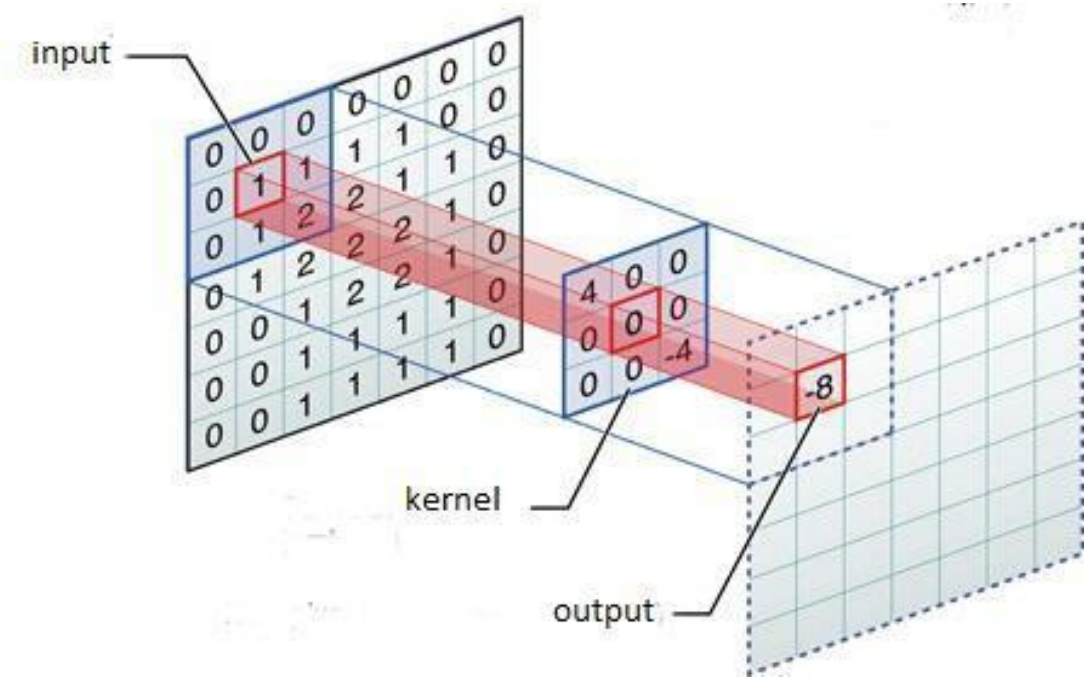
Back

Task

Pro

Out

Convolution in Neural Networks



Source:
https://miro.medium.com/proxy/0*dRD6PhKOnnClhz15.jpg

Related Work – Background

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

Mot

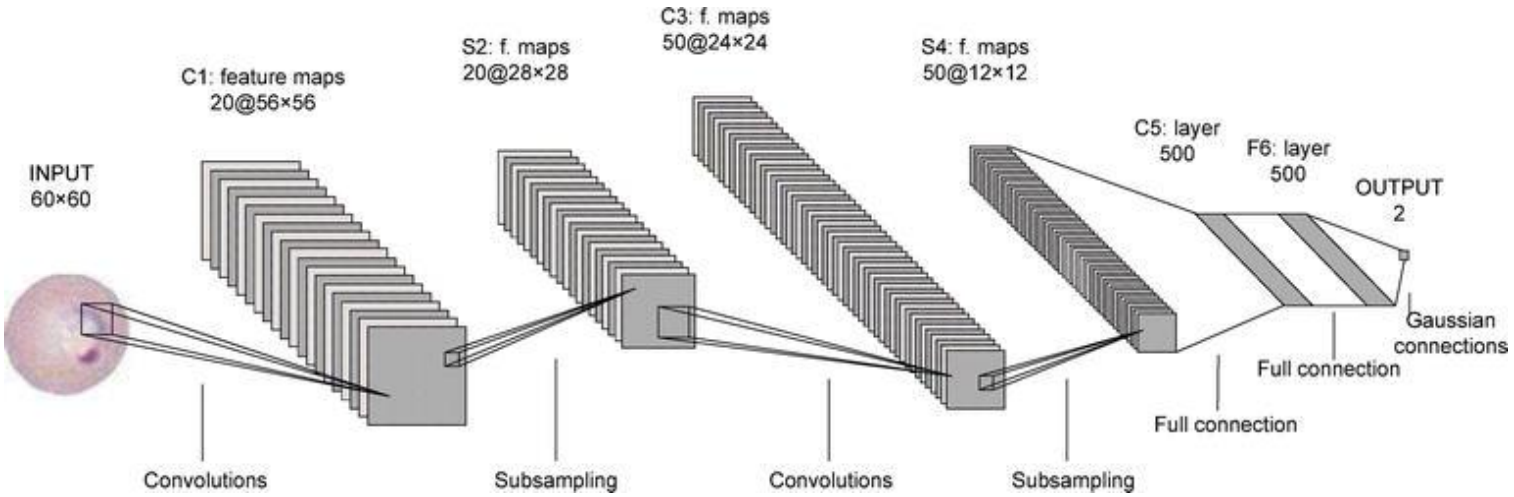
Back

Task

Pro

Out

Convolutional Neural Network Architecture (Lenet-5)



Source:
https://api.intechopen.com/media/chapter/58989/media/F4.png

Related Work – Background

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

Mot

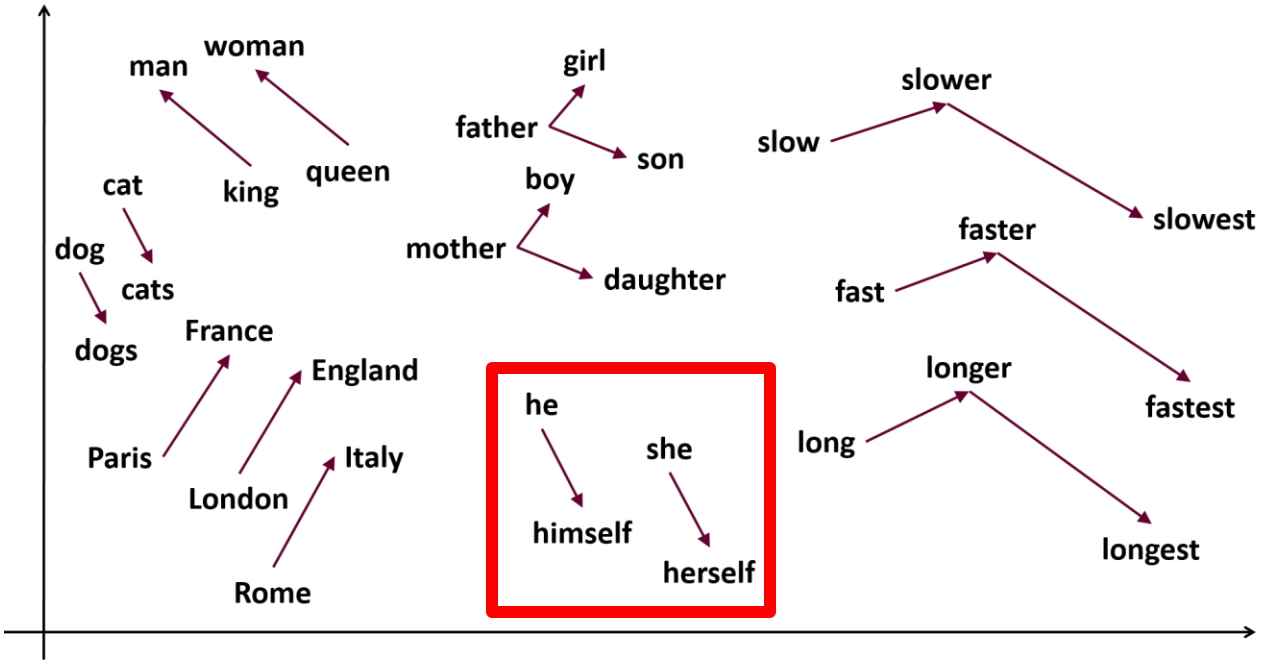
Back

Task

Pro

Out

Language Models



Source:
https://samyzaf.com/ML/nlp/word2vec2.png

Related Work – CNN in NLP

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

Mot

Back

Task

Pro

Out

"Convolutional Neural Networks for Sentence Classification"

- 2014
- Task:
  - Varying Classifications
- Datasets:
  - Movie reviews
  - SST-1, SST-2
  - Subjectivity dataset
  - TREC question dataset
  - Customer reviews
  - MPQA

Related Work – CNN in NLP

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

Mot

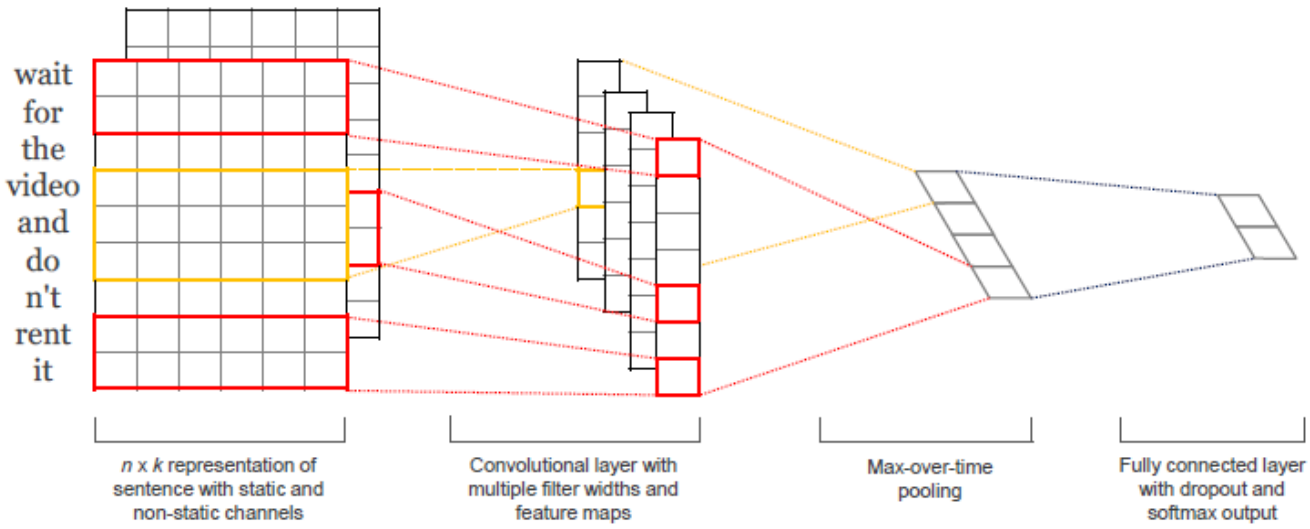
Back

Task

Pro

Out

"Convolutional Neural Networks for Sentence Classification"



Source:
"Convolutional Neural Networks for Sentence Classification"
Figure 1

Related Work - Pruning

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

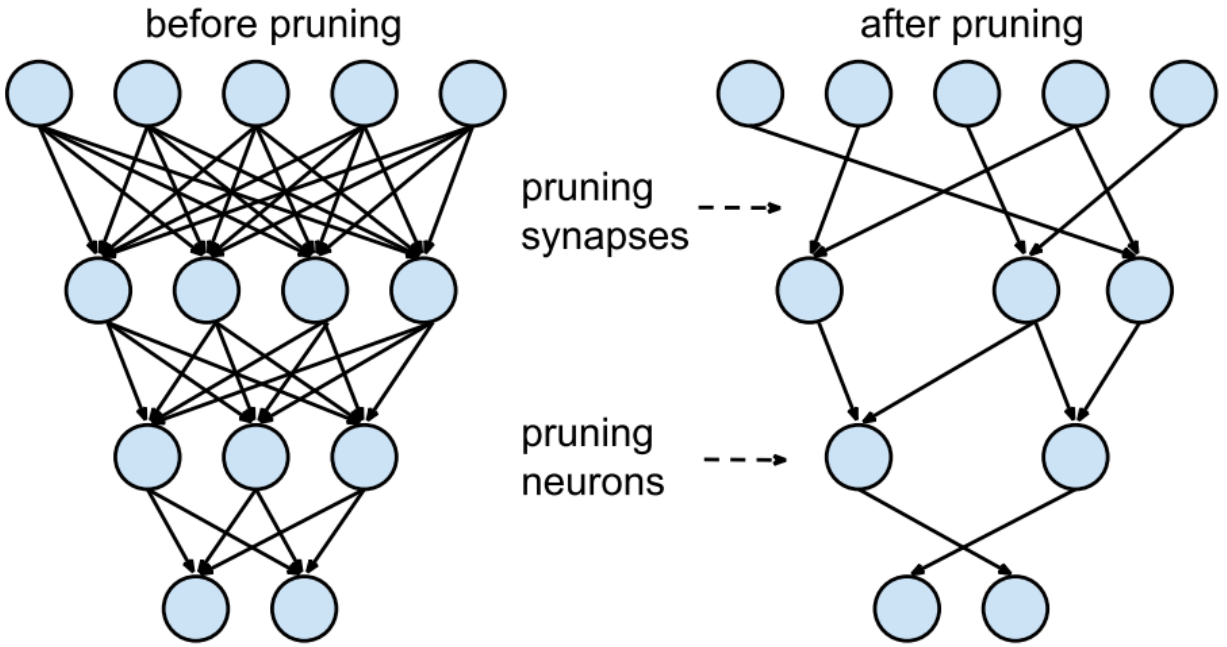
Mot

Rel

Task

Pro

Out



Source:
https://www.mdpi.com/applsci/applsci-09-03169/article_deploy/html/images/applsci-09-03169-g001-550.jpg

"Learning both Weights and Connections for Efficient Neural Networks"
Figure.3

Related Work – Pruning

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

Mot

Rel

Task

Pro

Out

"Learning both Weights and Connections for efficient Neural Networks"

- 2015
- Task:
  - Image Classification (ImageNet)
- Architectures:
  - LeNet (300-100-FC, 5-CNN)
  - AlexNet
  - VGG-16
- Compression:
  - 9x to 13x

Related Work – Pruning

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

Mot

Rel

Task

Pro

Out

"ThiNet: A Filter Level Pruning Method for Deep Neural Network Compression"

- 2017
- Task:
  - Image Classification (ImageNet)
- Architectures:
  - VGG-16
  - ResNet-50
- Compression:
  - Up to ~17x

Related Work – Pruning

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

Mot

Rel

Task

Pro

Out

"The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks"

- 2019
- Task:
  - Image Classification (MNIST)
- Architectures:
  - Lenet (300-100-FC, Conv-2, Conv-4, Conv-6)
  - VGG-19
  - ResNet-18
- Compression:
  - ~20x to ~50x

Related Work – Early Pruning

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

Mot

Rel

Task

Pro

Out

"Really should we pruning after model be totally trained? Pruning based on a small amount of training"

- 2019
- Task:
  - Image Classification (MNIST, CIFAR-10)
- Architectures:
  - Unspecified CNN
  - VGG-19
- Compression --- Training Speed-Up:
  - ~10x --- 10x

Related Work – Network Architecture Search

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

Mot

Rel

Task

Pro

Out

"Rethinking the Value of Network Pruning"

- 2018
- Observation:
  - Randomizing weights does not worsen a pruned network
- Weights are not essential to the quality of pruned network
- Pruning at its core is about finding suitable network architectures

Related Work – Network Architecture Search

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

Mot

Rel

Task

Pro

Out

"Network Architecture Search: A Survey"

- 2019
- […]

Task I

TECHNISCHE UNIVERSITÄT DARMSTADT

Reproduction

- No source-code available
  - Produce own source-code for the experiments in the paper
- Verify source-code by running experiments

Task II

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

Mot

Back

Task

Pro

Out

Transfer to NLP

- Original context for the paper
  - Task: Image Classification
  - Dataset: "MNIST"
  - Model: Varying FCNN and CNN
- Find comparable context in NLP
  - Task: Topic Classification
  - Dataset: "Reuters-21578"
  - Model: TBD
- Check if the Lottery-Ticket-Hypothesis holds

Task III

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

Mot

Back

Task

Pro

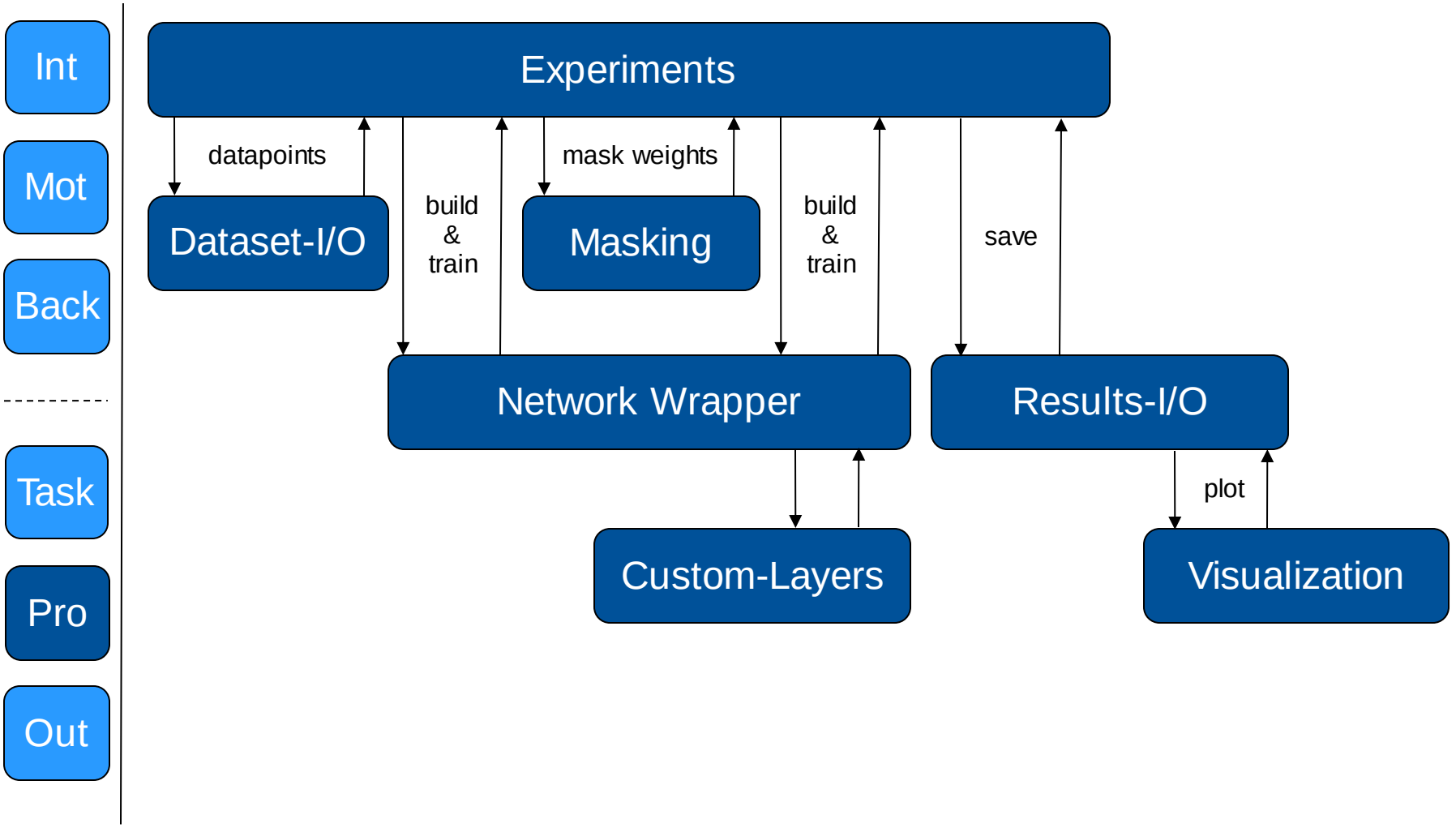
Out

Early Retrieval of Lottery Tickets

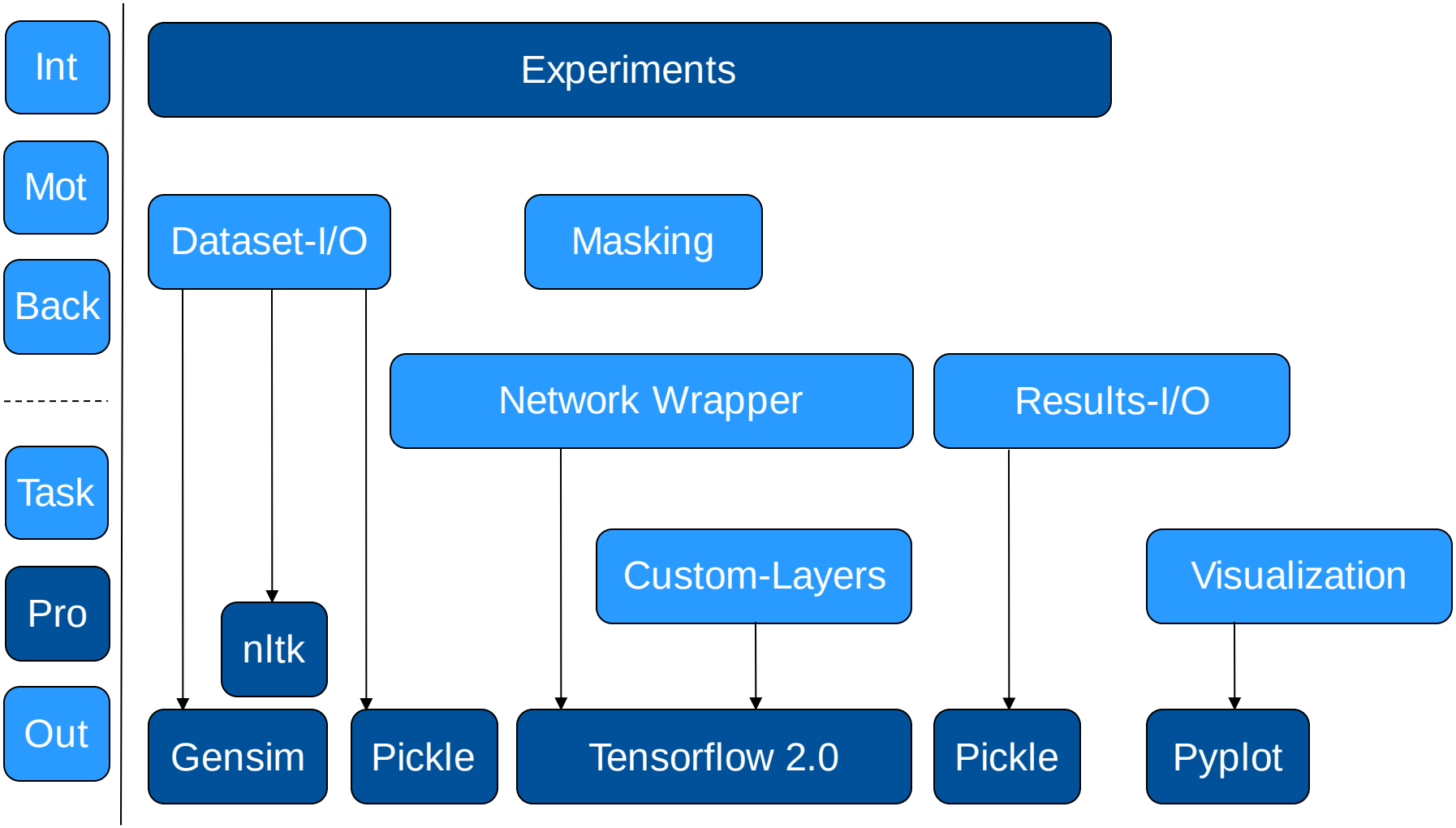
- Original method
  - "Select" all weights of the "fully trained" network over a certain percentile
  - Reset weights to original intial value
  - Retrain network
  - Repeat (Optional)
- Adaptation
  - "Select" weights earlier / develop early stopping criteria
  - "Select" weights based on other metrics (Optional)

Progress – Python-project

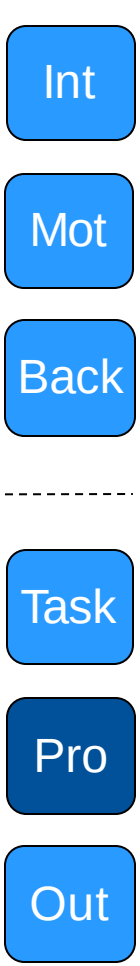
TECHNISCHE UNIVERSITÄT DARMSTADT



Progress – Backend



Progress – Experiment(s)



Remaining Work

TECHNISCHE UNIVERSITÄT DARMSTADT

Int

Mot

Back

Task

Pro

Out

Remaining parts of the framework

- Custom Convolutional Layer
- Support for iterative Pruning

More experiments

- MNIST / Lenet-CNN-6
- MNIST / VGG-18
- Reuters / TBD
- MNIST / Lenet-FCN / Early Pruning

Thank you for your attention! Questions?

TECHNISCHE UNIVERSITÄT DARMSTADT