RECHNER-ARCHITEKTUR
INFIII

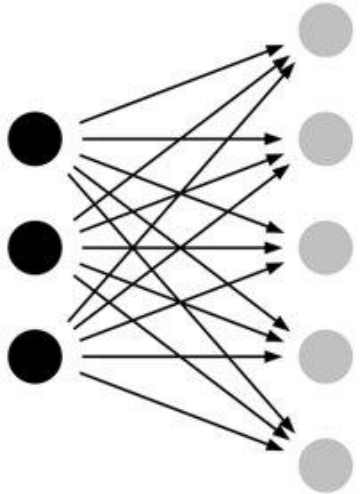
FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

2. Problem Characteristics of NNs

Philipp Holzinger

Content

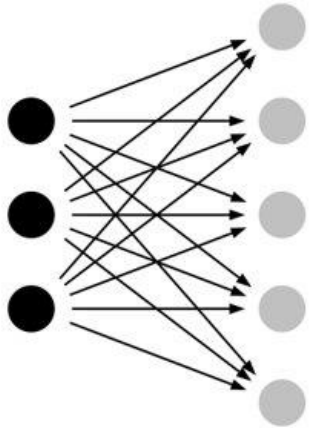
1. Algorithmic Description
2. Possible Optimizations



Pseudocode Description

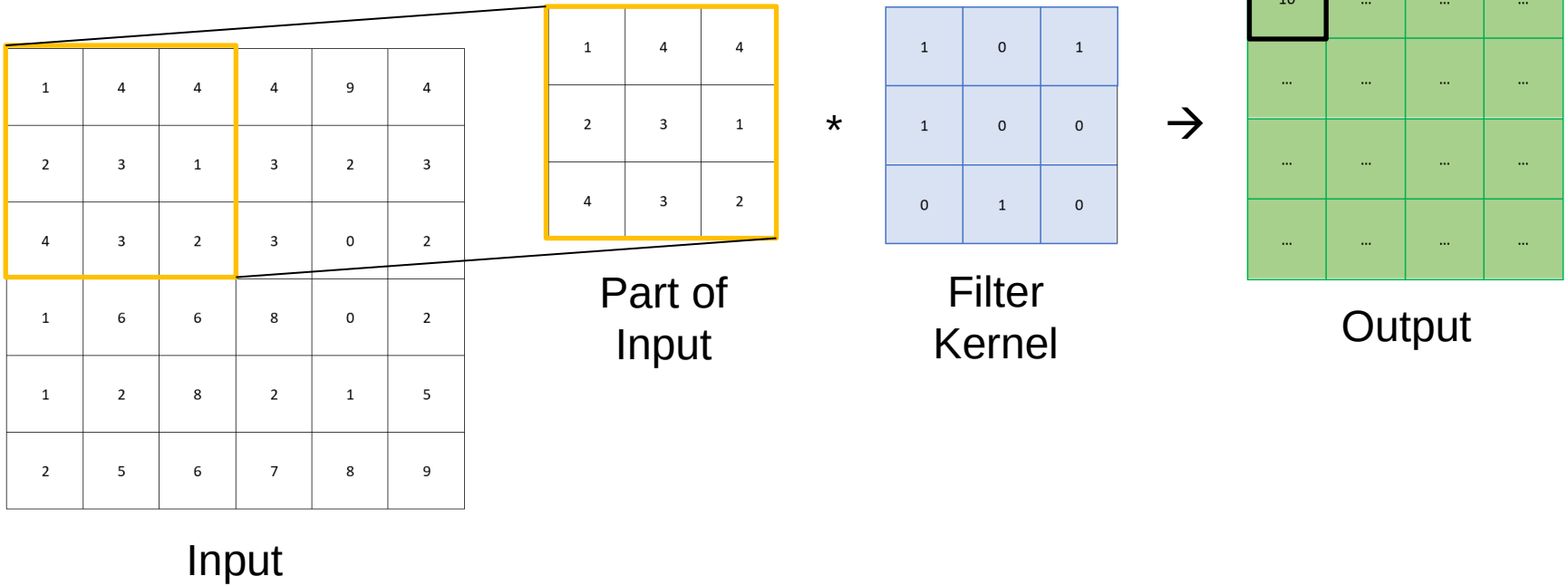
Fully connected layer

```
for (b = 0 to B-1)          // batch
  for (n = 0 to N-1)        // neuron
    for (i = 0 to I-1)      // input
```



Working with Filters

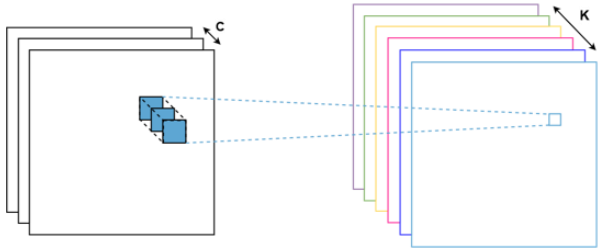
2D Example



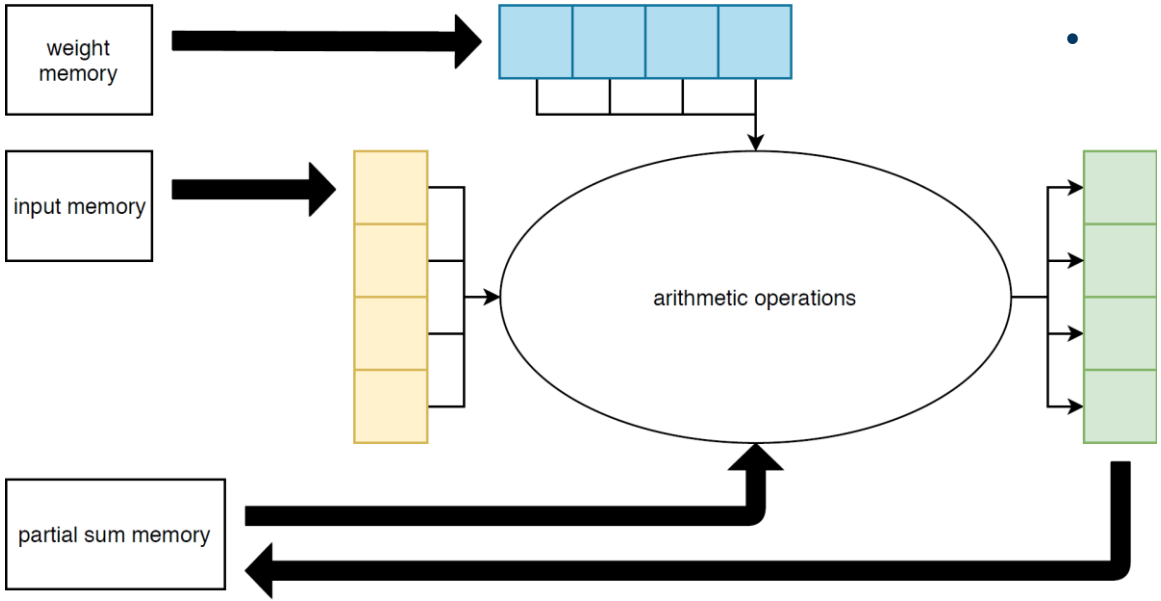
Pseudocode Description

Convolutional layer

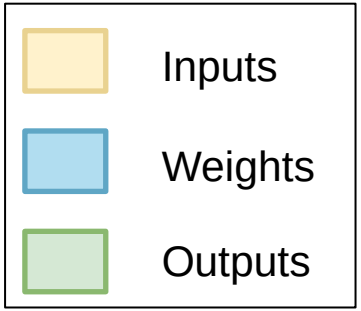
```
for (b = 0 to B-1)                  // batch
  for (k = 0 to K-1)                // output channels
    for (c = 0 to C-1)              // input channels
      for (x = 0 to X-1)            // input columns
        for (y = 0 to Y-1)          // input rows
          for (fx = 0 to FX-1)      // filter columns
            for (fy = 0 to FY-1)    // filter rows
```



Processing



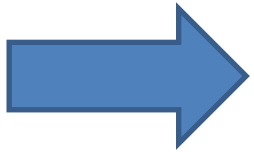
- What data is used in **one cycle**?
- What data is used in **consecutive cycles**?



Things can be Optimized

Fully connected layer

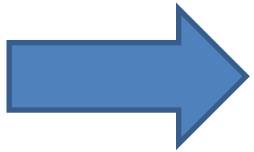
```
for (b = 0 to B-1)          // batch
  for (n = 0 to N-1)        // neuron
    for (i = 0 to I-1)      // input
```



Parallelism

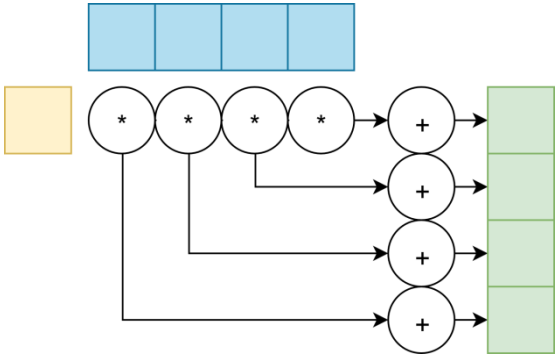
Convolutional layer

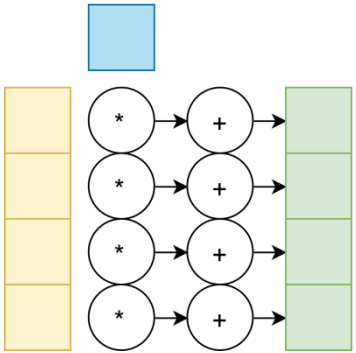
```
for (b = 0 to B-1)                  // batch
  for (k = 0 to K-1)                // output channels
    for (c = 0 to C-1)              // input channels
      for (x = 0 to X-1)            // input columns
        for (y = 0 to Y-1)          // input rows
          for (fx = 0 to FX-1)      // filter columns
            for (fy = 0 to FY-1)    // filter rows
```

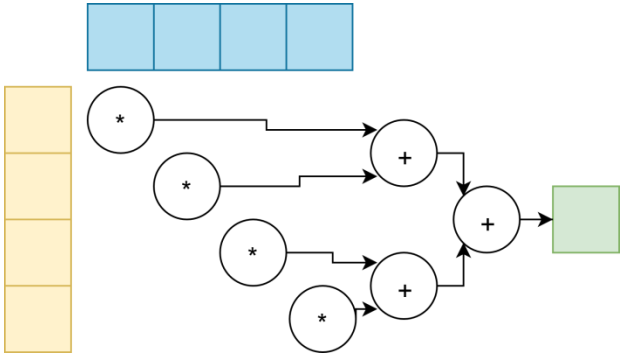


Data stationarity

Data Parallelism



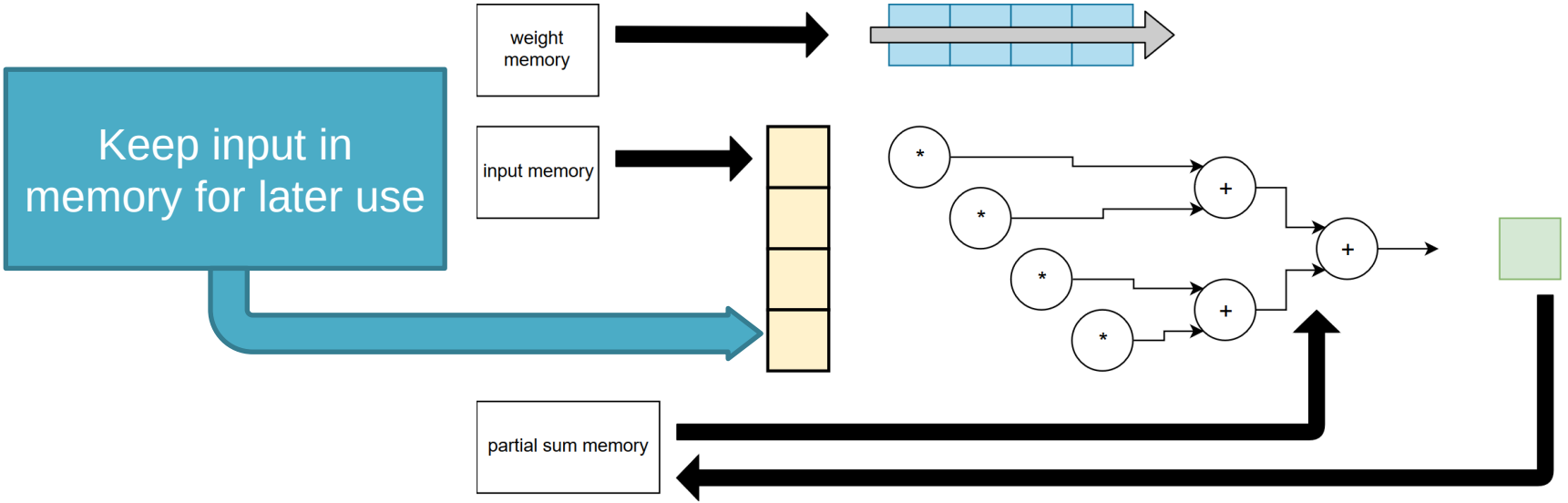




Input Parallel	Weight Parallel	Output Parallel
Reduce input bandwidth	Reduce weight bandwidth	Reduce output bandwidth

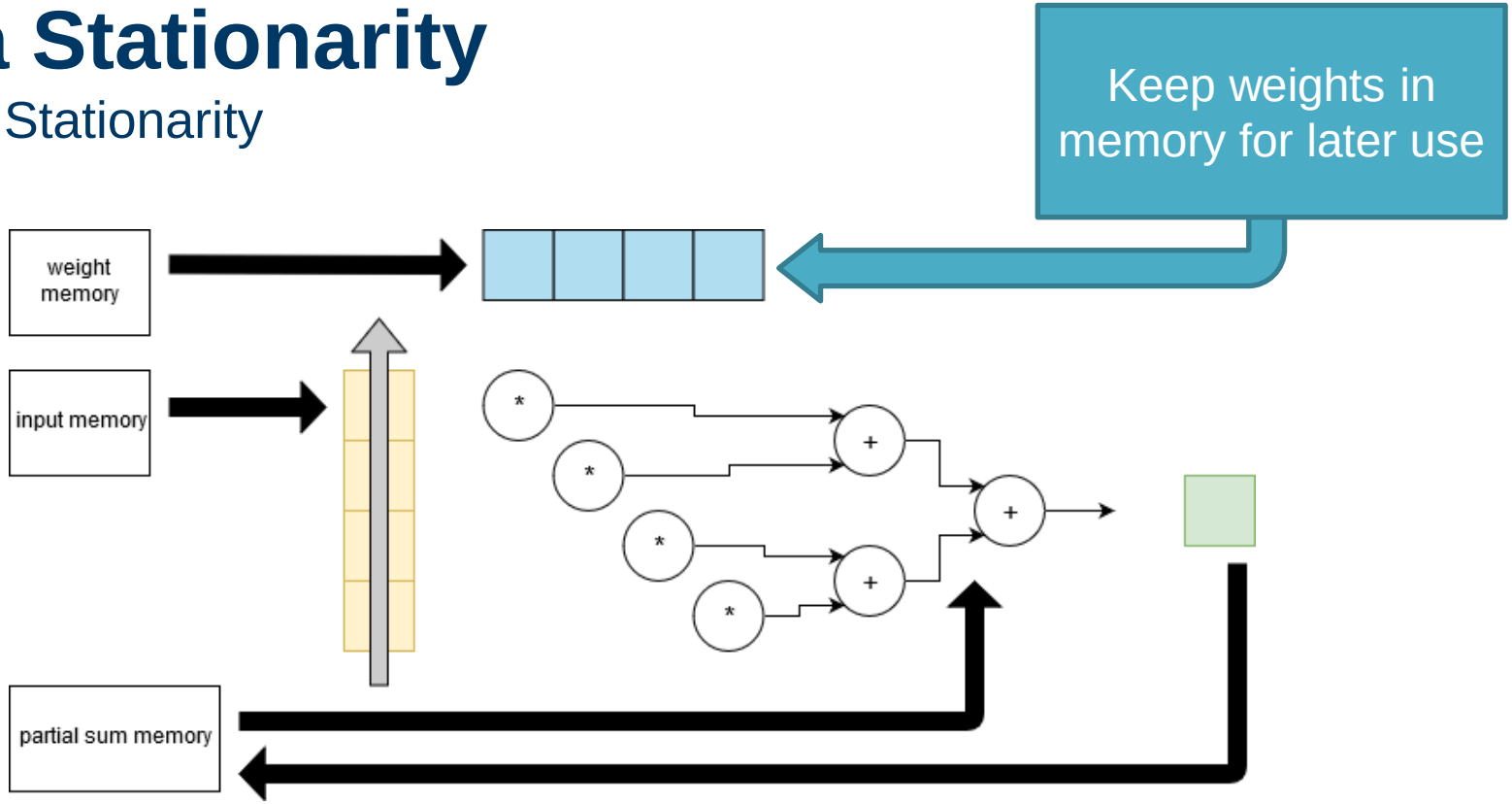
Data Stationarity

Input Stationarity



Data Stationarity

Weight Stationarity



Data Stationarity

Output Stationarity