# Architektur Basis-Modell

## Generell

* Modelle immer mit Binärer Klassifikation ausprobiert und getestet:
  + 10'000 Bilder pro Klasse
  + Bird vs Dog
  + Zweites Datenset vorhanden für Transfer-Learning und ausprobieren wie es mit 4 Klassen funktioniert
    - Fish, Lizard, Monkey, Snake
* Basis-Architektur:
  + Input: Grösse von 224 (gewählt weil guter Mix von Bildqualität aber Trainingsgeschwindigkeit)
  + Dann Normalisierung zu 0-1
  + Dann 4x Blöcke
    - Grösser-werdende Filter-Anzahl
    - Convolution
    - Batch-Normalisierung
    - Pooling
    - Dropout
  + GlobalMaxPooling
  + Flatten
  + 2x Dense-Layer mit 256 und 512 Neuronen
  + Normalisierung
  + Dropout
  + Dense-layer mit N-Anzahl-Klassen Neuronen
* **Architekturen welche signifikant grösser waren auf unserer Hardware nicht trainierbar. (unmöglich)**
* Verwenden direkt die logits für die Berechnung der Metriken. Sonst hatten wir Fehler bei der Berechnung mit Pytorch
* Filter-Anzahl von über 128 (basis Wert) war Training zu langsam. Deshalb wurde diese Grösse gar nicht erst miteinbezogen.
* Batch-Grösse von 64 hat bei unserem System das schnellste Training ermöglicht.
* Augmentation direkt im Modell ist ebenfalls vorteilhaft, da dann jede Epoche neuen Augments verwendet werden.
* RandomRotation aus der Augmentation entfernt, da damit die Resultate immer schlechter waren.
* Early Stopping wird eingesetzt, um ein Overfitting zu verhindern. Zusätzlich wird ein learning-rate scheduler verwendet, der die lr heruntersetzt, wenn die accuracy nicht besser wird. Der lr-scheduler hat eine tiefere patience als das early-stopping. Dies bedeutet, dass wenn sich die accuracy nicht ändert zuerst die learning-rate anpasst und wenn dann immer noch nichts passiert, wird gestoppt.

## Tuning

Wir haben nicht alles gleichzeitig getuned sondern einzelne Teile der Architektur. Damit sind wir schrittweise vorgegangen und haben so für einzelne Teile der Architektur die besten Parameter auswählen können. Anschliessend haben wir von den verschiedenen Durchläufen die besten Werte weiterverwendeten. Die Reihenfolge ist dieselbe wie im Dokument hier. Als Basis wurde die oben beschriebene Architektur genommen. Diese wurde dann durch das Tuning weiter angepasst. Für das Tuning haben wir immer die Validation-Accuracy verwendet, da dies der wichtigste Wert für ein Modell ohne Overfitting ist. Es wurde jeweils pro Durchlauf für 4 Epochen trainiert. Aus Zeitgründen wurden nicht mehr verwendet, da sonst das Tuning zu lange gehen würde.

## Tuning von Basis-Filter, Learning-Rate und Pool-Type

Hier wird die Learning-Rate und zwei CNN-Spezifische Entscheidungen getroffen. Die Anzahl Filter in jedem Block und der Typ von Pool (max oder avg). Bei den anderen Sachen (wie zum Beispiel stride, pool-grösse, etc…) waren wir uns sicher, dass diese gut gewählt sind. Basis-Filter ist der tiefste Wert der 4 Convolution-Blöcke. Dieser wird dann \*1, \*2, \*4, \*8 gerechnet und das sind die Anzahl-Filter in jedem Block.

* Beste Werte:
  + Basis-Filter 8, pool\_type max und learning-rate 0.001. Mit Tuning
* Kleinste Modellgrösse also beste Resultate 🡪 wahrscheinlich am wenigsten Overfitting

Trial 05 summary

Hyperparameters:

filter\_size: 8

pool\_type: max

learning\_rate: 0.001

Score: 0.8165000081062317

Trial 12 summary

Hyperparameters:

filter\_size: 64

pool\_type: max

learning\_rate: 0.001

Score: 0.8159999847412109

Trial 15 summary

Hyperparameters:

filter\_size: 64

pool\_type: max

learning\_rate: 0.0001

Score: 0.812749981880188

Trial 13 summary

Hyperparameters:

filter\_size: 32

pool\_type: max

learning\_rate: 0.001

Score: 0.8004999756813049

Trial 19 summary

Hyperparameters:

filter\_size: 32

pool\_type: avg

learning\_rate: 0.0001

Score: 0.7637500166893005

## Tuning von Output-Dense-Layern

Mit diesem Tuning wurde geschaut was die beste Kombination für die Output Layer ist. Dabei wurde einerseits untersucht wie viele Dense-Schichten (1-3) und wie viele Neuronen pro Schicht (64, 128 oder 256).

Dabei wurde auch hier wieder beispielsweise in der ersten Schicht 64\*1, 64\*2, 64\*4 gerechnet.

* Beste Werte: 3 Schichten mit 64 als Basis-Wert
* Also auch hier wieder sind kleinere Werte besser
* Resultiert in den folgenden Dense-Schichten:
  + Dense 64
  + Dense 128
  + Dense 256

Trial 03 summary

Hyperparameters:

dense\_layer\_amount: 3

dense\_layer\_base: 64

Score: 0.843999981880188

Trial 01 summary

Hyperparameters:

dense\_layer\_amount: 2

dense\_layer\_base: 128

Score: 0.8370000123977661

Trial 07 summary

Hyperparameters:

dense\_layer\_amount: 1

dense\_layer\_base: 128

Score: 0.8332499861717224

Trial 08 summary

Hyperparameters:

dense\_layer\_amount: 2

dense\_layer\_base: 64

Score: 0.8322499990463257

Trial 04 summary

Hyperparameters:

dense\_layer\_amount: 3

dense\_layer\_base: 256

Score: 0.8299999833106995

## Tuning von Regularisierung (Batch-Normalisierung und Dropout)

Damit wollten wir herausfinden was am besten hilft gegen Overfitting. Ob man in den einzelnen Blöcken schon Regularisierung soll und/oder erst nachher.

* Beste Werte konnten ohne jegliches DropOut und mit einer Batch-Normalisierung in den Schichten und in der Output-Schicht erreicht werden.

Trial 04 summary

Hyperparameters:

inter\_dropout: False

inter\_batch\_norm: True

output\_dropout: False

output\_batch\_norm: True

Score: 0.827750027179718

Trial 00 summary

Hyperparameters:

inter\_dropout: True

inter\_batch\_norm: True

output\_dropout: False

output\_batch\_norm: True

Score: 0.8180000185966492

Trial 05 summary

Hyperparameters:

inter\_dropout: False

inter\_batch\_norm: False

output\_dropout: True

output\_batch\_norm: False

Score: 0.815500020980835

Trial 13 summary

Hyperparameters:

inter\_dropout: False

inter\_batch\_norm: True

output\_dropout: True

output\_batch\_norm: False

Score: 0.8142499923706055

Trial 06 summary

Hyperparameters:

inter\_dropout: True

inter\_batch\_norm: True

output\_dropout: True

output\_batch\_norm: False

Score: 0.812250018119812

## Training Basis-Modell

Insgesamt 13 Epochen dann hat der Callback das Training gestoppt. Die Weights wurden von Epoche 7 übernommen, da dort der val\_accuracy am besten war. Training hat viele Spikes, also wahrscheinlich Overfitting.

* **Parameter: 71793**

A screenshot of a graph

Description automatically generated

# Architektur Grösseres Residual Modell

* Ohne Tuning Parameterauswahl. Sondern von Tuning-Resultate des Basis-Modells inspiereren lassen, dann durch trial und error.
* «Residual Connections». Mit diesen erhofft man sich in früheren Schichten erkannte Features wieder aufgreifen zu können.
* **358’353 Parameter**

Kontrast / Brightness Augmentation

* Rescaling auf 0-1
* Eine Input-Convolution + Batch-Normalisierung für anfangs-residual
* Dann Vier Convolution Schichten (16, 32, 64, 128)
  + Pro Schicht 2x Convolution
  + 2x Batch-Nprm
  + Max-Pooling
  + Residual berechnen und addieren
* Dann Klassifikation
  + GlobalMax
  + 2x Dense (128, 256)
  + BatchNorm
  + Dropout
  + Dense (Anzahl Klassen)

Resultate Training:

* 16 Epochen trainiert
* Weights von Epoche 10 wiederhergestellt
* Allgemein gut. Besser aber nicht viel besser als Basis-Modell aber hat viel mehr Parameter.
* Callback für early stopping und beste weights auswählen war auch hier wieder gute Entscheidung
* Training etwas «ruhiger»

A screenshot of a computer screen

Description automatically generated

# Transfer-Learning

* Basis-Modell wurde geladen
* Dann wurde der letzte Layer entfernt und durch ein Dense mit 4 Units ersetzt
* Dann wurden die Weights aller Layer ausser dem neu dazugefügten Dense gefreezed (layer.trainable = false)
* Dann wurde erneut ein Training mit fast gleicher Herangehensweise durchgeführt wie auch schon beim Basis-Modell
* Resultate schlecht. Stoppt bei 50% auf besser zu werden
  + Entweder Fehler von uns oder Modell einfach zu klein für transfer
* Insgesamt 17 Epochen trainiert. Aber beste Weights von Epoche 9 wurden gespeichert

A screenshot of a computer

Description automatically generated