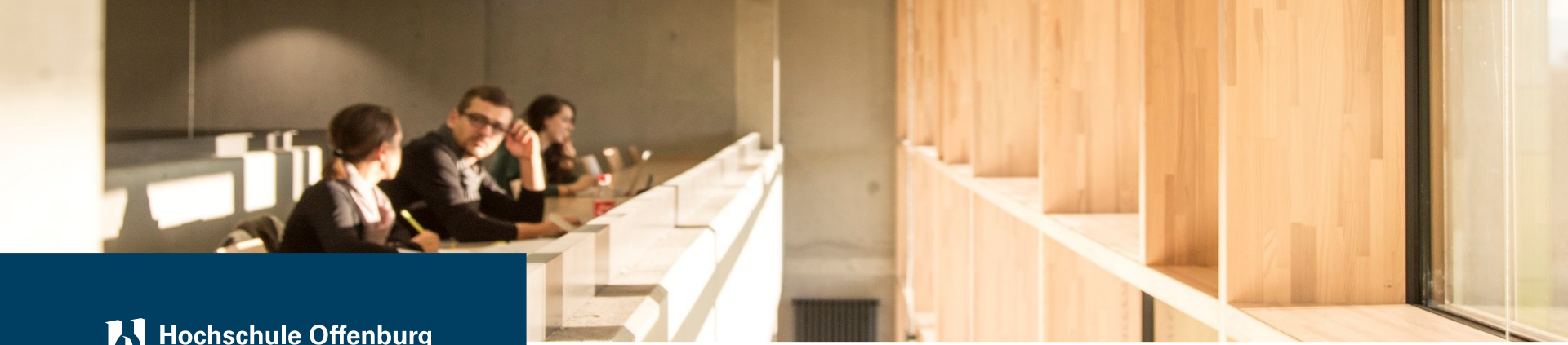
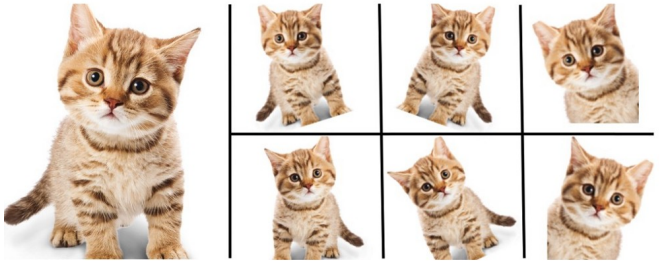


Hochschule Offenburg
offenburg.university

# Data Augmentation



Enlarge your Dataset

Seminar Deep Learning

Jonas Bräuer

08.01.2020

EMI Elektrotechnik, Medizintechnik und Informatik

# Agenda



- Theory
  - Definition
  - Different techniques
  - Which augmentation?
- Test with pytorch (practical part)
  - Comparison of different augmentations
- Conclusion

# Definition



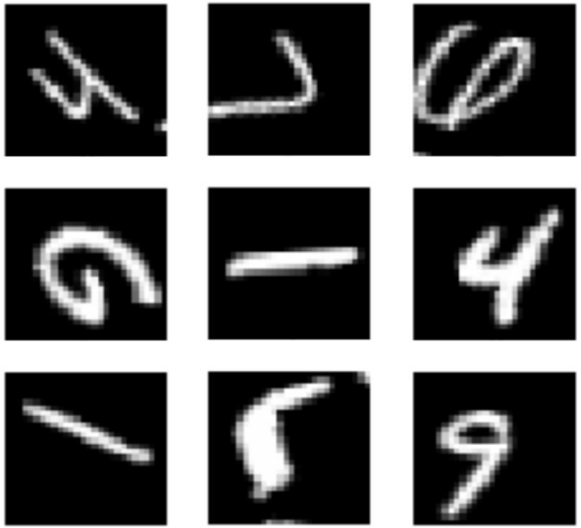
- „Approaches overfitting from the root of the problem, the dataset“
- Preserves the label
- Adds big data characteristics
- „Imagination or dreaming“ of data [2]
- Online or Offline augmentation

# Why Data Augmentation?



- Bigger datasets result in better Deep Learning models [2]
- Class-balancing oversampling (SMOTe)

  Synthetic Minority Over-sampling Technique
- Test time augmentation
- But be careful...




[2]

# Which augmentation to choose?

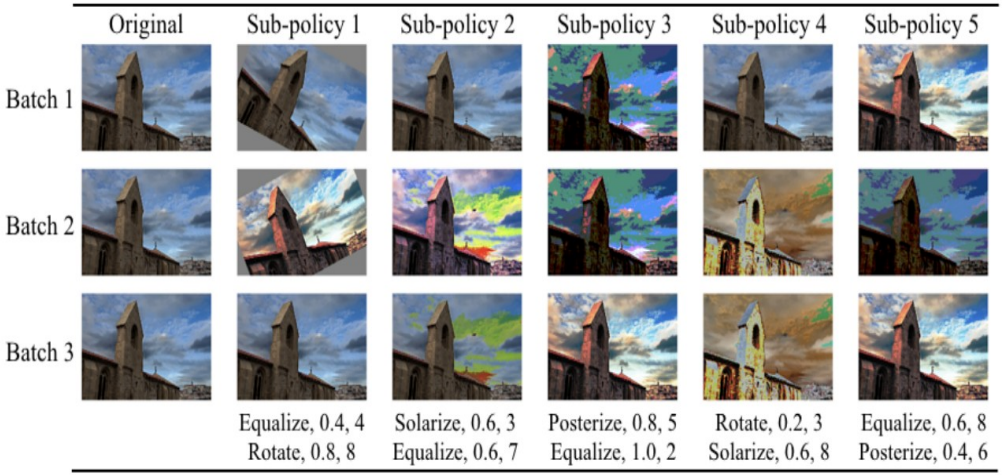
Hochschule Offenburg
offenburg.university

- No certain “best way”, context dependent, but...
  - think about what makes sense in context
  - try geometric/color based augmentation (simple and efficient)
  - try GAN augmentation
  - try combinations
  - consider test time augmentation
  - experiment and check latest publications
  - use Meta-Learning

# AutoAugment



- searches for optimal geometric transformations
  - ‘translateY 17 pixels’
- Improvements:
  - Fast Auto Augment [2]
  - PBA (Population Based Augmentation) [4]




[6]

# NVIDIA Dali



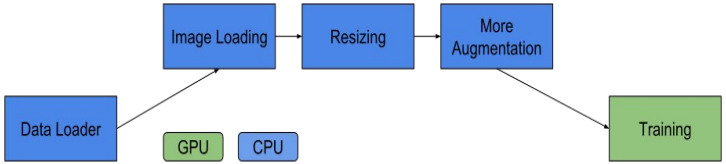
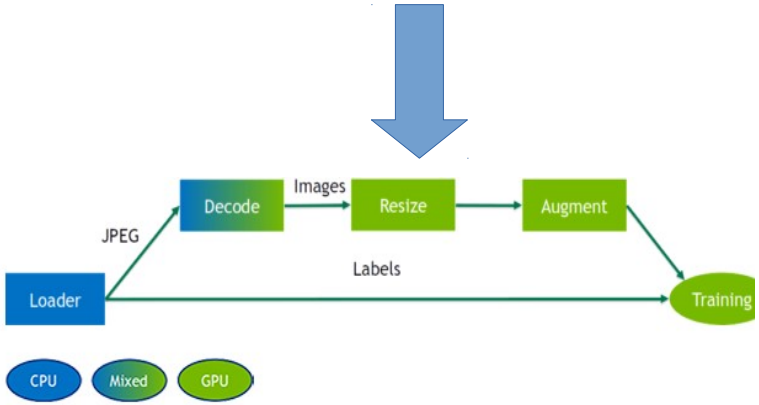
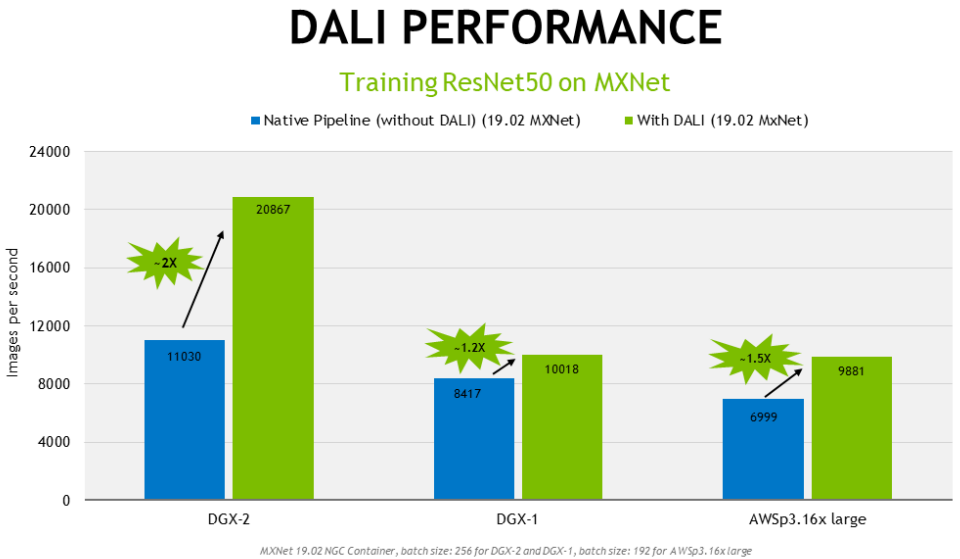


Image Loading
Resizing
More Augmentation
Data Loader
GPU
CPU
Training


Decode
Images
Resize
Augment
JPEG
Labels
Loader
Training
CPU
Mixed
GPU


DALI PERFORMANCE
Training ResNet50 on MXNet
Native Pipeline (without DALI) (19.02 MXNet)
With DALI (19.02 MxNet)
Images per second
24000
20000
16000
12000
8000
4000
0
11030
20867
~2X
8417
10018
~1.2X
6999
9881
~1.5X
DGX-2
DGX-1
AWSp3.16x large
MXNet 19.02 NGC Container, batch size: 256 for DGX-2 and DGX-1, batch size: 192 for AWSp3.16x large

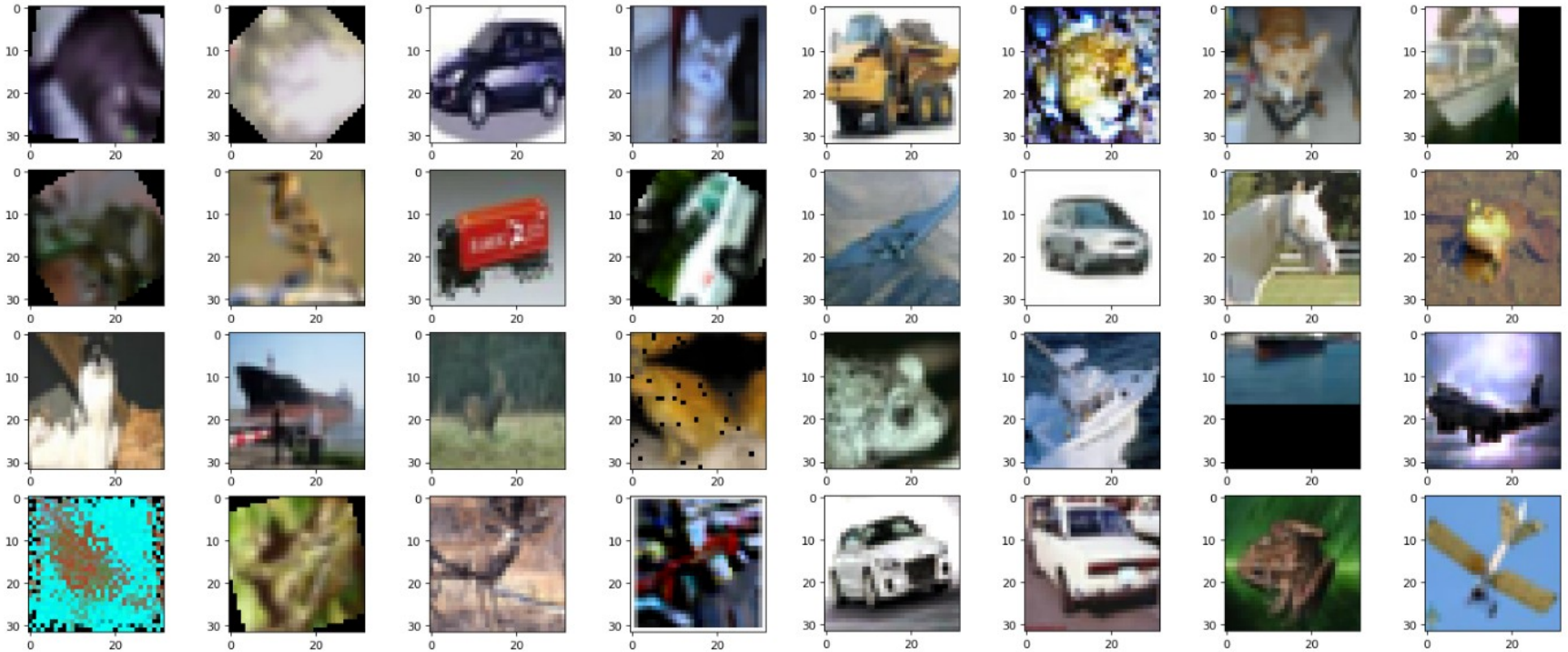
# Data augmentation tests with pytorch



- norm, imgaug, pytorch, autoaugment
- SmallNet, 3 conv. 2 fully connected layer
- Cifar10, 10 classes
- Online and offline augmentation

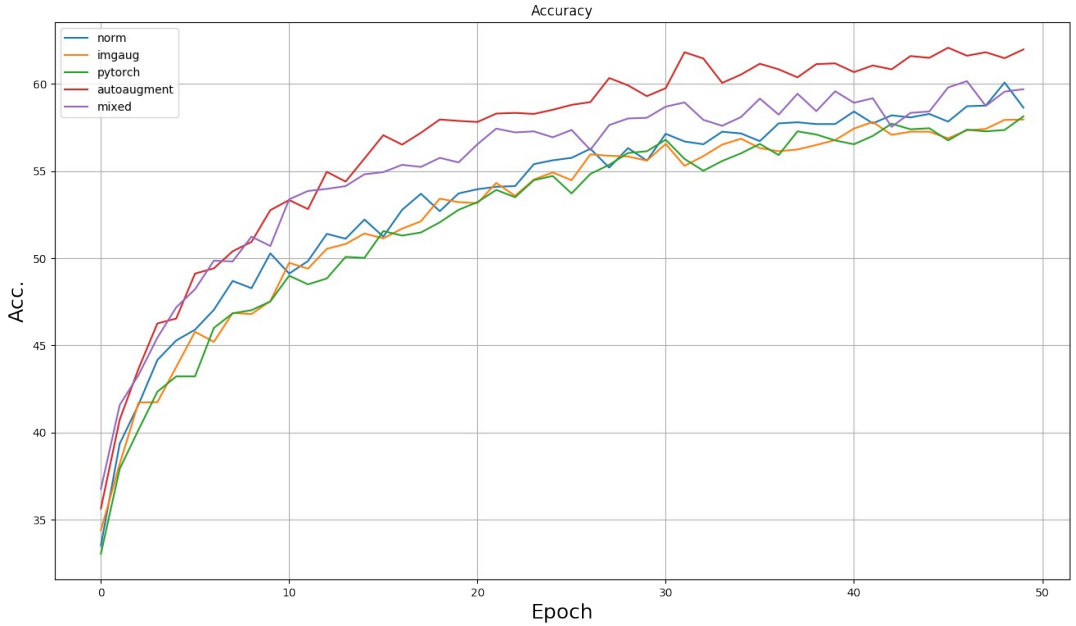
# Results

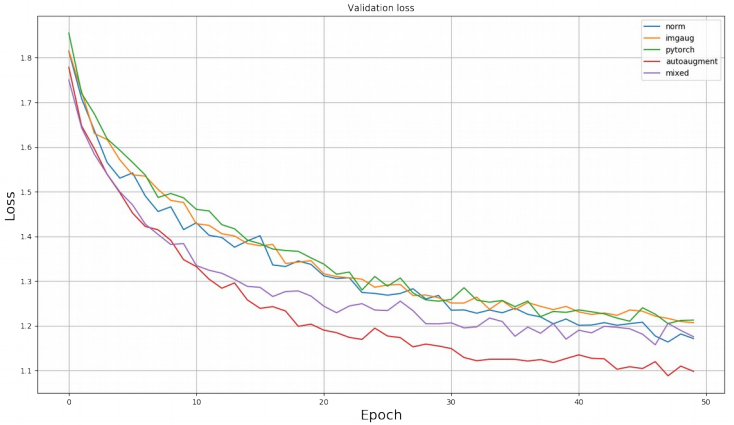




# Results




Accuracy
norm
imgaug
pytorch
autoaugment
mixed
Acc.
Epoch
35
40
45
50
55
60
0
10
20
30
40
50


Validation loss
norm
imgaug
pytorch
autoaugment
mixed
Loss
Epoch
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
0
10
20
30
40
50

# Conclusion

Hochschule Offenburg
offenburg.university

- Cons:
  - No consensus on ratio of original to final dataset size
  - Hard to find best way
  - Can’t correct poor diversity with respect to the testing data
- Pros:
  - Can drastically improve performance
  - Better test time performance
  - Understanding of data

# Information

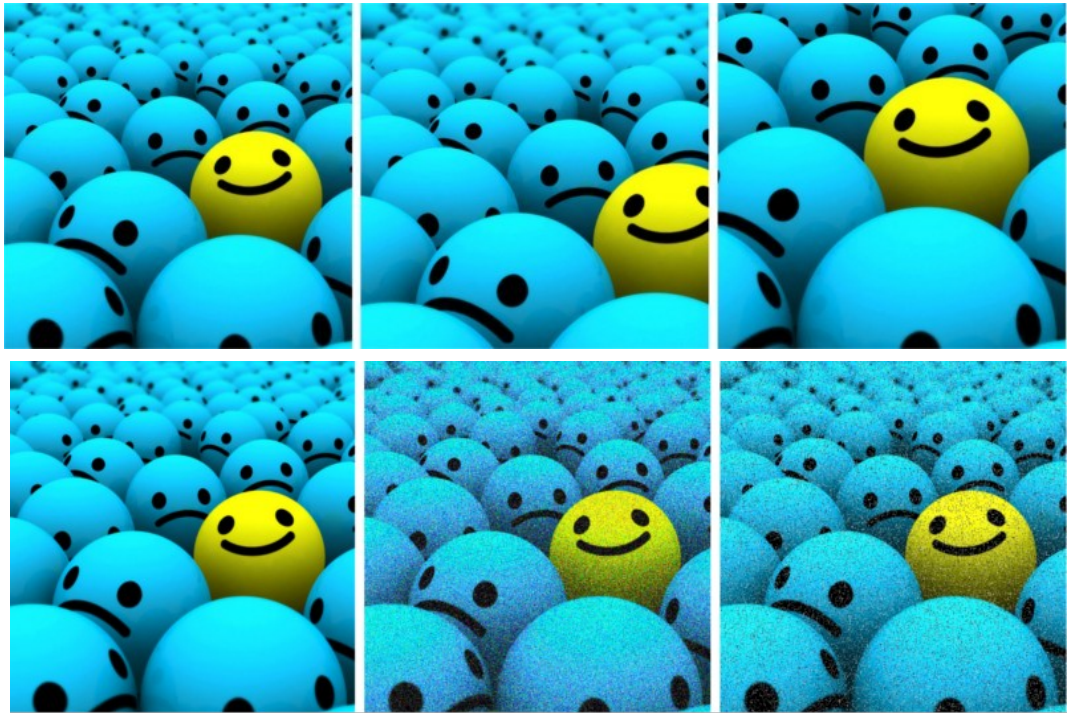
- **Github project repo:** https://github.com/Hexaa/dataaugmentation
- **Topic summary:** A survey on Image Data Augmentation for Deep Learning https://link.springer.com/article/10.1186/s40537-019-0197-0
- **Seminar paper:** The Effectiveness of Data Augmentation in Image Classification using Deep Learning https://arxiv.org/abs/1712.04621
- **Libraries:**
  - https://github.com/aleju/imgaug, https://github.com/albumentations-team/albumentations, https://github.com/kakaobrain/fast-autoaugment

    https://github.com/arcelien/pba

- **NVIDIA Dali:**

  https://docs.nvidia.com/deeplearning/sdk/dali-developer-guide/docs/index.html

# Thank you!



Questions?

# Sources


- [1]https://www.researchgate.net/figure/Data-augmentation-using-semantic-preserving-transformation-for-SBIR_fig2_319413978
- [2]https://link.springer.com/article/10.1186/s40537-019-0197-0
- [3]https://towardsdatascience.com/test-time-augmentation-tta-and-how-to-perform-it-with-keras-4ac19b67fb4d
- [4]https://github.com/kakaobrain/fast-autoaugment
- [5]https://docs.nvidia.com/deeplearning/sdk/dali-developer-guide/docs/index.html
- [6]https://github.com/arcelien/pba
- [7]https://github.com/DeepVoltaire/AutoAugment