Part 1 - How to Start and Common Issues

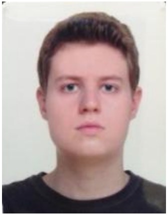
Training Deep Networks from Zero to Hero: avoiding pitfalls and going beyond

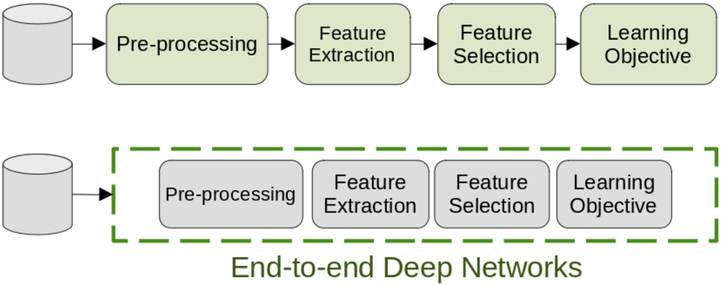
Moacir A. Ponti
Fernando P. dos Santos, Leo S. F. Ribeiro, Gabriel B. Cavalari

http://sites.google.com/moacirponti — moacir@icmc.usp.br

SIBGRAPI 2021







Pre-processing
Feature Extraction
Feature Selection
Learning Objective
Pre-processing
Feature Extraction
Feature Selection
Learning Objective
End-to-end Deep Networks

Agenda

How to Start

Common Issues

1 - Data quality

Basic principles

- Defined consistently with regards to the problem
- Cover all important cases
- Dataset size is appropriate

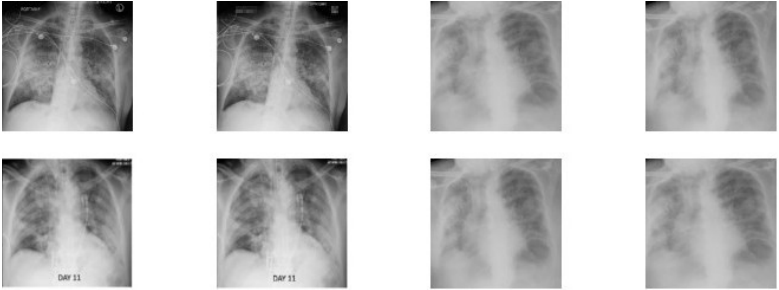
1 - Data quality

Basic principles

- Defined consistently with regards to the problem
- Cover all important cases
- Dataset size is appropriate

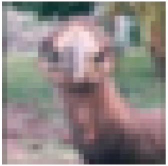
Problems to look for (even in benchmark datasets)

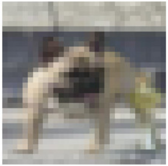
- Wrong labels
- Noise and outliers
- Duplicated instances

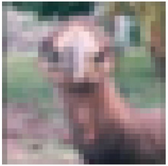

DAY 11
DAY 11

The ending made my heart jump up into my throat. I proceeded to leave the movie theater a little jittery. The movie was better than I expected. I don't know why it didn't last very long in the theaters or make as much money as anticipated. Definitely would recommend.

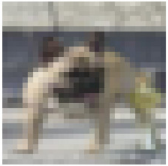
The ending made my heart jump up into my throat. I proceeded to leave the movie theater a little jittery. The movie was better than I expected. I don't know why it didn't last very long in the theaters or make as much money as anticipated. Definitely would recommend.
Label: Negative

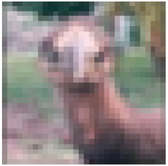




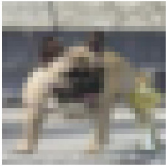


bird

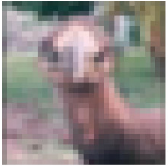




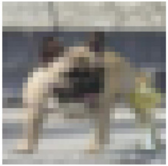
bird



cat

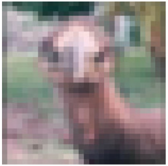


bird

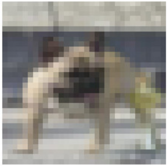


cat

automobile



bird



cat

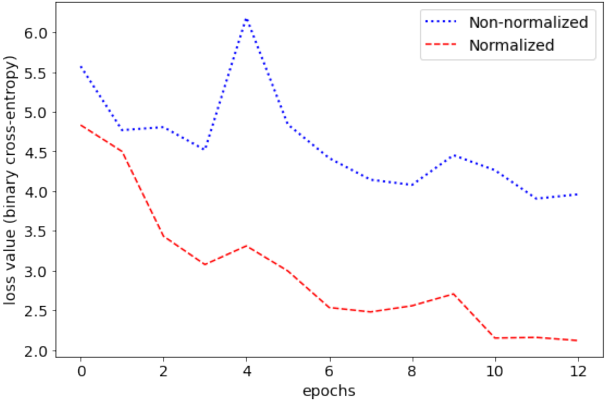
automobile

cat

2 - Normalization

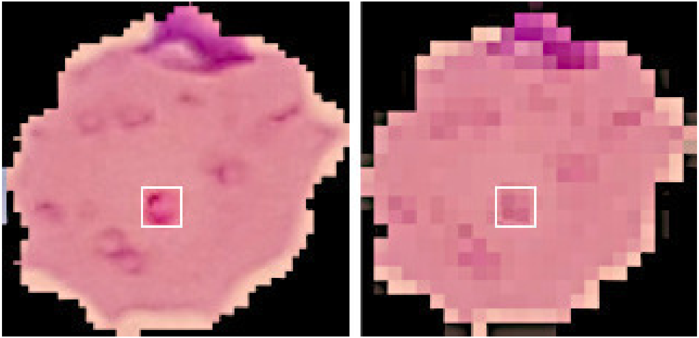
DNNs do not work well with arbitrary ranges.

- 0-1 MinMax Scaling
- z-Score Standardization



3 - Input Representation

After (or despite) pre-processing, the patterns required for learning the concept are present



4 - Loss Function and Evaluation Metrics

Accuracy does not fit all.

- How is output represented?
- What is the task to be solved?

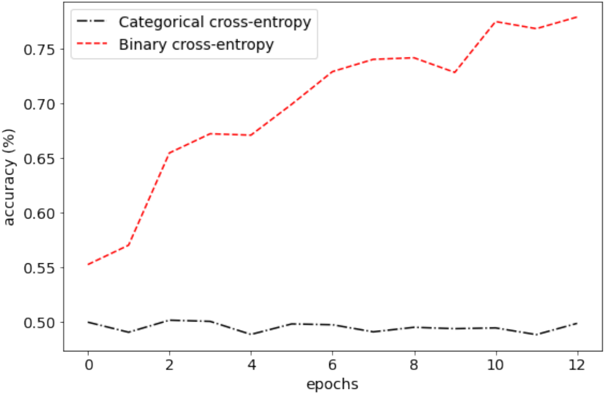
4 - Loss Function and Evaluation Metrics

Accuracy does not fit all.

- How is output represented?
- What is the task to be solved?

Tasks

- Binary classification
- Multi-class classification
- Regression
- Detection
- String matching
- Recommendation



...for a 2-class problem

Errors

- Mean squared error (MSE)
- Mean absolute error (MAE)
- Root of the mean squared error (RMSE)

Tendency to learn the "average" or mean output

Errors

- Mean squared error (MSE)
- Mean absolute error (MAE)
- Root of the mean squared error (RMSE)

Tendency to learn the "average" or mean output

Cross-entropy

Comparison between probability distributions

Errors

- Mean squared error (MSE)
- Mean absolute error (MAE)
- Root of the mean squared error (RMSE)

Tendency to learn the "average" or mean output

Cross-entropy

Comparison between probability distributions

Intersection over Union

Extent of overlap between boxes (regions)

Errors

- Mean squared error (MSE)
- Mean absolute error (MAE)
- Root of the mean squared error (RMSE)

Tendency to learn the "average" or mean output

Cross-entropy

Comparison between probability distributions

Intersection over Union

Extent of overlap between boxes (regions)

Cosine Distance

cosine of the angle between two vectors

Know your loss values

Example: 5 class problem

Categorical cross-entropy $-\sum y_i \ln(\hat{y}_i + 10^{-6})$

- All classes equally probable: $-1 \cdot \ln(1/5) = 1.60$
- Correct class with probability zero (and all other equally probable): 13.8

MSE $\sum(y_i - \hat{y}_i)^2$

- All classes equally probable: 0.80
- Correct class with probability zero (and all other equally probable): 1.25

5 - Model Tuning and Validation

You are given a training set and a test set.

There are: 2 architecture options and 4 learning rate values you want to investigate, what do you do?

5 - Model Tuning and Validation

You are given a training set and a test set.

There are: 2 architecture options and 4 learning rate values you want to investigate, what do you do?
An empirical evaluation:

1. For every architecture option you evaluate the 4 learning rates, generating 8 models
2. Evaluate those 8 models on the test set
3. Choose architecture and hyperparameter with highest score

5 - Model Tuning and Validation

You are given a training set and a test set.

There are: 2 architecture options and 4 learning rate values you want to investigate, what do you do?
An empirical evaluation:

1. For every architecture option you evaluate the 4 learning rates, generating 8 models
2. Evaluate those 8 models on the test set
3. Choose architecture and hyperparameter with highest score

WRONG. Why?

6 - Visualize feature space

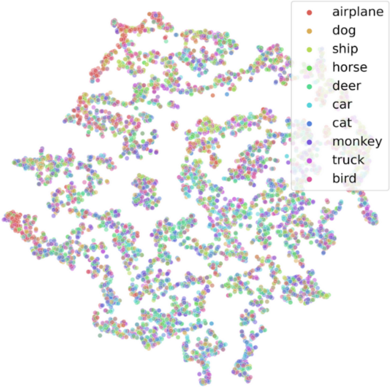
We trained a network for a difficult 10-class problem, in which we have a low number of instances.

Random classification would be 10%. We obtained 37% on a validation set. Great, right?

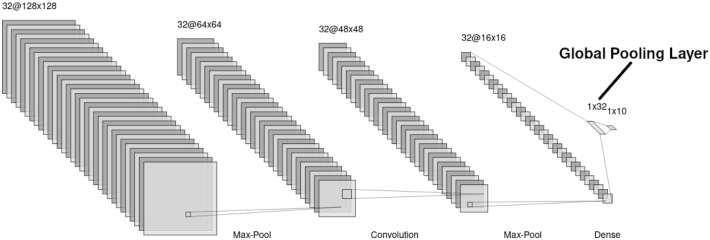
6 - Visualize feature space

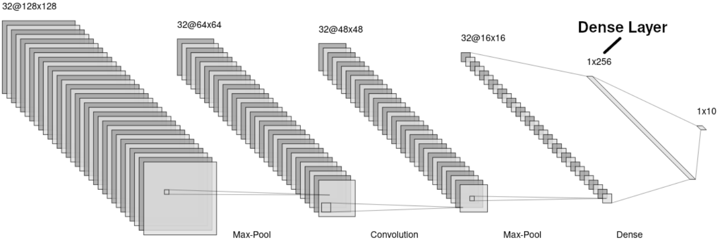
We trained a network for a difficult 10-class problem, in which we have a low number of instances.

Random classification would be 10%. We obtained 37% on a validation set. Great, right? Maybe not... no clear clusters are formed when projecting the feature space



Usually we visualize the output of the layer prior to prediction





7 - Internal and External Validation

A regression problem: data obtained from a given context. There is no fixed split for training and testing. What to do?

7 - Internal and External Validation

A regression problem: data obtained from a given context. There is no fixed split for training and testing. What to do?

Hold-out with 80/20 or k-fold cross-validation?

7 - Internal and External Validation

A regression problem: data obtained from a given context. There is no fixed split for training and testing. What to do?

Hold-out with 80/20 or k-fold cross-validation?

- this is *internal validation*: k-fold CV may help understand generalization.

7 - Internal and External Validation

A regression problem: data obtained from a given context. There is no fixed split for training and testing. What to do?

Hold-out with 80/20 or k-fold cross-validation?

- this is *internal validation*: k-fold CV may help understand generalization.

For sensitive data, it is important to have also **external validation**: use the model to evaluate on data acquire in other contexts

- in medical data this is crucial

Summary

- PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies
  https://www.acpjournals.org/doi/10.7326/M18-1376
  - participants, predictors, outcome, and analysis
- CLAIM: Checklist for Artificial Intelligence in Medical Imaging
  https://pubs.rsna.org/doi/10.1148/ryai.2020200029
  - study design, data, ground truth, evaluation, discussion

1. Data quality
2. Normalization
3. Input Representation
4. Loss Function and Evaluation Metrics
5. Model Tuning and Validation
6. Visualization
7. Internal vs External Validation

Agenda

How to Start

Common Issues

Small Datasets

- Regression: usually easier, require cover all output interval
- Classification: 1000-1500 instances per class to allow learning

If the dataset is too small, poor convergence will be observed

Small Datasets

- Regression: usually easier, require cover all output interval
- Classification: 1000-1500 instances per class to allow learning

If the dataset is too small, poor convergence will be observed

Alternatives

- Very small dataset ($n < 1000$): feature extraction using pre-trained networks and use of external classifier
- Small datasets ($1000 \leq n \leq 5000$): data augmentation and/or transfer learning

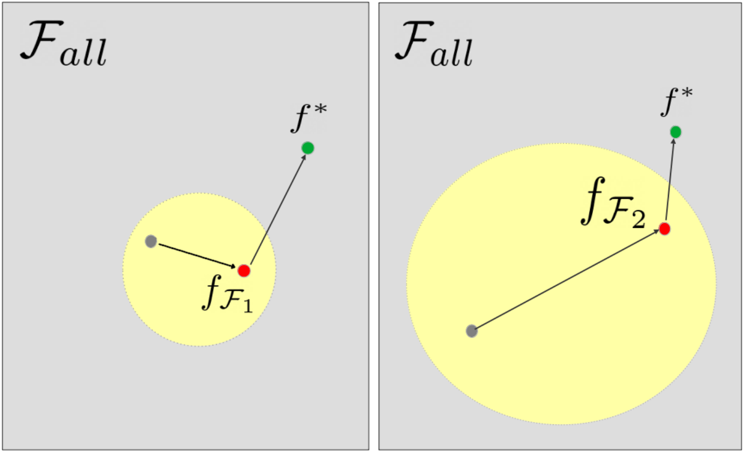
Imbalanced data

In this case be careful to use:

- metrics that take into account the unbalanced scenario
- make sure the distribution of future data will have similar distribution, or update the model if not
- instance weights by class so that the minority classes has more weight during training
- use data augmentation to mitigate the issue

Model complexity and generalization

The Bias-Variance dilemma and Deep Neural Nets



Model complexity and generalization

The problem is DNNs are often complex enough to include the **memory-based model**

The loss value after training is usually close to zero!

Model complexity and generalization

The problem is DNNs are often complex enough to include the **memory-based model**

The loss value after training is usually close to zero!

- Underfitting seldom occurs
- Overfitting is much more common: model performs worse in future data

Model complexity and generalization

The problem is DNNs are often complex enough to include the **memory-based model**

The loss value after training is usually close to zero!

- Underfitting seldom occurs
- Overfitting is much more common: model performs worse in future data

- Prefer lower capacity models if possible
- Give more importance to the data quality
- Evaluate rigorously

Models under attack

Training set without attack

CAR (56%)
AIRPLANE (30%)
HORSE (7%)

Pixel attack: if a malicious person has access to the training set...

Training set with pixel attack

CAR (56%)
AIRPLANE (30%)
HORSE (7%)

AIRPLANE (61%)
CAR (31%)
HORSE (6%)

Bibliography I

Moacir A. Ponti, Fernando dos Santos, Leo Ribeiro, Gabriel Cavallari. **Training Deep Networks from Zero to Hero: avoiding pitfalls and going beyond**.
SIBGRAPI, 2021. Tutorial.
https://arxiv.org/abs/2109.02752

Moacir A. Ponti, Leo Ribeiro, Tiago Nazaré, Tu Bui, John Collomosse. **Everything You Wanted to Know About Deep Learning for Computer Vision but were Afraid to Ask**.
SIBGRAPI-T, 2017. Tutorial.