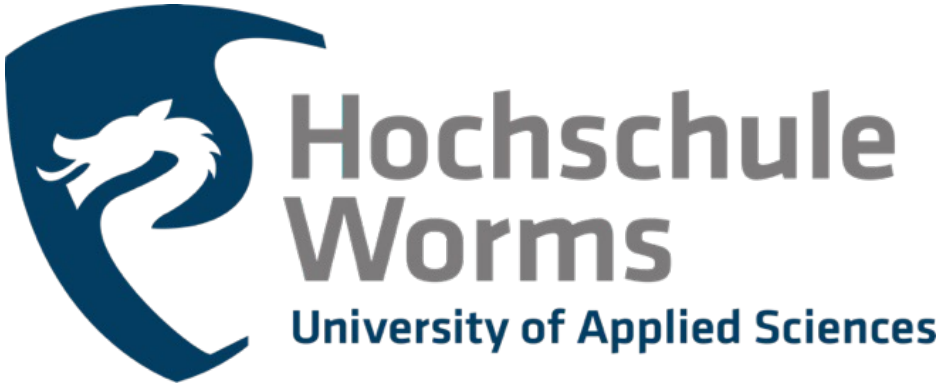
Architektur Neuronaler Netze für Generative KI

Vertiefungsaufgaben

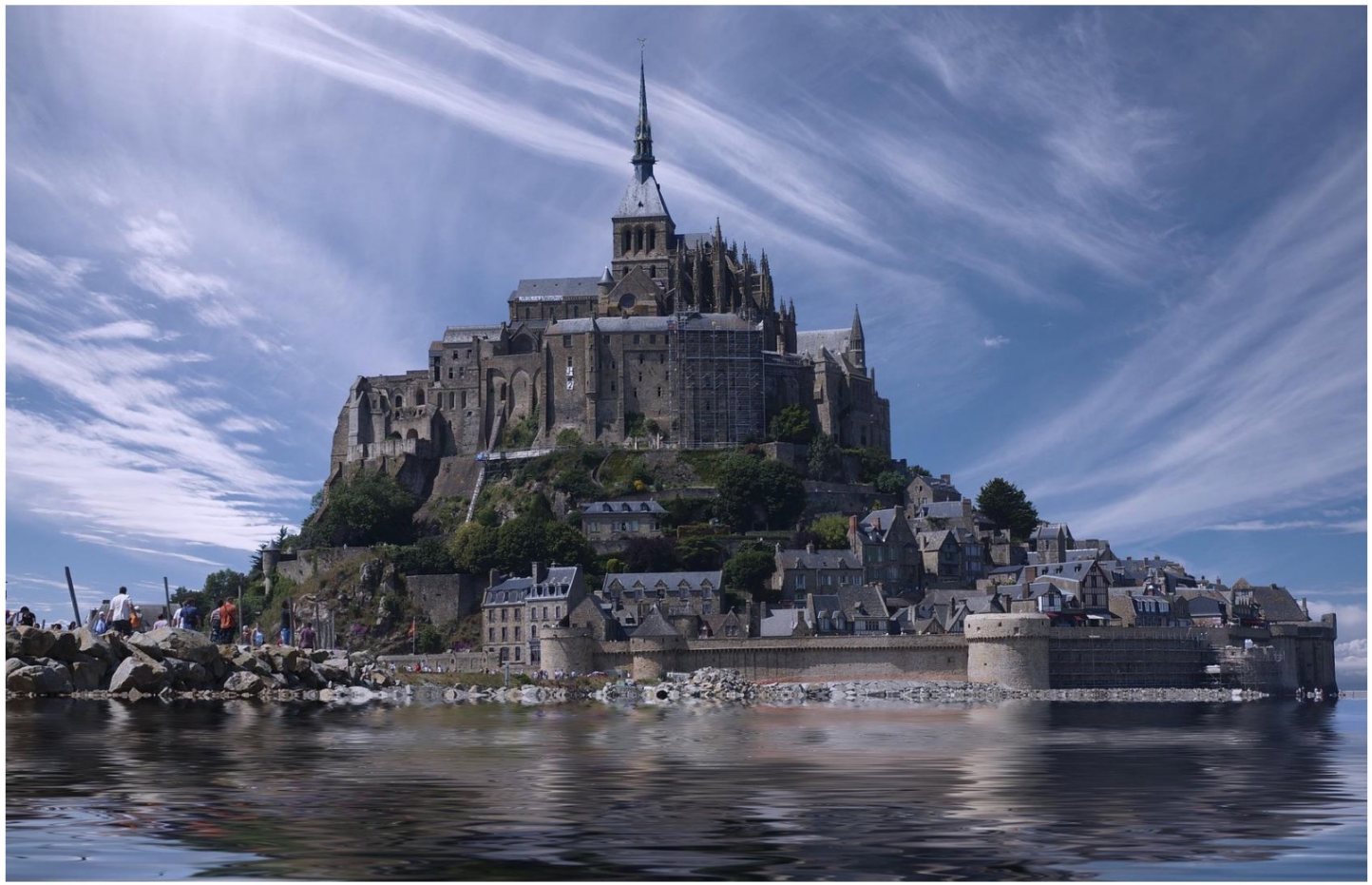
Hochschule Worms • Fachbereich Informatik

Prof. Dr. Stephan Kurpjuweit



1. Problem Statement

This example is adapted from a real production application, but with details disguised to protect confidentiality.



You are a famous researcher in the City of Peacetopia. The people of Peacetopia have a common characteristic: they are afraid of birds. To save them, you have **to build an algorithm that will detect any bird flying over Peacetopia** and alert the population.

Frage 1

The City Council gives you a dataset of 10,000,000 images of the sky above Peacetopia, taken from the city's security cameras. They are labeled:

- y = 0: There is no bird on the image
- y = 1: There is a bird on the image

Your goal is to build an algorithm able to classify new images taken by security cameras from Peacetopia.

There are a lot of decisions to make:

- What is the evaluation metric?
- How do you structure your data into train/dev/test sets?

Metric of success

The City Council tells you that they want an algorithm that

1. Has high accuracy.
2. Runs quickly and takes only a short time to classify a new image.
3. Can fit in a small amount of memory, so that it can run in a small processor that the city will attach to many different security cameras.

Note: Having three evaluation metrics makes it harder for you to quickly choose between two different algorithms, and will slow down the speed with which your team can iterate. True/False?

Frage 1

The City Council gives you a dataset of 10,000,000 images of the sky above Peacetopia, taken from the city's security cameras. They are labeled:

- y = 0: There is no bird on the image
- y = 1: There is a bird on the image

Your goal is to build an algorithm able to classify new images taken by security cameras from Peacetopia.

There are a lot of decisions to make:

- What is the evaluation metric?
- How do you structure your data into train/dev/test sets?

Metric of success

The City Council tells you that they want an algorithm that

1. Has high accuracy.
2. Runs quickly and takes only a short time to classify a new image.
3. Can fit in a small amount of memory, so that it can run in a small processor that the city will attach to many different security cameras.

Note: Having three evaluation metrics makes it harder for you to quickly choose between two different algorithms, and will slow down the speed with which your team can iterate. True/False?

Antwort: das ist korrekt.

Frage 2

The city revises its criteria to:

- "We **need** an algorithm that can let us know a bird is flying over Peacetopia as accurately as possible."
- "We *want* the trained model to take no more than 10 sec to classify a new image."
- "We *want* the model to fit in 10MB of memory."

Given models with different accuracies, runtimes, and memory sizes, how would you choose one?

Find the subset of models that meet the runtime and memory criteria. Then, choose the highest accuracy.

Take the model with the smallest runtime because that will provide the most overhead to increase accuracy.

Accuracy is an optimizing metric, therefore the most accurate model is the best choice.

Create one metric by combining the three metrics and choose the best performing model.

Frage 2

The city revises its criteria to:

- "We **need** an algorithm that can let us know a bird is flying over Peacetopia as accurately as possible."
- "We *want* the trained model to take no more than 10 sec to classify a new image."
- "We *want* the model to fit in 10MB of memory."

Given models with different accuracies, runtimes, and memory sizes, how would you choose one?

Find the subset of models that meet the runtime and memory criteria. Then, choose the highest accuracy.

Take the model with the smallest runtime because that will provide the most overhead to increase accuracy.

Accuracy is an optimizing metric, therefore the most accurate model is the best choice.

Create one metric by combining the three metrics and choose the best performing model.

Frage 3

Which of the following best answers why it is important to identify optimizing and satisficing metrics?

Identifying the metric types sets thresholds for satisficing metrics. This provides explicit evaluation criteria.

Identifying the optimizing metric informs the team which models they should try first.

Knowing the metrics provides input for efficient project planning.

It isn't. All metrics must be met for the model to be acceptable.

Frage 3

Which of the following best answers why it is important to identify optimizing and satisficing metrics?

- **Identifying the metric types sets thresholds for satisficing metrics. This provides explicit evaluation criteria.**
- Identifying the optimizing metric informs the team which models they should try first.
- Knowing the metrics provides input for efficient project planning.
- It isn't. All metrics must be met for the model to be acceptable.

Frage 4

Structuring your data

Before implementing your algorithm, you need to split your data into train/dev/test sets. Which of these do you think is the best choice?

Train	Dev	Test
6,000,000	1,000,000	3,000,000

Train	Dev	Test
9,500,000	250,000	250,000

Train	Dev	Test
6,000,000	3,000,000	1,000,000

Train	Dev	Test
3,333,334	3,333,334	3,333,334

Frage 4

Structuring your data

Before implementing your algorithm, you need to split your data into train/dev/test sets. Which of these do you think is the best choice?

Train	Dev	Test
6,000,000	1,000,000	3,000,000

Train	Dev	Test
9,500,000	250,000	250,000

Train	Dev	Test
6,000,000	3,000,000	1,000,000

Train	Dev	Test
3,333,334	3,333,334	3,333,334

Frage 5

Now that you've set up your train/dev/test sets, the City Council comes across another 1,000,000 images from social media and offers them to you. These images are different from the distribution of images the City Council had originally given you, but you think it could help your algorithm. You should add the citizens' data to the training set. True/False?

- True
- False

It is not a problem to have different training and dev distributions. In contrast, it would be very problematic to have different dev and test set distributions.

Frage 6

One member of the City Council knows a little about machine learning, and thinks you should add the 1,000,000 citizens' data images to the test set. You object because:

A bigger test set will slow down the speed of iterating because of the computational expense of evaluating models on the test set.

The test set no longer reflects the distribution of data (security cameras) you most care about.

This would cause the dev and test set distributions to become different. This is a bad idea because you're not aiming where you want to hit.

The 1,000,000 citizens' data images do not have a consistent x-->y mapping as the rest of the data.

Frage 6

One member of the City Council knows a little about machine learning, and thinks you should add the 1,000,000 citizens' data images to the test set. You object because:

A bigger test set will slow down the speed of iterating because of the computational expense of evaluating models on the test set.

The test set no longer reflects the distribution of data (security cameras) you most care about.

This would cause the dev and test set distributions to become different. This is a bad idea because you're not aiming where you want to hit.

The 1,000,000 citizens' data images do not have a consistent x-->y mapping as the rest of the data.

Frage 7

You train a system, and its errors are as follows (error = 100%-Accuracy):

Training set error	4.0%
Dev set error	4.5%

This suggests that one good avenue for improving performance is to train a bigger network so as to drive down the 4.0% training error. Do you agree?

- Yes, because having a 4.0% training error shows you have a high bias.
- No, because this shows your variance is higher than your bias.
- No, because there is insufficient information to tell.
- Yes, because this shows your bias is higher than your variance.

Frage 7

You train a system, and its errors are as follows (error = 100%-Accuracy):

Training set error	4.0%
Dev set error	4.5%

This suggests that one good avenue for improving performance is to train a bigger network so as to drive down the 4.0% training error. Do you agree?

- Yes, because having a 4.0% training error shows you have a high bias.
- No, because this shows your variance is higher than your bias.
- No, because there is insufficient information to tell.
- Yes, because this shows your bias is higher than your variance.

Frage 8

You ask a few people to label the dataset so as to find out what is human-level performance. You find the following levels of accuracy:

Bird watching expert #1	0.3% error
Bird watching expert #2	0.5% error
Normal person #1 (not a bird watching expert)	1.0% error
Normal person #2 (not a bird watching expert)	1.2% error

If your goal is to have “human-level performance” be a proxy (or estimate) for Bayes error, how would you define “human-level performance”?

0.4% (average of 0.3 and 0.5)

0.75% (average of all four numbers above)

0.3% (accuracy of expert #1)

0.0% (because it is impossible to do better than this)

Frage 8

You ask a few people to label the dataset so as to find out what is human-level performance. You find the following levels of accuracy:

Bird watching expert #1	0.3% error
Bird watching expert #2	0.5% error
Normal person #1 (not a bird watching expert)	1.0% error
Normal person #2 (not a bird watching expert)	1.2% error

If your goal is to have “human-level performance” be a proxy (or estimate) for Bayes error, how would you define “human-level performance”?

0.4% (average of 0.3 and 0.5)

0.75% (average of all four numbers above)

0.3% (accuracy of expert #1)

0.0% (because it is impossible to do better than this)

Frage 9

Which of the below shows the optimal order of accuracy from worst to best?

- ○ Human-level performance -> Bayes error -> the learning algorithm's performance.
- ○ The learning algorithm's performance -> human-level performance -> Bayes error.
- ○ Human-level performance -> the learning algorithm's performance -> Bayes error.
- ○ The learning algorithm's performance -> Bayes error -> human-level performance.

Frage 9

Which of the below shows the optimal order of accuracy from worst to best?

- ○ Human-level performance -> Bayes error -> the learning algorithm's performance.
- ○ The learning algorithm's performance -> human-level performance -> Bayes error.
- ○ **Human-level performance -> the learning algorithm's performance -> Bayes error.**
- ○ The learning algorithm's performance -> Bayes error -> human-level performance.

Frage 10

- You find that a team of ornithologists debating and discussing an image gets an even better 0.1% performance, so you define that as “human-level performance.” After working further on your algorithm, you end up with the following:

Human-level performance	0.1%
Training set error	2.0%
Dev set error	2.1%

Based on the evidence you have, which two of the following four options seem the most promising to try? (Check two options.)

- [ ] Get a bigger training set to reduce variance.
- [ ] Try increasing regularization.
- [ ] Try decreasing regularization.
- [ ] Train a bigger model to try to do better on the training set.

Frage 10

- You find that a team of ornithologists debating and discussing an image gets an even better 0.1% performance, so you define that as “human-level performance.” After working further on your algorithm, you end up with the following:

Human-level performance	0.1%
Training set error	2.0%
Dev set error	2.1%

Based on the evidence you have, which two of the following four options seem the most promising to try? (Check two options.)

- [ ] Get a bigger training set to reduce variance.
- [ ] Try increasing regularization.
- [ ] Try decreasing regularization.
- [x] Train a bigger model to try to do better on the training set.

Frage 11

11. You've now also run your model on the test set and find that it is a 7.0% error compared to a 2.1% error for the dev set. What should you do? (Choose all that apply)

- [ ] Increase the size of the dev set.
- [ ] Get a bigger test set to increase its accuracy.
- [ ] Try increasing regularization to reduce overfitting to the dev set.
- [ ] Try decreasing regularization for better generalization with the dev set.

Frage 11

11. You've now also run your model on the test set and find that it is a 7.0% error compared to a 2.1% error for the dev set. What should you do? (Choose all that apply)

- [ ] Increase the size of the dev set.
- [x] Get a bigger test set to increase its accuracy.
- [x] Try increasing regularization to reduce overfitting to the dev set.
- [ ] Try decreasing regularization for better generalization with the dev set.

Frage 12

After working on this project for a year, you finally achieve:

Human-level performance	0.10%
Training set error	0.05%
Dev set error	0.05%

What can you conclude? (Check all that apply.)

- [ ] This is a statistical anomaly (or must be the result of statistical noise) since it should not be possible to surpass human-level performance.
- [ ] If the test set is big enough for the 0.05% error estimate to be accurate, this implies Bayes error is ≤ 0.05
- [ ] It is now harder to measure avoidable bias, thus progress will be slower going forward.
- [ ] With only 0.05% further progress to make, you should quickly be able to close the remaining gap to 0%

Frage 12

After working on this project for a year, you finally achieve:

Human-level performance	0.10%
Training set error	0.05%
Dev set error	0.05%

What can you conclude? (Check all that apply.)

- [ ] This is a statistical anomaly (or must be the result of statistical noise) since it should not be possible to surpass human-level performance.
- [ ] If the test set is big enough for the 0.05% error estimate to be accurate, this implies Bayes error is ≤ 0.05
- [ ] It is now harder to measure avoidable bias, thus progress will be slower going forward.
- [ ] With only 0.05% further progress to make, you should quickly be able to close the remaining gap to 0%

Frage 13

- It turns out Peacetopia has hired one of your competitors to build a system as well. Your system and your competitor both deliver systems with about the same running time and memory size. However, your system has higher accuracy! However, when Peacetopia tries out your and your competitor's systems, they conclude they actually like your competitor's system better, because even though you have higher overall accuracy, you have more false negatives (failing to raise an alarm when a bird is in the air). What should you do?

- ◯ Look at all the models you've developed during the development process and find the one with the lowest false negative error rate.
- ◯ Ask your team to take into account both accuracy and false negative rate during development.
- ◯ Rethink the appropriate metric for this task, and ask your team to tune to the new metric.
- ◯ Pick false negative rate as the new metric, and use this new metric to drive all further development.

Frage 13

- It turns out Peacetopia has hired one of your competitors to build a system as well. Your system and your competitor both deliver systems with about the same running time and memory size. However, your system has higher accuracy! However, when Peacetopia tries out your and your competitor's systems, they conclude they actually like your competitor's system better, because even though you have higher overall accuracy, you have more false negatives (failing to raise an alarm when a bird is in the air). What should you do?

- ○ Look at all the models you've developed during the development process and find the one with the lowest false negative error rate.
- ○ Ask your team to take into account both accuracy and false negative rate during development.
- ○ Rethink the appropriate metric for this task, and ask your team to tune to the new metric.
- ○ Pick false negative rate as the new metric, and use this new metric to drive all further development.

Frage 14

You've handily beaten your competitor, and your system is now deployed in Peacetopia and is protecting the citizens from birds! But over the last few months, a new species of bird has been slowly migrating into the area, so the performance of your system slowly degrades because your model is being tested on a new type of data. There are only 1,000 images of the new species. The city expects a better system from you within the next 3 months. Which of these should you do first?

- ◯ Add hidden layers to further refine feature development.
- ◯ Add the new images and split them among train/dev/test.
- ◯ Put them into the dev set to evaluate the bias and re-tune.
- ◯ Augment your data to increase the images of the new bird.

Frage 14

You've handily beaten your competitor, and your system is now deployed in Peacetopia and is protecting the citizens from birds! But over the last few months, a new species of bird has been slowly migrating into the area, so the performance of your system slowly degrades because your model is being tested on a new type of data. There are only 1,000 images of the new species. The city expects a better system from you within the next 3 months. Which of these should you do first?

- ○ Add hidden layers to further refine feature development.
- ○ Add the new images and split them among train/dev/test.
- ○ Put them into the dev set to evaluate the bias and re-tune.
- ○ Augment your data to increase the images of the new bird.

Frage 15

15. The City Council thinks that having more Cats in the city would help scare off birds. They are so happy with your work on the Bird detector that they also hire you to build a Cat detector. (Wow Cat detectors are just incredibly useful, aren't they?) Because of years of working on Cat detectors, you have such a huge dataset of 100,000,000 cat images that training on this data takes about two weeks. Which of the statements do you agree with? (Check all that agree.)

Needing two weeks to train will limit the speed at which you can iterate.

Having built a good Bird detector, you should be able to take the same model and hyperparameters and just apply it to the Cat dataset, so there is no need to iterate.

If 100,000,000 examples is enough to build a good enough Cat detector, you might be better off training with just 10,000,000 examples to gain a $\approx$10x improvement in how quickly you can run experiments, even if each model performs a bit worse because it's trained on less data.

Buying faster computers could speed up your teams' iteration speed and thus your team's productivity.

Frage 15

15. The City Council thinks that having more Cats in the city would help scare off birds. They are so happy with your work on the Bird detector that they also hire you to build a Cat detector. (Wow Cat detectors are just incredibly useful, aren't they?) Because of years of working on Cat detectors, you have such a huge dataset of 100,000,000 cat images that training on this data takes about two weeks. Which of the statements do you agree with? (Check all that agree.)

- [x] Needing two weeks to train will limit the speed at which you can iterate.
- [ ] Having built a good Bird detector, you should be able to take the same model and hyperparameters and just apply it to the Cat dataset, so there is no need to iterate.
- [x] If 100,000,000 examples is enough to build a good enough Cat detector, you might be better off training with just 10,000,000 examples to gain a $\approx$10x improvement in how quickly you can run experiments, even if each model performs a bit worse because it's trained on less data.
- [x] Buying faster computers could speed up your teams' iteration speed and thus your team's productivity.

Frage 16

The essential difference between an optimizing metric and satisficing metrics is the priority assigned by the stakeholders. True/False?

- False
- True

Frage 16

The essential difference between an optimizing metric and satisficing metrics is the priority assigned by the stakeholders. True/False?

- False
- True

Stakeholders must define thresholds for satisficing metrics, leaving the optimizing metric unbounded.