What do you get when you combine computer vision, synthetic data, and curiosity?

Aleksei Terin


It all begun
biking past
Linnanmäki
when I thought to myself...

01

PROBLEM

“Wait, how many wristband tickets are there on the wall?”

TAKE A GUESS!

I mean it! How many tickets can you see on the wall?
342? 1337? Remember your number! I initially guessed 2000.

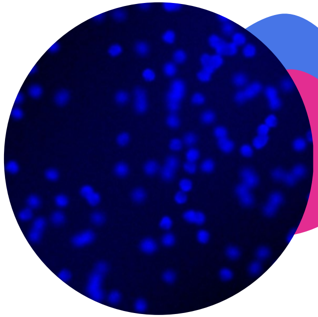
BUT WHY?

Okay, so you’ve picked a number.
But why care how many tickets there are?

It turns out that besides curiosity,

OBJECT COUNTING

is a salient problem spanning numerous disciplines!



Counting the number of cells can be crucial in diagnosis of e.g. cancer stages.

— MEDICAL IMAGING

Counting agricultural crops or livestock can help optimize resource allocation.

— AGRICULTURE

02

SOLUTION

So how can we count the tickets?

Well, let's consider

SOME APPROACHES

Approach	OK?	Reasoning
BY HAND	✗	**Too slow!**
ALGORITHM*	✗	**Too fragile, not likely to be accurate.**
ARTIFICIAL INTELLIGENCE	✓	**Incredibly powerful, lots of past/emerging research.**

*For instance, using Canny edge detection and a maximum filter.
It's ok if you've never heard of this, it's not crucial to understanding the work!

03
BACKGROUND

Ok, so how do we build the AI (machine learning)?

THE INGREDIENTS

for machine learning (simplified)

MODEL

How the 'brains' of our artificial intelligence will be structured

1

FEEDBACK

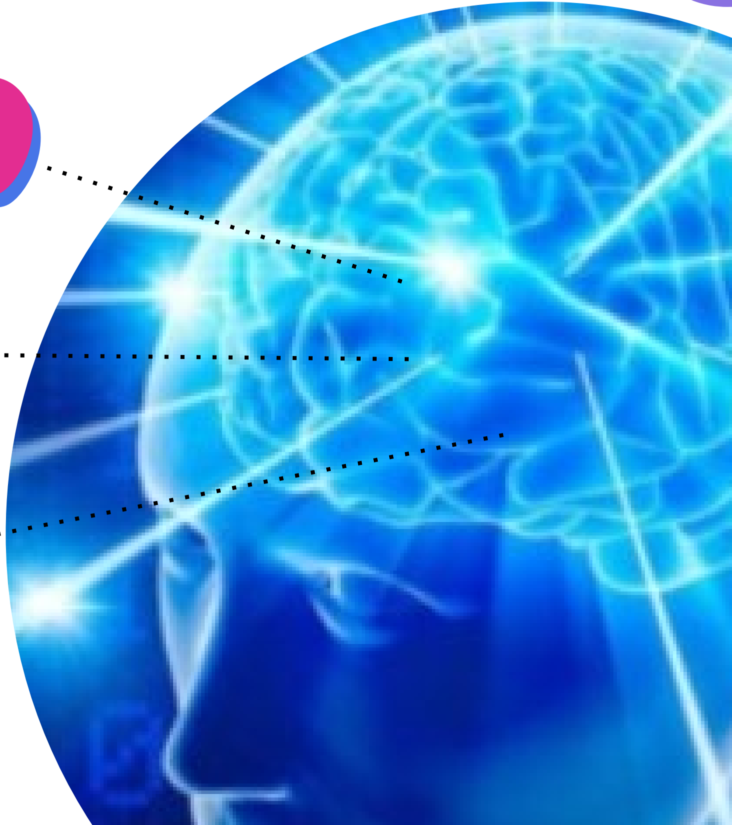
We need to tell the model whether it's doing good or bad, and how to improve

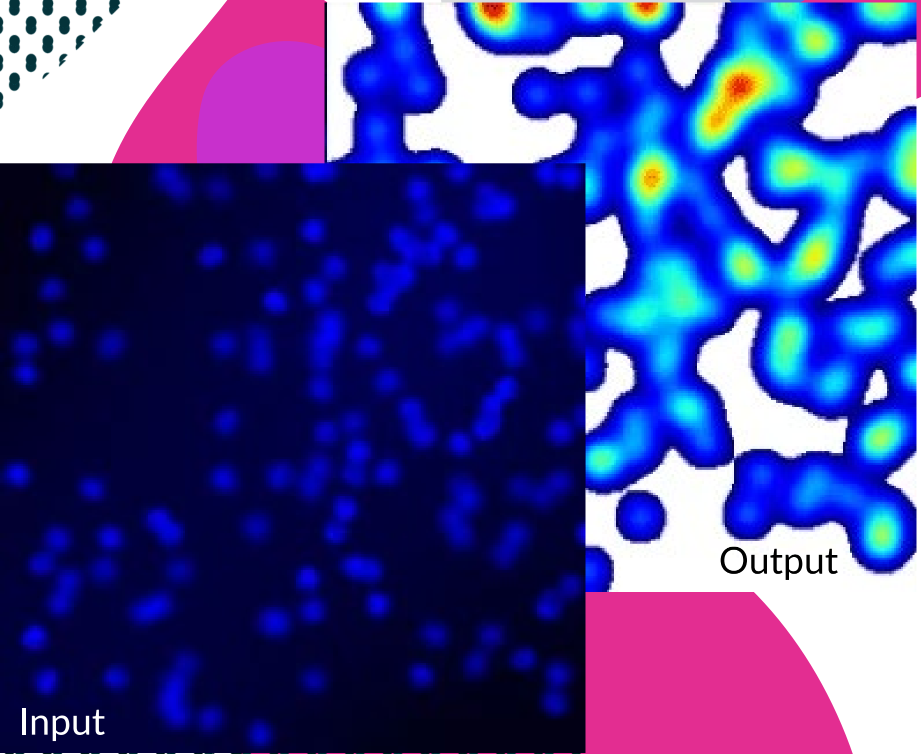
2

DATA!!!

Lots of teaching material that the model can internalize and use to make predictions

3





MODEL

For the purposes of this study, I opted for a convolutional neural network architecture known as U-Net*.

In a nutshell, the model takes an image as an input, and spits another image back out, which we will use as a heatmap for the count (and position) of detected tickets.

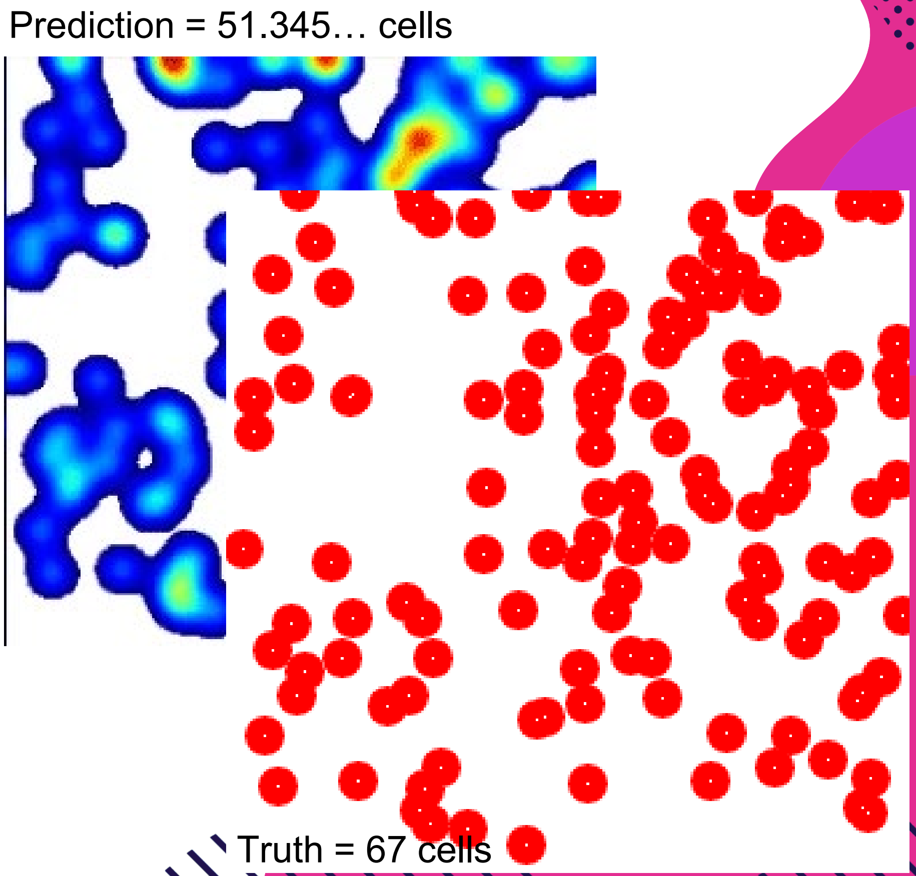
*It's not crucial to understand how this specific neural network works in detail.

FEEDBACK

Like a human, a model needs feedback to learn.

In this study, we will tell the model how close its guess of the number of tickets is to the actual number of tickets in the picture.

Additionally, we will evaluate how well the model predicts the locations of these tickets on the heatmap!



BUT WAIT!

We have a problem...

WHERE DO WE GET THE DATA?

CHATGPT

Models like ChatGPT use mountains of text/content compiled from nearly the entire Internet!

But there aren’t readily available pictures of Linnanmäki tickets that have been annotated with the correct counts/positions of tickets...

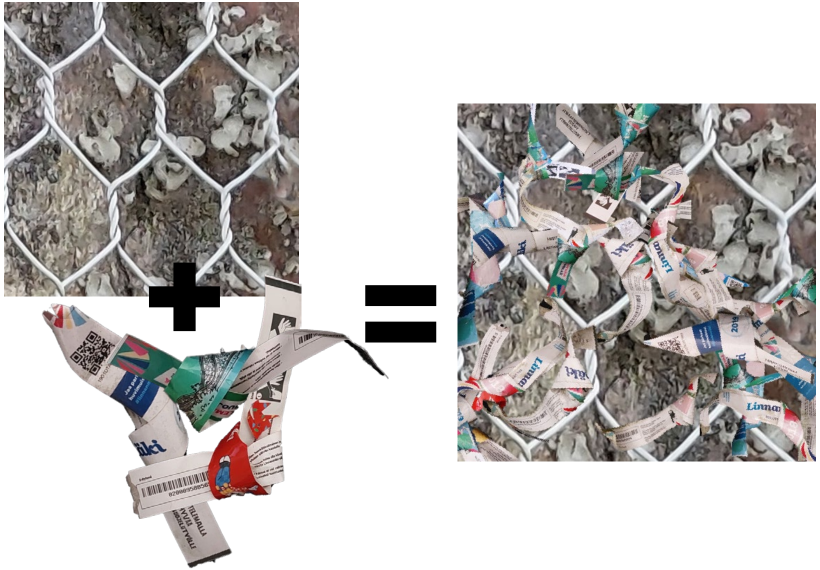
INTR
ODU
CING

SYNTHETIC DATA!

AN EXAMPLE

The picture of tickets on the right was never ‘taken’. It was filled with tickets by a computer algorithm!

This means we can easily generate thousands of pictures, and label how many tickets are in each picture!



Synthetic data is “one of the most promising general techniques on the rise in modern deep learning, especially computer vision.”

— **NVIDIA**

Source: https://blogs.nvidia.com/blog/what-is-synthetic-data/

and so a

RESEARCH QUESTION

was born...

To what extent is it viable to use **synthetically generated data** to train a computer
vision model for **object counting** and **localization***?

*localization refers to determining the positions/locations of objects in an image

04

METHODOLOGY

So how do we find out if synthetic data ‘works’?

STEP-BY-STEP

01 NATURAL

Use natural (not synthetic) data as a benchmark.

02 SYNTHETIC

Generate synthetic data to match its natural counterpart.

03 TRAIN

Train models on different proportions of natural to synthetic data.

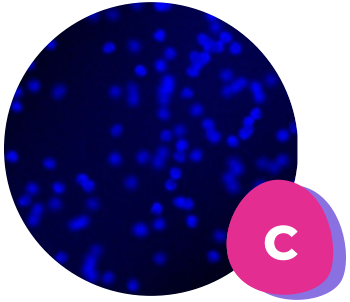
04 EVALUATE

Evaluate performance gain/loss in response to synthetic data.

Besides the ‘tickets’ dataset, this study incorporates a total of

THREE NATURAL DATASETS

to obtain more generalized results



CELLS

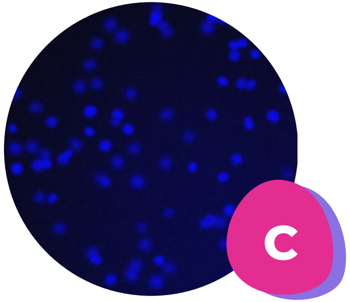
BLUEBERRIES

TICKETS

And here are their

SYNTHETIC COUNTERPARTS

generated to not only visually, but also statistically match the natural datasets



CELLS

BLUEBERRIES

TICKETS

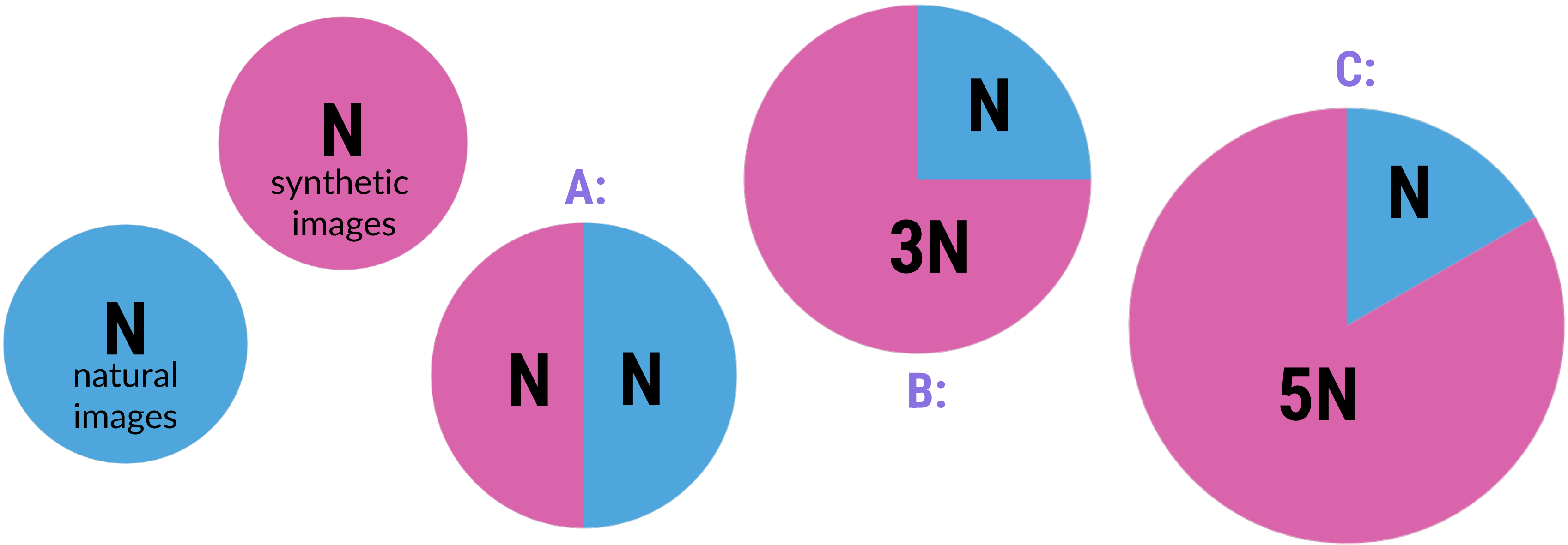
04.1

VARIABLES

So what are we varying/measuring?

INDEPENDENT:

the proportion of synthetic to natural data, varied across five qualitative categories



N = the number of images in the original, natural dataset, e.g. 100.

You may notice this means the size of training data is intentionally ***not*** constant, since synthetic data can easily boost the quantity of training images.

DEPENDENT:

the model's ability to (1) count and (2) localize objects

1

2

Mean Average Error

Intuitively, MAE tells us how far away the model's predicted number of objects is from the ground truth. In other words, the uncertainty (±) in its numeric count prediction.

F1-score

The F1-score originates from object detection. In a nutshell, it tells us how well the model predicts the locations of objects, ranging from 0% (worst) to 100% (best).

CONTROL:

a complete understanding of these is **not** necessary, rather they are provided for completeness

Variable	Value
Learning rate	0.01
Batch size	8
U-Net filters	64
Convolution layers per block	2

Table 1: The control parameters used for training and validation of the model.

05

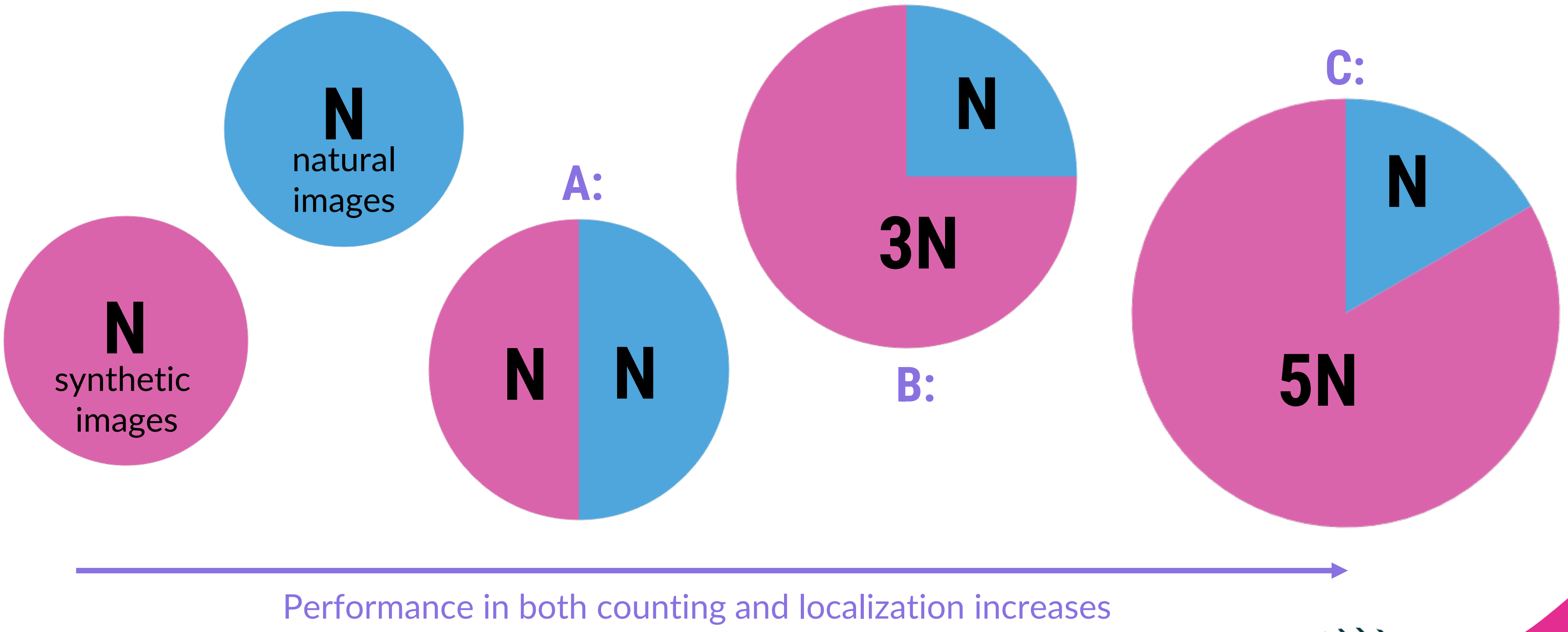
HYPOTHESIS

Place your bets!

My contention was based on a simplistic generalization that more training data generally leads to better performing models.

Hence, I hypothesized that the performance in both object counting and localization will **increase linearly** with the volume of training data (and so proportion of synthetic data).

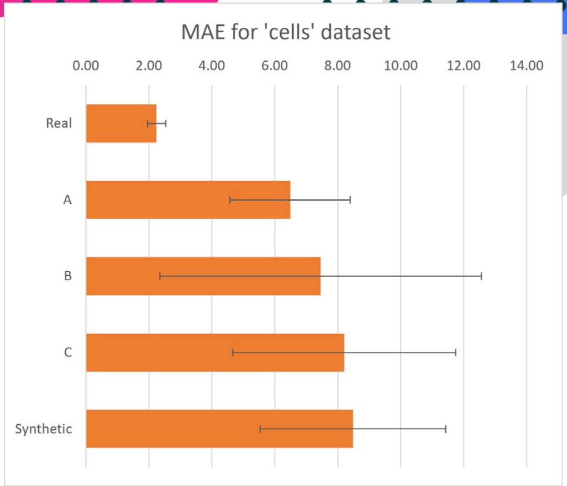
MY HYPOTHESIS
in a nutshell

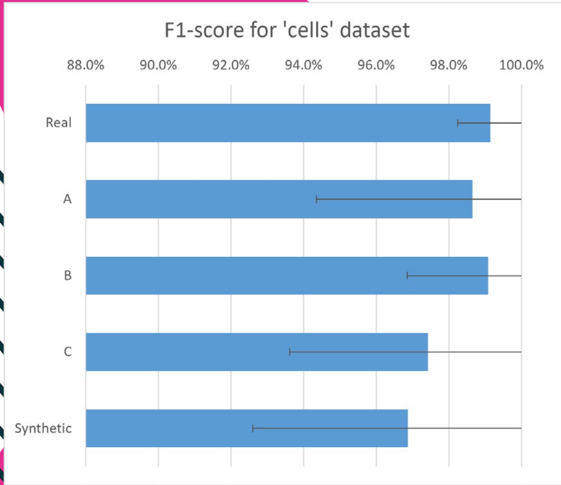


06

RESULTS

The most exciting part!





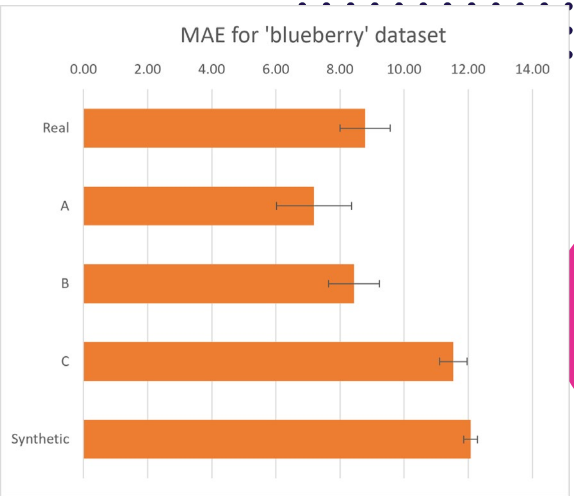
CELLS

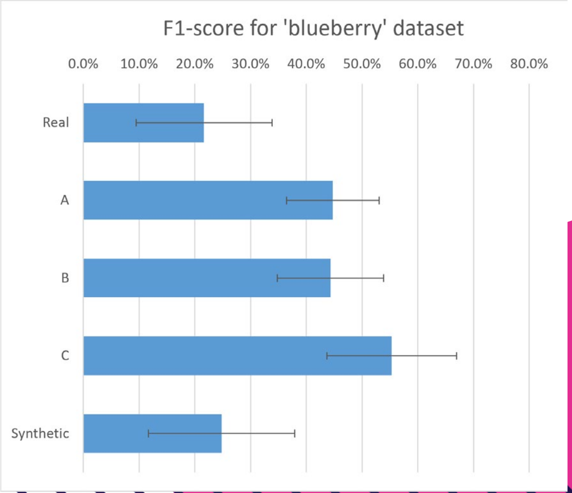
- The model trained on purely real data outperforms its purely synthetic counterpart in both object counting (MAE) and localization (F1-score).
- Despite an increase in the volume of total training data ('A', 'B' and 'C'), synthetic data does not help and in fact hinders* the model.

*the likely explanation for this is overfitting, which is addressed in the paper, but somewhat difficult to condense into a presentation

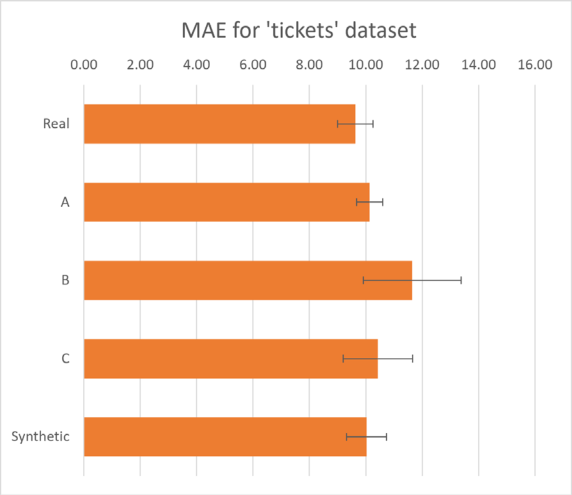
BLUEBERRIES

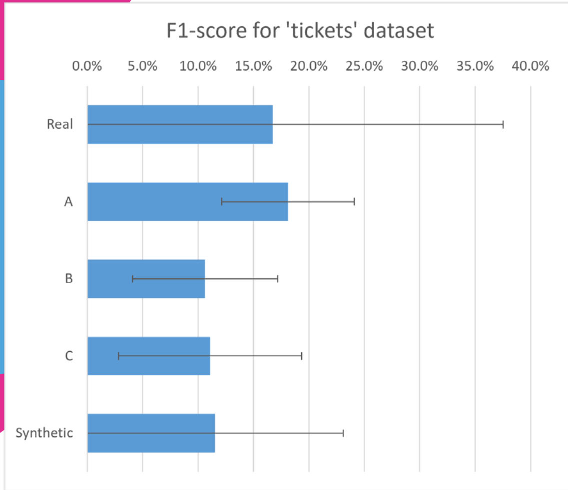
- The networks trained on ‘A’ and ‘B’ outperform the purely natural training data across both counting (MAE) and localization (F1-score), supporting the hypothesis.
- As proportion of synthetic data increases (‘C’), performance in counting declines, but in localization improves.*





*this is a curious finding that would have to be investigated further.





TICKETS

- Results are closely tied across both counting (MAE) and localization (F1-score).
- This suggests the task hand is just super difficult (many overlapping tickets), irrespective of the nature of training data.*

*a different approach to counting/localizing tickets would likely be required to see better results

SUMMARY:

As proportion of synthetic data increased…

CELLS

- Mean average error in counting increased
- Localization (F1-score) worsened

👎

BLUEBERRIES

- Mean average error in counting was minimal at 1 to 1 ratio
- Localization improved continuously

👍

TICKETS

- Both MAE and F1-scored remained tied; no significant change in counting and localization

Inconclusive

In cases where natural data was incomplete (like the blueberries dataset), synthetic data could provide substantial performance improvements.

However, otherwise or in high concentrations,
it could also hinder the model by overly ‘diluting’ the training set.

— Takeaway

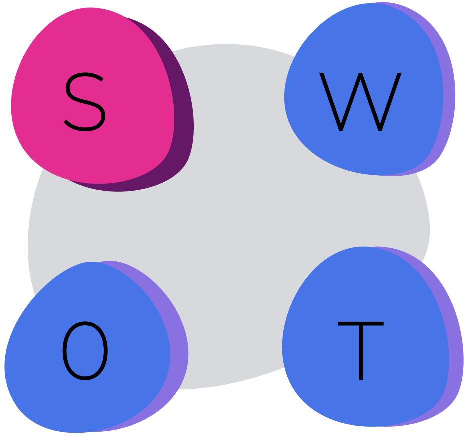
07
REFLECTION

What was good in this study? What could be more developed?

STRENGTHS

Diverse range of datasets (blueberries, cells, tickets)

Results computed and averaged over numerous trials, with proper error bars

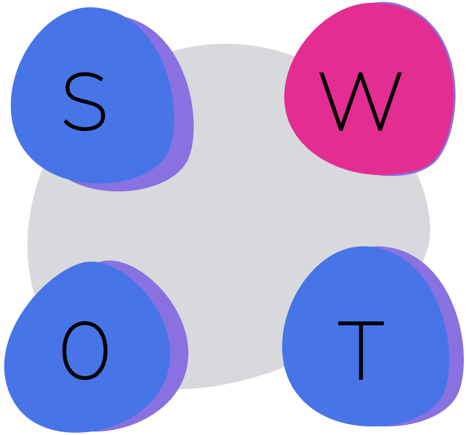


Critical reflection on the implications of results

Fuses personal curiosity with a real-world problem

WEAKNESSES

Synthetic data is generated somewhat arbitrarily on a case-by-case basis

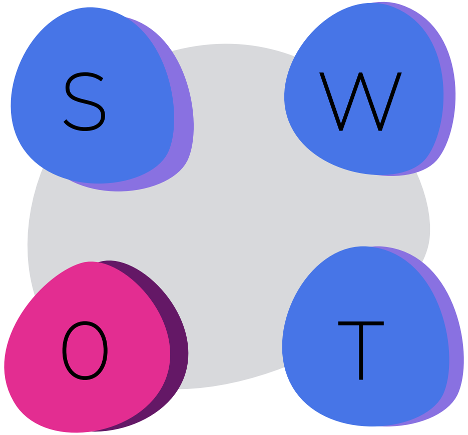


Only one neural network architecture explored (U-Net)

The five qualitative categories for proportions of synthetic to natural data are rather arbitrary

Inconclusive results for the ‘tickets’ dataset

FUTURE OPPORTUNITIES



Design an algorithm (possibly another neural network) that can generate synthetic data systematically

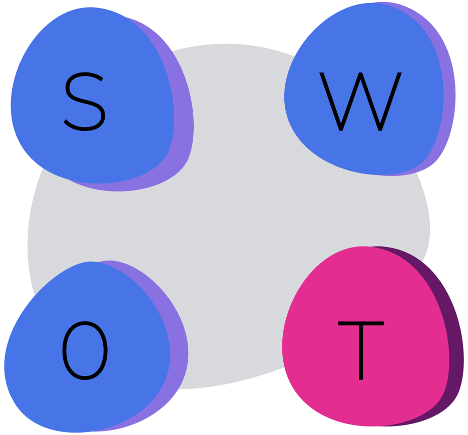
Explore other neural network architectures (e.g. FCRN-A)

Further analyze more categories/proportions of synthetic to natural data

Evaluate synthetic data in fields besides computer vision (e.g. natural language processing)

THREATS

- There exists some concern over research into computer vision due to fears of heightened government surveillance or the existential threat of AI.



However, I believe that with timely regulation, the net effect of research into computer vision and machine learning lies much farther on the positive side.

So what have I learned?

- A lot of machine learning is trial-and-error
- Quantity of data without quality is meaningless
- Scientific reporting principles (referencing, reading others' publications)
- Science is fun ☺

08

WAIT!

There's more?

DO YOU REMEMBER WHICH NUMBER YOU PICKED?

the

CORRECT ANSWER

is...


3844
TICKETS!


3844
TICKETS!
± 1472*
*ye the uncertainty is pretty large :)

THANKS!

Want to find out how this number (3844±1472) was calculated or have other questions?

Come talk to me at the stand!