# Finding the most beneficial annotation strategy

## Public Project Summary

Péter Paziczki

I am working at a company that is developing an autonomous driving software, highly relying on neural networks when it comes to perception. I am part of a team that creates the labelled data for these neural networks and am facing questions, such as what frames to label to achieve the highest possible neural network performance, of course considering the financial limits.

Our hypothesis was that we could find an optimal frame distance, a maximum point where the highest network performance was achieved considering a given budget. We were assuming there was a point where the images we have labelled were uncorrelated enough, so the neural networks could learn the most from them and number of annotated frames were enough to provide the highest possible performance. A maximum point had to exist for every budget, because eventually too much operational cost would have been spent on recording and there wouldn't have been enough money left to annotate images. If said optimum could be found, that information could be used to optimize the recording and annotation strategy given a certain budget. We had a few other initial assumptions as well, such as the less correlated frames we used, the better network performance was achieved, and larger train sets allowed for better network performance.

To find that optimal frame distance, the best frame selecting strategy for a given budget and to validate our initial assumptions, we designed an experiment with three steps. We were looking at our costs budget wise, meaning that we had a limited amount of money per financial period to spend on recording and annotation, so first, we determined the CAPEX and OPEX cost items of recording and annotation and understood their relation. I interviewed all the stakeholders to understand what cost items were the most significant, with what overhead they worked with. Once I explored all the cost items, I was able to create a formula that described the relation between the CAPEX of recording and the OPEX of recording and annotation. It showed that the higher the frame distance was, the more of OPEX was spent on recording and the remained to spend on annotating images.

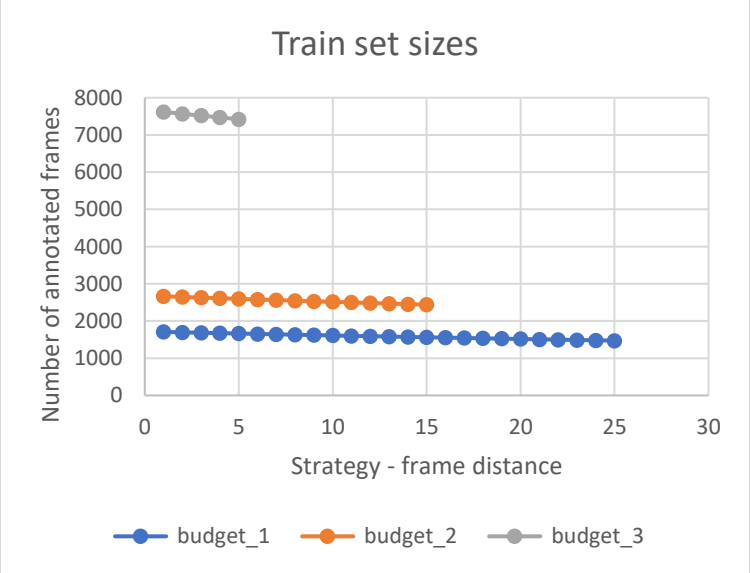


*Figure 1 Train set sizes*

Then we could create three arbitrary budgets and put training sets together within those budgets with two key differences between them, such as the number of frames kept as distance between the key frames and the number of frames that could be annotated. When this so-called frame distance is 1, direct neighbor frames are labelled. When the distance is 25, every 25$^{th}$ frame is labelled. Considering 25 frames per second, it would mean that basically 1 frame per second is labelled. We had an

issues with data availability, we only had enough data to experiment in the frame distance range of 1-25 frames.

The train data used for training have completely been recorded internally in a highway environment, taken by different cameras and different lenses, from different angles, and no publicly available annotations have been mixed into them.

The neural network retrained in the experiment was already being used in production (and has been developed internally), in actual testing on public roads.

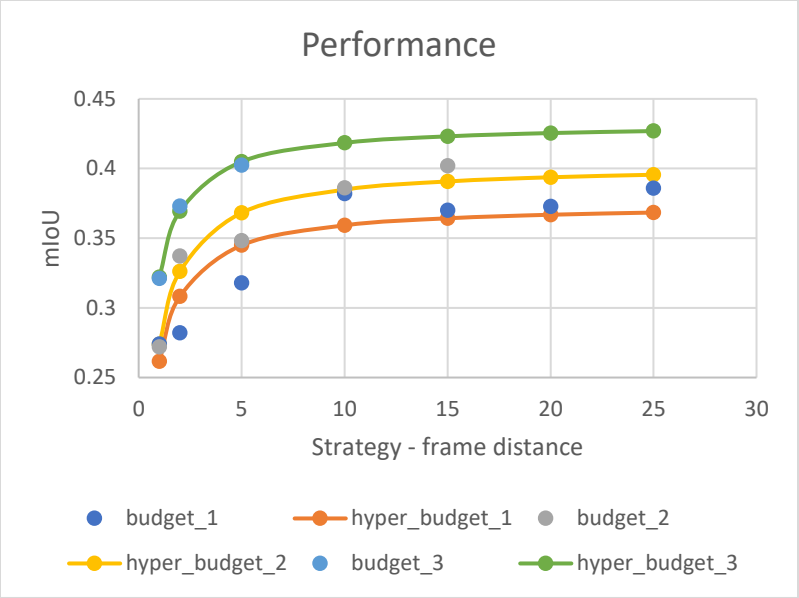


*Figure 2 Neural network performance measured by mIoU*

Once we had the train sets, we run trainings with every one of them. We measured the performance of our newly trained networks by mIoU (mean intersection over union).

The data set used to evaluate the performance of the newly trained networks was a public dataset, the images of it have been taken by different cameras, lenses, from different angles, at the different locations and none of these images have been used for training, thus the result were representative, even if they were not that outstanding.

To better illustrate the results, I have fitted a hyperbolic curve on the measured performance values of all three budgets, please find it on Figure 2. If we had repeated the training several times, the results would have scattered, but from experience we assumed that a hyperbolic curve fitted the points well enough.

Our initial assumptions were validated, the higher the frame distance was, the better network performance could be achieved. The larger the sets were, the better network performance was achievable, so both frame distance and frame count (number of annotated images) are powerful drivers.

The key findings and observations that I wanted to mention are highlighted on Figure 3. When looking at finding number 1, we could clearly see the single effect of train set sizes. By simply increasing the train set size we could achieve a significantly better performance. For budget_1 and budget_2 we practically had the same results, but for budget_3 we had a more than 15% better result.

When inspecting finding number 2, we could clearly see the evidence of how much frame distance mattered. Train set size of budget_3 (at frame distance of 1 frame) was 4,5 times higher than for budget_1 (at frame distance of 5 frames) and to create the train set for budget_3 cost 4 times more in terms of OPEX, yet we practically had the same mIoU results for both. It showed us how important it was to feed as uncorrelated images to the networks as possible. The more unique the frames were, the more the networks could learn from them. Finding number 3 was a similar case, for budget_3 we had a 3 times larger train set size and cost more than 2 times more in terms of OPEX, yet we had the same performance results. Frame distance is a very significant variable.

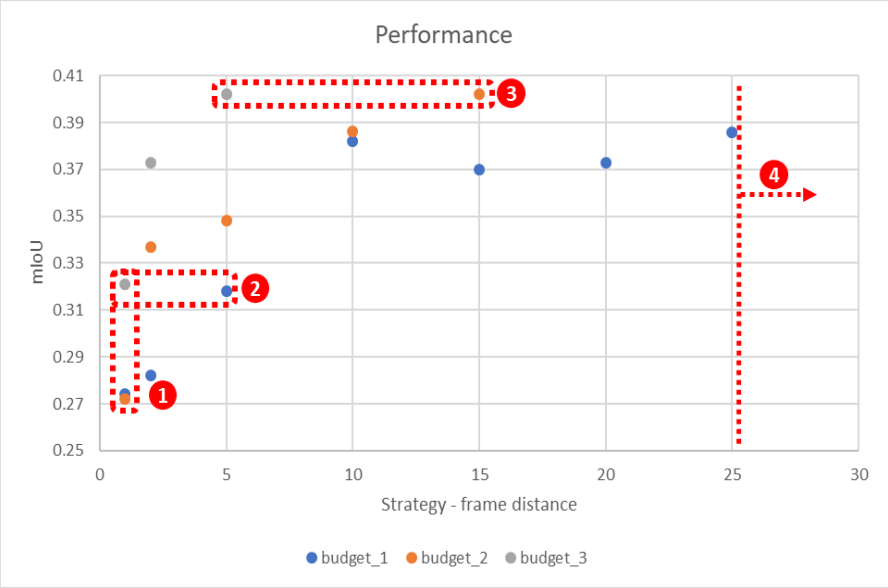


*Figure 3 Key findings*

Finding number 4 referred to the fact that the optimum we were looking for, practically the maximum of the curves, couldn't be found in the range we have experimented with. As I have mentioned earlier, we have experimented with this amount of data, because that was only available. Prior to this experiment a different strategy was followed (adjacent frames were annotated) and thus there was not enough data to experiment with frame distances higher than 25 frames. If we had more data, other frame distances would have been assessed.

If we wanted to choose a frame distance to continue our work with, it still would be an arbitrary choice. Our educated guess based on experience and domain knowledge was that the optimal frame distance for highway environment might be between 500-750 frames.

My key suggestion was to continue the experiment with targeted annotations first, so we'd have enough data to repeat the experiment but assessing another range that time, 25 to 750 frames. The experiment could be conducted only with a limited sized data set because we saw that train set sizes less than 2k images could provide meaningful insights. But it certainly is to be repeated with orders of magnitude more data, once it is available, because that would provide much more reliable results.

Polygon annotation was not the only form of annotation done in house, polyline annotation made up to a significant part of our efforts, but the results couldn't be generalized to that. It would be quite beneficial to repeat the experiment with polyline annotations, that would instantly bring some optimization to the field.