Capstone Project Summary

SIAMESE NETWORKS:

RANKING OBJECTS BY POORLY DEFINED SIMILARITY

An industrial use-case

By
Péter Endes-Nagy

Submitted to Central European University - Private University
Department of Economics and Business

In partial fulfilment of the requirements for the degree of Master of Science in Business Analytics

Supervisor: János Divényi

Budapest, Hungary
2022

Business case

A machine learning tool was developed for a Hungarian SME, specialised on manufacturing custom-made machine tools and metal products. As the products are custom-made and manufactured in low serial numbers, the unit cost of creating operation plans (series of manufacturing steps) is relatively high. Approximately 50 new objects arrive a day, many of them are the same or at least very similar to previously manufactured items. Previous operation plans are occasionally re-used when then engineers remember them, but 1) in an ad-hoc manner, 2) knowledge isn't transferable among engineers, and 3) it becomes less effective as the number of objects increases.

Building a ML tool that retrieves Top-k similar objects from the historical database has high added value as it saves time and energy: engineers don't need to create the operation plan from scratch but use the previous plan as it is or slightly modified.

Proposed solution: A Siamese network trained on similar and dissimilar object pairs. The model's output is a similarity score calculated pairwise between the new object and items in the historical dataset, then Top-k object is retrieved from the ordered list.

Challenges and work done

Similarity should be understood in terms of how to manufacture the object and not per its geometrical properties – when a new object arrives, the geometrical properties are available, so the task is predicting similarity in terms of how to manufacture, using geometrical features. Identifying similar operation plans is less than straightforward for various reasons: 1) step naming isn't standardized, 2) continuous technological change, 3) involvement of sub-contractors due to lack of capacity or technology at a given point of time, 4) some steps can be arbitrary, ambiguous, redundant, and interchangeable.

Labels (similar or not) are indispensable for supervised learning; therefore, similar and dissimilar pairs were identified with conservative heuristics. With these strict heuristics, over a quarter of the label space can be covered, predominantly with negative labels. The main issue with this identification strategy is that it focuses on very similar/dissimilar objects and positive pairs are easier to be found among simple objects. It results in potential overspecialization on simple objects while retrieving similar examples for more complex objects has higher added value for the company. The issue can be partially mitigated with down-sampling, but we can't assess the degree and prevalence of possible overspecialization.

Evaluating models and choosing between them is also challenging, therefore a custom evaluation metric was developed. Difficulty of evaluating the model's performance stems from the following reasons: 1) label space is partially covered, 2) labels are unbalanced and strict 3) an object might have no similar pair in the dataset at all 4) the task is ranking the full historical dataset by similarity and retrieve the Top-k results and unlabeled items might populate the Top-k.

Siamese neural networks fit well the business case as they predict similarity score between 2 inputs instead of classifying them into the unknown number of classes (classes are "operation plan types" in this business case). They are widely used in image classification and facial-recognition use-cases as 1) small number of training data is enough 2) they can deal with unknown number of classes 3) and new classes.

Three Siamese network candidates were built, and 10 different experiment (length of training and size of training sample) was conducted for each of the three neural network architectures. Models were evaluated with the custom metric and classic binary classification metrics, like Recall and Precision. Ordering the historical items by cosine similarity, calculated from the raw and standardized object features served as reference models.

Key results and insights

Siamese networks are a viable solution, most of them manage to converge and yield good results, but choice of network architecture is key – simpler network works better for this business case.

The smart choice of standardization (during feature engineering) already yielded surprisingly good results, the Siamese networks further improved the positive effect of standardization by bringing up more of the similar items to the top, but not to the very top. Recall and Precision metrics inspected over the learning process also support the insight that the Siamese networks manage to distinguish between similar and dissimilar objects (the more they learn), but at the price of not bringing them to the very top (lower similarity score predicted). Therefore, a wider Top-k (25 instead of 10) is recommended in order to fully exploit the model's merits.

The network's architecture opens the door for further online training. Putting the model into production and give it feedbacks during usage would be particularly beneficial in shedding light on the grey area. The model was trained with extremely similar object pairs (identical operation plans), items that are "close to similar" or "similar but unidentified with the heuristics" could be precious inputs for the model, particularly for complex objects.