Counterfactual Explanations for Deep Learning-Based Traffic Forecast

Rushan Wang
ETH Zurich

Deep learning models are widely used in traffic forecast tasks and achieve high accuracy. However, the black-box nature of those models makes the results hard to be explained and trusted by users. On the one hand, the lack of interpretability and explainability makes machine learning developers not aware of how the model learns from the data. On the other hand, domain users need more explanations on using the model to gain a more insightful understanding of the real world.

In recent years, the issues of interpretability and explainability in AI have gained more and more attention from researchers. Interpretability focuses on creating models whose internal workings are inherently understandable. Explainability aims to furnish post hoc insights into existing "black-box" models to elucidate their decision-making processes. Local explanations for a machine learning model are important for people to interpret its output. Techniques for explaining an ML model often involve a simpler surrogate model that yields interpretable information, such as feature importance scores. However, these techniques suffer from an inherent fidelity-interpretability trade-off due to their use of a simpler model for generating explanations.

Highly interpretable explanations may end up approximating too much and be inconsistent with the original ML model (low fidelity), while high-fidelity explanations may be as complex as the original ML model and thus less interpretable. To solve this dilemma, counterfactual explanations have been introduced. It maintains consistency with the original machine learning model, offering arguably interpretable insights. Counterfactual explanations reveal the minimal changes required in the original input features to alter the model's prediction, thus providing understanding without sacrificing fidelity or complexity.

This study aims to leverage explainable AI to enhance the explainability and usability of the deep learning-based traffic forecast model. Specifically, the goal is to elucidate the relationships between various input features and their corresponding output predictions. The research aims of this study are summarized by the following overriding research questions:

- What is the impact of input variables on deep learning-based traffic forecast?
- How can we modify the input variables to achieve the desired prediction for various scenarios?

This thesis present a comprehensive framework that utilizes counterfactual explanations for traffic forecasting and provides actionable insights through the proposed scenario-driven counterfactual explanations.

The study first establishes a deep learning model to predict traffic speed based on various context features. The Attribute Augmented Spatial Temporal Graph

Convolutional Neural Network model is built. Several context features were used for the prediction model, which includes static features and dynamic features. Static features are location-based, which means they generally only vary with regard to different road segments. The study includes nearby POI data for each road segment, speed limit data for each road segment, and lane configuration for each road segment. Particularly, the number of POIs includes the nearby gas station, charging station, parking lot, and restaurant. Dynamic features are time-based, indicating that they change over time. Calendar data, including day of the week, hour of the day, and weather condition data are included in this study. In this study, the traffic forecast model is built to predict the future traffic speed for each of the road segments on the traffic graph.

The core focus lies in the generation of counterfactual explanations to illuminate how alterations in these variables could affect predicted outcomes, thereby enhancing the model's transparency. This study considers the task of generating counterfactual explanations as a multi-objective optimization problem. To guide the search for counterfactuals, four key quantitative metrics were employed, which are validity, proximity, sparsity, and plausibility.

- Validity: A counterfactual is valid if it produces a predicted outcome closely approximating the target speed.
- Proximity: The ideal counterfactual should differ minimally from the original feature set, thereby ensuring that the changes suggested are modest and realistic.
- Sparsity: A counterfactual gains in feasibility when the number of altered features is minimized.
- Plausibility: For a counterfactual explanation to be considered plausible, it should be close to the nearest observed data points.

It is important to recognize that a counterfactual example, while perhaps optimal in feature space, may not be practically feasible due to real-world constraints. Therefore, users have the flexibility to specify constraints on feature manipulation, including range constraints and changeable variables. Range constraints define feasible ranges for each feature. For instance, a constraint might specify that "Speed limit on the road should be larger than 30 km/h." Changeable variables define which specific variables can be altered in the search for a counterfactual explanation.

The presence of multiple objectives in a problem gives rise to a set of optimal solutions, instead of a single optimal solution. Without additional information, it's hard to say one of the solutions is better than the other. To efficiently address this problem, the Non-dominated Sorting Genetic Algorithm II (NSGA-II) was introduced as a fast and elitist multi-objective evolutionary algorithm. In the context of this study, the performance of a counterfactual is represented by its vector of objective values, validity, proximity, sparsity, and plausibility, corresponding to the criteria outlined in the previous section. Lower objective values signify better counterfactuals.

After the generation and selection of counterfactual explanations, a comprehensive evaluation is essential to assess their performance and broader impact. It is crucial to verify that the counterfactual explanations achieve the desired speed improvement for the targeted road segment. Beyond the targeted segment, it is also necessary to ensure that localized changes don't negatively impact the speed of other road segments within the entire road network. This thesis involves the generation of counterfactual explanations for different spatial settings, under different time periods.

This study also delves into the practical implications of these features by integrating user-defined constraints to generate targeted counterfactual explanations. Two distinct methods for scenario constraints, directional and weighting constraints, are proposed to tailor counterfactual explanations to specific use cases. These tailored explanations benefit machine learning practitioners who aim to understand the model's learning mechanisms and domain experts who seek actionable insights for real-world applications. For directional constraints, users have the option to specify the direction, either increase or decrease, in which they would like specific features to move. For weighting constraints, users can assign weights to individual features to prioritize their importance during the counterfactual generation process.

Throughout the experiments, there are some significant patterns in the distribution of objectives. First, there appears to be a negative correlation between the validity loss and the proximity loss, which suggests that as counterfactual predictions approach the target speed more closely, the divergence of the generated counterfactual features from the original features increases. Second, validity loss and plausibility loss are negatively correlated. This implies that when the counterfactual predictions come closer to the target speed, they tend to deviate more from plausible, observed points in the feature space. Third, a positive correlation between proximity loss and plausibility loss. Generally, a greater proximity loss is accompanied by a larger plausibility loss. However, an interesting cluster of points exists in the bottom-right corner of this figure. These points show that there are counterfactual explanations that differ substantially from the original features but still maintain a close distance to overall observed data points.

Our findings underscore the integral relationship between traffic speed predictions and the spatial-temporal dynamics of road settings, revealing that varied patterns emerge across suburban and urban roads, as well as between weekdays and weekends.

Impact of contextual features on highway traffic: Counterfactual explanations generated for highway road segments failed to yield improvements in speed. This suggests that the static features investigated in this study, namely the number of POI, the number of lanes, and speed limits, do not substantially influence traffic patterns on highways within the scope of this road network. This outcome can be interpreted that highway speeds are primarily influenced by dynamic factors such as weather conditions and events, rather than by the static features examined here. In the case of nearby POIs, their presence appears to have negligible impact on highway speeds, as highways generally lack direct access to these facilities. Regarding the number of lanes and speed limits, isolated adjustments to these parameters on specific highway segments seem ineffective at altering overall speed. This is likely because highway traffic pattern is highly dependent on inflow conditions; altering the attributes of only a section of the highway would not significantly impact the overall traffic demand or the carrying capacity of the entire highway network. Therefore, it won't be able to enhance the speed in this situation.

Impact of contextual features on suburban road: When aiming to increase speeds on suburban road segments, counterfactual explanations suggest an increase in the number of POIs nearby. This is because the model associates road segments with a higher density of nearby POIs with lower levels of traffic congestion. The geographical location of a suburban road appears to significantly influence its traffic patterns. For instance, suburban roads adjacent to residential neighborhoods may experience lighter traffic but with more nearby POIs. In contrast, other suburban roads might be part of arterial routes and, despite having fewer nearby POIs, experience higher traffic volumes, leading to increased congestion or reduced speeds. Therefore, if the goal is to improve speeds on specific suburban roads, the model recommends increasing the number of nearby POIs. This alteration aims to make these road segments contextually similar to quieter, residential suburban roads, where lower traffic volumes and less congestion are observed. With regard to the number of lanes, the model does not suggest any significant modification pattern, except for the weekday afternoons, when the original traffic is the most congested and experiences the lowest speed. During these hours, the counterfactual explanations recommend reducing the number of lanes. Specifically, by reducing the number of lanes at the beginning of the road segment, less traffic would be able to enter the road segment, leading to more fluid traffic flow. Therefore, it can alleviate congestion and result in higher speeds. During weekends, the counterfactual explanations did not recommend alterations to the speed limit. This suggests that speed limits are not a significant factor affecting suburban road traffic during these times.

Impact of contextual features on urban road: In contrast to the suburban road, when targeting to increase speeds on urban road segments, counterfactual explanations suggest a decrease in the number of POIs nearby. This discrepancy between urban and suburban roads could be interpreted in two ways. Firstly, it reflects the inherently different traffic patterns between suburban and urban settings. Secondly, it's important to note that the initial number of POIs near the studied urban road segments is already quite high. Unlike in suburban areas where an increase in POIs seems to alleviate congestion, urban roads appear to benefit from a reduction in POIs, presumably because fewer attractions would lead to less traffic. Interestingly, an exception arises during weekday afternoons, where the counterfactual explanations do not recommend a reduction in the number of POIs for urban roads. This could be because, during these peak hours, the number of POIs does not have a significant influence on the speed of traffic on urban roads.

The use of deep learning models, coupled with Counterfactual Explanations, provides a powerful combination for uncovering complex relationships between variables. These relationships may be subtle or intricate enough for humans to notice, thus highlighting the novel capabilities of explainable AI and deep learning in data analysis. However, the efficacy of this approach is bound by certain limitations. Primarily, the model's predictive and interpretative strengths are dependent on the quality and diversity of the training data. Like all the data-driven methods, a dataset lacking in variability may limit the model's generalizability, resulting in recommendations that are not universally applicable. In the context of this study, a noteworthy limitation lies in the restricted exploration of different road graphs and a limited set of contextual features. This narrow scope may influence the robustness of the generated counterfactuals and their applicability to other scenarios. One potential avenue for mitigating these limitations involves the incorporation of domain-specific knowledge into the data-driven models. This can enhance both the generalizability and reliability of the model's recommendations. In light of this, scenario-driven counterfactual explanations are proposed in this study.

The experimental results, obtained by incorporating various scenario constraints into the counterfactual explanation generation process, are highly promising for several reasons. Firstly, all generated counterfactual explanations demonstrate reasonable validity and plausibility scores. This indicates that the method retains its efficacy even when additional constraints are applied, thereby affirming the feasibility and effectiveness of the approaches proposed in this study. Secondly, some constraints facilitate more efficient counterfactual generation. On the one hand, the collection of generated Counterfactual Explanations generally exhibits lower validity loss, implying enhanced performance in aligning the predicted speeds with target speeds. On the other hand, underweighting constraints, not only do the colors in the set of CFEs become more vibrant, but the scatter points also converge within a smaller area. This indicates increased efficiency after adding the scenario constraint, as the algorithm is more adept at identifying optimal counterfactuals within a constrained search space. Overall, the integration of user-defined prior knowledge into post-hoc explanations has proven to be invaluable. This not only addresses the initial research questions posed but also has profound implications for future work in the field of Explainable AI.

While this work demonstrates that scenario-driven counterfactual explanations offer significant benefits in the context studied, a key question that remains is how to ensure the practical utility and broader applicability of these methods in real-world settings. In this study, the quality of counterfactual explanations is solely evaluated based on objective metrics such as proximity and plausibility loss. We make the assumption that lower scores on these metrics indicate that implementing the counterfactual features in practice would be easier and more feasible. However, real-world applications often prove to be far more complex and challenging. To bridge this gap, future research should focus on collaborating with domain experts, such as urban planners, to gain insights into the actual challenges and constraints involved in modifying contextual settings.

In conclusion, the thesis introduces a comprehensive framework that advances the use of counterfactual explanations in spatio-temporal prediction tasks, effectively bridging the gap between theoretical understanding of models and their practical implications for actionable insights. In this study, a deep learning-based traffic forecast model was trained at first, using the state-of-the-art architecture, attribute augmented spatiotemporal graph convolutional networks. Subsequently, we generated diverse sets of counterfactual explanations by targeting various spatial and temporal settings.

On the one hand, by suggesting minimal alterations to input features, counterfactual explanations enhance our understanding of the model's behavior and elucidate the role of various contextual variables in deep learning-based traffic forecasting. This provides invaluable insights for AI practitioners, aiding in a deeper comprehension of what the model has learned from the data. More specifically, by examining a variety of spatial settings—such as suburban roads, urban roads, and highways, as well as different time slots, this study reveals that the impact of static contextual features on traffic speed is influenced by distinct spatial and temporal conditions.

On the other hand, this study advances the field by introducing scenario-driven counterfactual explanations, which offer domain experts like urban planners actionable and validated recommendations tailored to specific scenarios. By integrating user-defined constraints into our framework, we can provide nuanced insights that are directly applicable to a range of real-world conditions. Specifically, it introduces two methods for incorporating these scenario constraints: directional and weighting constraints. Both approaches effectively align the generated counterfactual explanations with users' prior knowledge and expectations, thereby making the search for optimal solutions more efficient. Importantly, we observed that some scenarios, particularly those incorporating weighting constraints, expedited the generation process and yielded more precise and useful CFEs. This is manifested through a more focused distribution of CFEs, indicating a clearer pathway for the algorithm to identify optimal counterfactual conditions.

The results indicate that counterfactual explanations can be useful in understanding the underlying patterns affecting traffic speed, showing potential for future applications in spatial-temporal predictive tasks. However, the study also reveals limitations concerning the model's geographical and feature-specific generalizability, suggesting avenues for future research.