Review8: Effective and General Evaluation for Instruction Conditioned Navigation using Dynamic Time Warping

Vivswan Shitole

CS539 Embodied AI

1 Summary

This paper introduces new metrics for evaluation of instruction conditioned navigation based on a classical approach for measuring similarity between two time series, known as Dynamic Time Warping (DTW). The paper defines a normalized Dynamic Time Warping (nDTW) metric that softly penalizes deviations from reference path, is sensitive to the order of nodes composing each path, is suited for both continuous and graph-based evaluations and can be efficiently calculated using dynamic programming. The paper further defines SDTW, which constrains nDTW to only successful paths. The authors motivate the metric by discussing the shortcomings of existing metrics for instruction conditioned navigation tasks: Path Length (PL), Navigation Error (NE), Oracle Navigation Error (ONE), Success Rate (SR), Oracle Success Rate (OSR) don't measure the fidelity between two paths and are very sensitive to the last node in the reference path. Success Weighted Edit Distance (SED) heavily penalizes a query path that is very close to the reference path but not exactly on it. Coverage Weighted by Length Score (CLS), Average Deviation (AD) and Max Deviation (MD) take account of both the paths in their totality but don't account for the order of the nodes.
DTW is computed by aligning elements from a reference and a query series, preserving the order in which elements appear in each of them, and forcing the initial and final elements of both the series to be aligned. Hence the warping is defined as a set of pairs of nodes from both the series, with one node of a series can be paired with the node from the other series that is exactly opposite to it or an immediate neighbour of the exactly opposite node. This warping scheme induces a monotonicity property, ensuring the warping preserves ordering. The optimal warping between two series is one which minimizes the cumulative cost of warping. The optimal warping can be calculated in quadratic space and time complexities using dynamic programming since the solution for a n-length series can be derived from the solution for (n-1) length series and three comparison operations. FastDTW approximates DTW in linear space and time complexitites using a multilevel approach that recursively refines the warping from coarser resolutions and projects them to higher resolutions. DTW is adopted to the navigation task by using the shortest distance between the nodes along the graph as the warping cost. To make the metric invariant to scale and density of nodes along the trajectories, DTW is normalized by the length of the reference path and the goal distance threshold. Finally a negative exponential of the normalized value is taken to yield the normalized variant nDTW as a bounded score. To account for success in the task, nDTW is multiplied by the success rate to yield Success weighted normalized Dynamic Time Warping (SDTW).
The introduced metrics are compared with other dominant metrics for the instruction conditioned navigation tasks: Room-to-Room (R2R) and Room-for-Room (R4R). Two types of evaluations are conducted: (1) Human evaluation where humans use their judgement to gauge the similarity of the query and the reference trajectories. 2525 annotated samples (505 sets of 5 query and reference pairs) were collected from 9 human raters, each set ranked according to the metrics under study. The pre-computed scores for the metrics are compared with the rankings obtained from humans by calculating the average Spearman's rank correlation. nDTW and SDTW yeild the highest average and lowest standard deviation in the rank correlation. (2) Comparison of goal-oriented agents with fidelity oriented agents that receive an additional reward proportional the gain in nDTW. Fidelity-oriented agents yield comparable or better performance as measured by previous metrics, and a strictly better performance when measured using nDTW and SDTW metrics.

2 Strengths

The authors have done a good job of motivating the introduced metrics by discussing the shortcomings of other prevalent metrics and listing the favourable properties of DTW for the instruction guided navigation task. The background

theory on DTW and its adaptations for the instruction guided navigation task are well specified. The authors have conducted thorough evaluations, comparing and reporting the average scores and standard deviations of all the metrics under study. The idea of extending the introduced metric as a RL reward signal is a good one.

3 Weaknesses

The paper seems to be built as a story around the improvements obtained by the introduced metrics over the instruction conditioned navigation tasks and not as a conceptually driven fundamental contribution. The improvements reported in the result tables are nominal in average when compared to the second best method. In table 3, SPL seems to differentiate between goal and fidelity orientations. The authors undermine this by stating this is only due to the fact that fidelity oriented agents produce paths that have more similar length to the reference path rather than fidelity to them. This reason is not justified clearly. The proof and pseudocode for FastDTW could have been provided as an appendix. An additional table could be provided justifying the lower computation time of DTW based metrics when compared to that of previous metrics, as claimed in the paper.

4 Reflections

In effect, this paper contributes a slightly better evaluation metric for evaluating performance in instruction conditioned navigation tasks. It is a broad contribution rather than a fundamental contribution as it provides for better evaluation results for most of the papers to come in the instruction conditioned navigation domain. Even better evaluation metrics may be possible by building on the introduced metrics.

5 Most Interesting Thought

I did not know that a new evaluation metric can be contributed as a paper in itself.