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

This paper introduces a new metric “normalized Dynamic Time Warping” to evaluate the path similarity of navigating agents in Visual and Language Navigation tasks. The paper shows the fundamental flaws in the existing metrics used for evaluating performance in such tasks, such as Path length (PL), navigation error (NE), Oracle Navigation error (ONE), Success Rate (SR), Oracle Success Rate (OSR), Average Deviations (AD), a Success weighted by Path length (PL) and so on.
The proposed metrics addresses these flaws by measuring similarity between the entirety of reference and the query trajectory (element wise distance function), rather than just being measured by the last nodes (goal position) in the reference path. It aligns elements in both trajectories while preserving the order in which elements appear in both the trajectories. Further, the metric can be used for both continuous and discrete actions in navigation space. The paper additionally defined an analogous metric to SPL, for evaluating the success of reaching the goal as Success weighted by the Normalized Dynamic Time Warping (SDTW)

The effectiveness of the proposed metric is evaluated firstly by comparing it with human rankings as to which metric is in close proximity to the one ranked by humans as the best query path, given a reference path. Further, the practical application of nDTW was evaluated by using this metric as a reward signal for agents in the Matterport 3D environment on R2R and R4R datasets. The experiments demonstrate that the metric as a reward signal for RL agent in VLN tasks outperforms the one with other metrics.

The strengths of the paper as as following. The paper clearly demonstrates the fundamental flaws in the existing approaches both intuitively and mathematically, which is the key strength of this paper. The proposed metric based on Dynamic Time Warping is quite interesting as it ensures that the performance is not just evaluated based on the sparse feedback of whether the task has been accomplished. The paper has clearly stated the desirable properties of the proposed metric which shows the wide application areas of the metric, such as Human-Robot interactive tasks, trajectory optimization in robotics tasks, etc. Further, the use of metric for both continuous and graph-based path evaluations, and the quadratic time complexity of the approach makes it quite feasible for real time applications. Additionally, the clarity of the approach is well demonstrated in all the figures.

However, there are few weaknesses in the paper. Firstly, the evaluations demonstrated to compare the metrics with human judgement is not fair, as the human judgements are likely to be biased towards path which is as close as possible to the reference trajectory, and it seems that the humans are not aware of the fact that the task is to reach the goal as instructions given to human raters does not say anything about goal, so it is likely that humans would have chosen the query paths which are close to overlapping path, while last node might deviate largely from the goal. Secondly, in the evaluation in VLN environments, it is not clear why the performance with SPL as reinforce signal for RL agent performs better than the one with SDTW in fidelity oriented tasks. The VLN tasks experiments are not well explained and lacks explanation of RL algorithm used, and why the reward with nDTW performs better than SDTW.

Overall, the idea of introducing a new metric for evaluating the performance of navigation tasks with Dynamic time warping is interesting. The metric clearly stands out of all the existing SOTA metric and has set a good benchmark for similar future works for VLN tasks. It would be interesting to extend this work for tasks where finding the cost of warping is non-trivial, for instance in case of dynamic environments where some areas in the environments are occluded in the query trajectories.