Visual Odometry (VO) means estimating an agent’s ego-motion from an image sequence captured with a camera attached to the agent.

Deep Mono VO is an end-to-end approach attacking this problem with deep learning for mono camera setups.

Our Contribution
First, we address drift in VO by introducing a global component to the loss function. Second, we develop the concept of stethoscopes to reduce overfitting.

Following the architecture of Wang et al., the network is made up of a CNN pre-trained for optical flow followed by an RNN.

Our Hypothesis: VO algorithms overfit by extracting location-specific information from the input. E.g.: “Passing the blue house, the agent has a speed of 30mph.”

With Stethoscopes, we introduce a mechanism for assessing the amount of location-specific information at any part of the network. They can be attached to any layer of the network and are trained on the task of inferring the absolute position. The inspiration for this technique was taken from.

- We treat the stethoscope attached to the RNN as an adversary and add a confusion loss to the main network.
- This penalises the network for extracting location-specific information.
- The results show less overfitting (reduced gap between training and test performance) and a minor increase in test performance.

Acknowledgements
We would like to thank Ronnie Clark, recent graduate from the Sensor Networks Group, for sharing his valuable insights and his generous help to implement the DeepVO algorithm from.

Drift - the accumulation of local errors - is an inherent problem of visual odometry. We found that adding a global component to the loss function which penalises drift improves the accuracy.

Only Local Loss: 3.8%
With global Loss: 2.6%

<table>
<thead>
<tr>
<th>Error (displacement over traveled distances)</th>
<th>only local loss</th>
<th>with global loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>VO trajectory</td>
<td>3.8%</td>
<td>2.6%</td>
</tr>
<tr>
<td>Predicted Ground Truth</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted Ground Truth</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ L_{VO-\text{local}} = \frac{1}{N} \sum_{k=1}^{N} (|| p_k - x_k ||^2 + \beta || \phi_k - \phi_k ||^2) \]

\[ L_{VO-\text{global}} = \frac{1}{N} \sum_{k=1}^{N} (|| \delta x_k - x_k ||^2 + || \delta \phi_k - \phi_k ||^2) \]

Where \( p \) and \( \phi \) denote the relative translation and rotation between frames and \( (x,y) \) is the agent’s absolute position.

References: