WORKING NOTE AC125

Alan Creak

1998 July 24

THE TURTLE TAUGHT

It is proposed to use the turtle in an experiment on reinforcement learning, using Paul Qualtrough's algorithm. Given a repertoire of simple moves and rotations, it will be rewarded for pushing an object forwards. Variations on the reward patterns will encourage it to learn different behaviours. Rate of learning and other phenomena will be investigated.

Paul Qualtrough[1] has worked out a scheme in which a reinforcement learning technique is used by a learning system in conjunction with a proposed memory structure in such a way as to lead the learner to develop an internal model of its environment in terms of its current and earlier perceptions.

The principle is to keep a list of observed sensor states and actions taken, with the rewards received. The system must associate actions with sensor states so that it can choose actions with high rewards when they are available. In some cases, a reward is not determined solely by the immediate state, but by one or two preceding states as well; if so, the system must build structures – essentially memories – with which it can resolve the ambiguities.

Paul has tested his ideas on simple simulated "robots"; it would be interesting to try them on a real one. We have a real one – the turtle. It moves, and is equipped with sensors, so can in principle be used as a mobile robot by a system resembling Paul's.

THE EXPERIMENT.

The turtle has a "wheelchair" drive with two driven wheels, each operated by a stepper motor; the two motors can be controlled independently. The turtle can therefore easily move forwards and backwards, and turn round on the spot by driving both wheels at the same speed in the same or opposite directions. More ambitious control of the motors can be used to produce paths which approximate to circular arcs of different radii. The drive is essentially continuous; compared with the size of the turtle the steps of motion corresponding to single motor steps are very small.

The turtle's sensory input comes from a contact sensor in the form of a ring round its body, which operates a microswitch – or, occasionally, two microswitches – when pushed far enough inwards. There are four switches set at 90° angles, so pushes from front, back, left, and right can be detected.

If the turtle is pushing an object which is not directly in front, the object is likely to slip sideways as the turtle moves forwards, so that eventually the front microswitch will be switched off and either the left or right switch will come on. By rewarding only forward pushing, it should be possible for the turtle to learn to turn in the direction of the object and to resume pushing it forwards, though in a new geographical direction.

The learning task which is to be accomplished by the turtle (a term which includes its controlling computer whenever the context so requires) is to observe associations between the rewards which it receives, its immediate sensations, and the recent history of actions and sensations, and to develop from these observations a strategy for maximising its rewards.

What actions should be made available to the learner ? The turtle can at any instant move either of its wheels forwards or backwards by one step of the associated stepper motor; it does not seem likely that these are suitable for the actions controllable by the learner, because they make hardly any difference to its state. In the first instance at least, we want a small set of actions which are likely to make some change to the state which is of significance in the task to be accomplished. There are at least two ways of constructing such a set of actions.

The first action set is derived by, essentially, guesswork : we inspect the problem, and decide what sorts of action are likely to be useful. A good guess is perhaps the combination of linear moves forwards and backwards, and clockwise and anticlockwise rotations. The lengths of the moves and the angles of the rotations can be set by experiment; sensible starting points might be 5 cm and 60°. (90°, though perhaps more obvious, is not particularly sensible, because it is likely to convert a position in which a pushed object has just slipped round the side in – say – a clockwise direction into a position where the object is just about to slip round the side in an anticlockwise direction.)

The second action set is rather different. It is composed of a few (probably two or three in the first place) speeds for the stepper motors, with any combination of left speed and right speed permitted. This results in motion which is in general a succession of circular arcs. The length of the arcs might be fixed, or an action might be deemed to continue until the sensor input changes. An interesting starting point would be to provide two speeds for each motor, equal but of opposite sign; with judicious choice of arc length, this would be equivalent to the other action set, and learning rates could be compared.

What rewards should be given ? Generally, good behaviour should be rewarded and bad behaviour penalised. Pushing the object forwards should clearly be rewarded; backing away from an object once found, or rotating so that an object in front becomes an object on one side should be penalised. In the first instance, it would probably be wise to reward all actions which contribute to the desired result, to encourage quick learning – so a rotation in the appropriate direction when the object has slipped round the side of the turtle can be rewarded, even though it is not strictly part of the pushing job. Once it is clear that the system works, a less pushy style of rewards could be tried.

THE GOAL.

First : construct the system, and investigate the simple pushing problem described above.

Second : study the same system, but with the added constraint that the turtle must push the object in a specified direction. The turtle has no sense of direction, but the controlling computer can keep track of it by calculating the angles turned, making the results available as a "sensory" input to the turtle. Only the four cardinal directions will be used.

Third : continue the experiment, but with a real direction, given in terms of angles (or, better, sines and cosines). This will require a more "intelligent" learner, able to optimise in terms of real numbers.

COMMENTS.

The turtle is very much slower than Paul's simulated machines, and in his experiments quite long runs were sometimes necessary. It is clearly necessary to control the problems set so that they are soluble in a reasonable time span. The first experiments should therefore be designed so that learning is as fast as possible; giving a reward at every step which explicitly encourages good behaviour and discourages deviation from the desired path should ensure that something is learnt fairly quickly. From this base, we can investigate the effect of varying the reward scheme to make the learning more difficult, but perhaps more "realistic".

PRACTICALITIES.

There are several respects in which the nature and behaviour of the turtle match the requirements of the experiment less well than those of Paul's simulated robots. This is not necessarily a bad thing, and one of the good features of this experiment is that it forces us to pay attention to real devices, not artificially simulated entities. Two examples follow; it is essentially certain that there will be others.

- The turtle will be moving in a finite space, and if it is to succeed in learning to push an object in a straight line it is likely to attempt to run out of the space. To simulate an infinite space, it will therefore be necessary to intervene in the experiment and move turtle and object from time to time. There is no particular difficulty in doing so, but provision for intervention must be built into the turtle control programme, including some way of observing the turtle's state so that it can be reestablished after the move before restarting the experiment.

  It would clearly be more rewarding if the turtle could learn for itself how to cope with walls as well as with its pushing task. It seems unlikely that this is practicable, though, partly because of the very limited speed of the turtle, but also because its sensory abilities are not sufficient to notice the wall. The sensation of pushing the object forwards doesn't change if the object is stopped by the wall; as the turtle can't tell that its wheels are slipping, it has no information which it can use to detect the change of state.

- In early stages of learning, and perhaps at later times too, the turtle is quite likely to wander away from the object. Given that the chances of finding the object by Brownian motion are slim, it is sensible to intervene if this happens. Judicious help can be given without messing up the learning algorithm by moving the object in front of the turtle after it has lost contact with the object for while. In so doing, though, the turtle must not receive any information which could be interpreted in the sense that wandering away from its current object will quickly result in a reward for finding another object.

In each case of intervention, it might be necessary to build some appropriate provision into the learning and control programme to ensure that the intervention either is invisible to the learner, or that the information received by the learner is realistic.

REFERENCES.

1 : Paul Qualtrough : PhD thesis, currently in preparation (1998).