How BECCA works

BECCA 0.6.0

Brandon Rohrer, December 2014

BECCA's long term goals are captured in its name: brain-emulating cognition and control architecture. This is a detailed description of the architecture and the algorithms that underlie it. It is a summary of the python code. If any discrepancies exist, the code is the authoritative source. This description omits some of the details of implementation and computational bookkeeping that aren't essential to understanding its principles of operation.

BECCA belongs to a class of machine learning algorithms called reinforcement learners. The distinguishing characteristic of a reinforcement learning agent is that it receives an explicit reward signal, which it tries to maximize. (Figure 1)

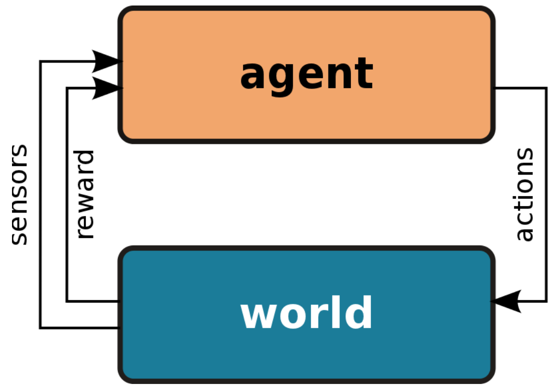

Figure 1: A reinforcement learning agent.

To be more precise, BECCA consists of both an unsupervised learning algorithm that extracts features from the data, and a reinforcement learning algorithm that learns which sequences of features tend to result in positive outcomes. (Figure 2)

BECCA is designed so that it doesn't know (or need to know) anything about the world it's operating in. This

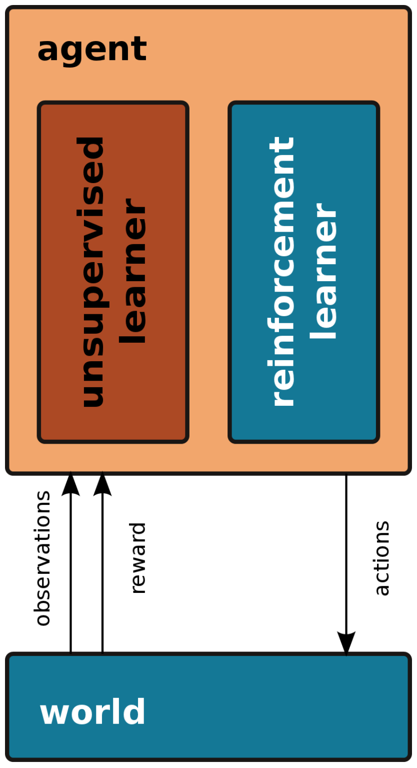

Figure 2: BECCA's agent contains both an unsupervised feature extractor and a reinforcement learner.

is so that it can be used as the learning portion of any embodied agent, a general purpose robot brain.

AGENT

The agent is, in essence, the brain. Everything else—the hardware, the sensors, and the rest of the universe—is included in the world. In fact, any system-specific pre-processing of sensor inputs and post-processing of actions are part of the world. The agent handles only sensor and reward inputs and produces action outputs. It does this in discrete time steps.

The agent contains one or more gearboxes in a hierarchical arrangement. A central collection of elements–the

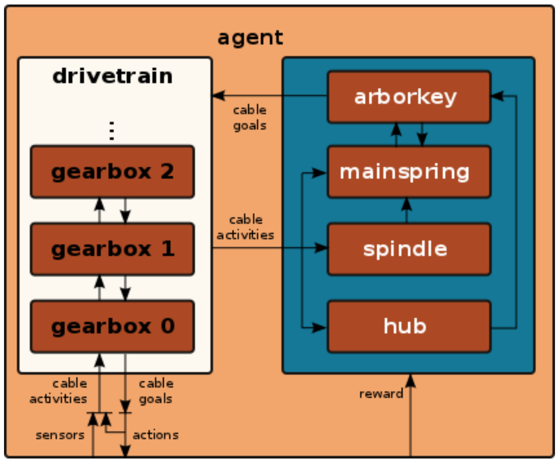

Figure 3: The agent consists of a hierarchy of gearboxes, arranged to form the drivetrain (the unsupervised learner), all of which are connected to a central complex of consisting of the hub, spindle, mainspring, and arborkey (the reinforcement learner).

hub, spindle, mainspring, and arborkey–is connected to them all. (Figure 3) The gearboxes build the sensory information into spatio-temporal features whose complexity and extent grows greater in each subsequent level. The first gearbox takes an array of cable activities as inputs at each time step, which are composed of the latest sensor activities, and a copy of the actions from the previous time step. Here, cables are a physical metaphor for channels or lines that carry a signal. Cable activities are the set of signal values on an array of cables at one point in time. Each gearbox takes a set of cable activities as inputs and produces a set of cable activities as outputs. They are cascaded, such that the cable activity outputs of one gearbox are also the inputs to the next.

At each time step, the cable activities are propagated upward through the entire drivetrain. Then, copies of the cable activity inputs to each gearbox are sent to the hub, spindle, and mainspring. The hub uses cable activities together with the reward to select a hub cable goal for one of the gearboxes. The spindle chooses one cable to attend to. The mainspring uses the attended cable and reward as a basis for constructing short term and long term memory and for learning delayed rewards. The arborkey evaluates potential goals from the hub with the help of the mainspring and periodically issues one of those goals back to the drivetrain. This centrally-selected goal is analogous to deliberate behavior in humans. The gearboxes then propagate this goal downward through the drivetrain, together with any internal goals (similar to reactions or auto-pilot behavior in humans) that they generate. A subset of the cable goals that emerge from the lowest gearbox correspond to an array of actions. These are passed back to the world. They are they only decisions or influence that the agent has on the world around it.

GEARBOX

Gearboxes are core structural elements of the architecture and naturally form a hierarchy, the drivetrain, in which each is supported by the one below. Each gearbox contains some cogs and a ziptie, a cable-clustering element. (Figure 4)

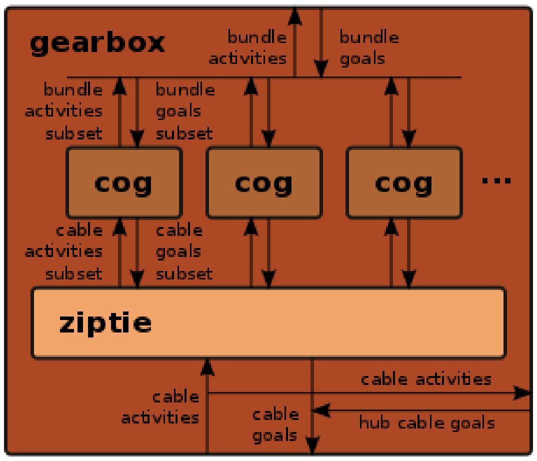

Figure 4: Each gearbox consists of a ziptie cable clusterer and a number of cogs, all operating in parallel.

The gearbox takes a set of cable activities as inputs. The ziptie organizes them into groups over time and assigns each group of cables to a single cog. Cogs are where cable activities are combined together across cables and over time to build spatio-temporal features. They take arrays of cable activities as inputs and combine those into bundles. Their outputs are the activities in each of those bundles. The bundle activities of all the cogs are combined together into a single array to form the output of the gearbox.

The cable activities are normalized such that they fall on an interval from zero (the lowest cable activity ever experienced on that cable) to one (the highest). This helps keep every bock insensitive to the range of cable activities it takes in, including sensor values from the world.

The number of gearboxes in the drivetrain grows as higher-level features are created. Once the number of bundles created collectively by all the cogs in gearbox n exceeds a fraction threshold of the gearbox's total capacity (.5), then gearbox $n + 1$ is created automatically. This allows the feature hierarchy to adaptively scale to represent arbitrary complexity in the data being observed.

COG

Like their mechanical counterparts, cogs are relatively simple devices whose true value is realized when they are used in parallel with a number of their fellows. Each cog consists of a daisychain and a ziptie. (Figure 5)

The daisychain identifies the relative prevalence of all possible cable activity sequences from one time step to the next, i.e. chains. The ziptie clusters commonly co-occurring chains into bundles. A bundle of chains is a pattern of signals that covers multiple cables and multiple time steps. Another way of expressing this is to say that is has both spatial and temporal extent or is a spatio-temporal pattern. The bundles that the cog generates are data-driven features. Taken in aggregate, all the bundles from all the cogs form a feature set that Becca can use to represent its relationship to its world and to choose its future actions.

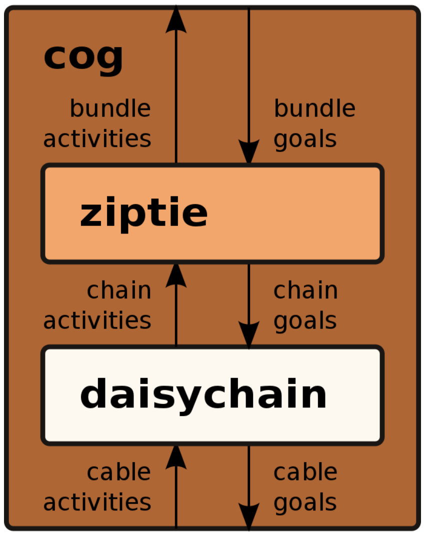

Figure 5: Each cog contains a daisychain and a ziptie.

ZIPTIE

Ziptie is a clustering algorithm used throughout Becca. In contrast to many unsupervised learning algorithms which cluster individual data points, ziptie clusters signals or time series of data points or dimensions in many-dimensional input stream. This is analogous to bundling groups of cables with zipties. Like all algorithms in Becca, it is incremental, performing its functions iteratively at each time step. Ziptie identifies cables whose activities tend to be high at the same time. Although at any given moment the cable activities may be very different, over time they will tend to be co-active.

The ziptie updates and maintains a map from cables to bundles. (Figure 6) This map dictates how cable activities are translated into bundle activities and how bundle goals are translated into cable goals.

The ziptie groups cables into bundles based on how often the cables' activity levels tend to be simultaneously high. Any cable can be a member of one bundle, many bundles, or no bundles. Membership is binary (all or nothing), but there is no reason the algorithm couldn't

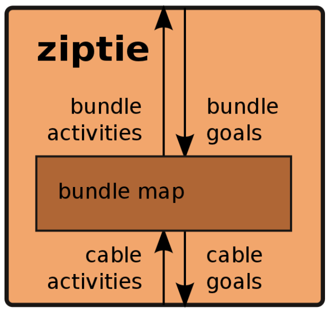

Figure 6: A ziptie clusters its cables together into bundles.

be extended to represent partial membership.

Bundle activity

Bundle activity is a weighted combination of its constituent cable activities. Some inhibition occurs between bundles that share cables. Initial bundle activities are calculated by taking the generalized mean of the cable activities with a negative exponent. The generalized mean, M_p, is given by:

$$M_p(x) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^p\right)^{\frac{1}{p}}$$

For $p = 1$, M_p is the arithmetic mean. For $\lim_{p\to 0}$ it is the geometric mean. For $\lim_{p\to\infty}$ it gives the maximum value. And for $\lim_{p\to-\infty}$ it gives the minimum. A negative exponent weights the lower cable activities more heavily.

Ziptie makes a first pass at each bundle's activity .

$$\begin{aligned}
a &= M_m(c)\\
a &: \text{initial bundle activity}\\
c &: \text{activities of cables in the bundle}\\
m &: \text{generalized mean exponent} = -4
\end{aligned}$$

If a cable contributes to multiple bundles, the activity it contributes to each will be somewhat inhibited.

$$\begin{aligned}
d &= c\left(\frac{a}{b}\right)^k\\
a &: \text{initial bundle activity}\\
b &: \text{highest initial bundle activity}\\
c &: \text{original cable activity}\\
d &: \text{inhibited cable activity}\\
k &: \text{inhibition exponent} = 6
\end{aligned}$$

The final bundle activity is calculated in the same way as the initial bundle activity, except that it uses the inhibited cable activities. A single cable ends up contributing different activation levels to the different bundles of which it's a member.

A cable can also have a residual non-bundle activity, signifying that it has signal energy that has not been applied to activating any bundle.

$$\begin{aligned}
f &= \max\left(0, c - \sum d\right)\\
c &: \text{original cable activity}\\
d &: \text{inhibited cable activity for each bundle}\\
f &: \text{non-bundle activity}
\end{aligned}$$

Bundle nucleation

Non-bundle cable activity is accumulated over time. If it accumulates to a high enough level, that signals the ziptie that a new bundle needs to be created.

$$g = g + (f(1-g) - cgp)q$$

f	:	non-bundle activity
g	:	nucleation energy
c	:	cable activity
p	:	energy decay rate $= .01$
q	:	nucleation energy rate $= .0001$

A cable's nucleation energy exceeds a threshold (.05), then it nucleates a new bundle, that is, it becomes the sole member of a new bundle.

Bundle agglomeration

Bundles accumulate agglomeration energy from cables whose non-bundle activity is correlated with their own.

$$h = fa$$

h	:	co-activity
f	:	non-bundle activity
a	:	bundle activity

Each cable's non-bundle activity is distributed to agglomeration energy with each bundle proportionally to their co-activities.

$$s = s + f\frac{a}{\sum a}(h(1-s) - csp)t$$

s	:	agglomeration energy between a cable and a bundle
f	:	non-bundle activity
a	:	bundle activity
$\sum a$	:	sum of all bundle activities
h	:	co-activity
c	:	cable activity
p	:	energy decay rate $= .01$
t	:	agglomeration energy rate $= .01$

When the agglomeration energy exceeds a threshold (.05), the cable is added to the bundle.

Cable goals

Bundle goals are translated back into cable goals as they propagate down the drivetrain of gearboxes. For each cable, the goals for the bundles it contributes to are summed in such a way that the magnitude of the sum is never greater than one.

$$u = S(v)$$

u	:	cable goal
v	:	bundle goals

The bounded sum function, S, never produces a total greater than one, given that its arguments are all less than one.

$$
\begin{aligned}
S(x) &= A\left(\sum B(x)\right) \\
A(x) &= 1 - \frac{1}{x+1} \\
B(x) &= \frac{x}{1-x} - 1 \\
A &: \text{maps the interval} [0,\infty) \\
&\quad \text{onto the interval} [0,1) \\
B &: \text{maps the interval} [0,1) \\
&\quad \text{onto the interval} [0,\infty)
\end{aligned}
$$

DAISYCHAIN

Dasiychain is an incremental algorithm that estimates the probability of one cable being active following another. It represents this as a conditional probability: given that one cable is active what is the expected activity of a second cable in the next time step. (Figure 7) High expected chain activities indicate sequences of cable activities that co-occur regularly. They identify temporal structure in the data.

Expected chain activities are similar to transition probabilities in Markov models. The difference is that in a Markov model, only one state can be occupied at each time step. This is analogous to just one cable being active. In a daisychain, many cables can be completely or partially active at once. As a result, transition probabilities can sum to much more than one.

Expected chain activity

A temporal sequence of one cable being active, followed by another, is a chain. The activity of a chain is given by the product of the two cable activities involved.

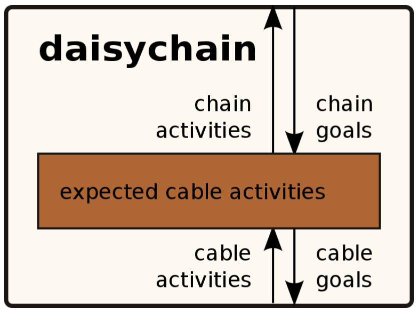

Figure 7: A daisychain identifies commonly occurring cable sequences, or chains.

$$
\begin{aligned}
c &= ab \\
c &: \text{chain activity} \\
a &: \text{cable activity} \\
b &: \text{another cable activity,} \\
&\quad \text{from the previous time step}
\end{aligned}
$$

A leaky accumulation of the activity on each cable and on each chain is also maintained.

$$
\begin{aligned}
d &= d + c - \frac{1}{df} \\
d &: \text{accumulated chain activity} \\
c &: \text{chain activity} \\
f &: \text{aging time constant} = 10^6
\end{aligned}
$$

$$
\begin{aligned}
g &= g + b - \frac{1}{gk} \\
g &: \text{accumulated cable activity} \\
b &: \text{previous cable activity} \\
k &: \text{chain update rate} = .1
\end{aligned}
$$

The expected chain activities are maintained and updated based on the current chain activities.

$$h = h + (c - h)b\left(\frac{1-k}{g} + k\right)$$

h	:	expected chain activities
c	:	chain activity
b	:	previous cable activity
k	:	chain update rate $= .1$
g	:	accumulated cable activity

Expected chain activity deviation

In addition, the expected *deviation* from the expected chain activities are maintained and updated based on the difference between the current and expected chain activities.

$$m = m + (|c - h| - m)b\left(\frac{1-k}{g} + k\right)$$

m	:	expected chain activity deviation
c	:	chain activity
h	:	expected chain activities
b	:	previous cable activity
k	:	chain update rate $= .1$
g	:	accumulated cable activity

Predictions

The temporal structure captured in the expected chain activities provide a basis for making short-term predictions.The reaction is the predicted next set of cable activities.

$$n = \frac{\sum ha}{\sum a}$$

n	:	predicted cable activity
h	:	all expected chain activities that include the cable activity being predicted
a	:	current cable activities

The most recently observed cable activities can be compared to those that would have been predicted from the previous cable activities to find surprising events.

$$p = \frac{\sum \frac{a|a-n|}{m}}{\sum \frac{a}{m}}$$

p	:	surprise
a	:	current cable activities
n	:	predicted cable activity, based on previous cable activity
m	:	expected chain activity deviation

Cable goals

As chain goals are propagated down through the daisychain, they are combined to form the cable goals. Each cable goal is a weighted, bounded sum of all the goals of the chains it belongs to.

$$s = S(qa + n)$$

s	:	cable goal
S	:	bounded sum operator
q	:	chain goals
a	:	current cable activities
n	:	predicted cable activity

HUB

The hub is where reinforcement learning is implemented and decisions are made. It takes in copies of all gearboxes' input cable activities, as well as the reward. It maintains an estimate of expected reward for every cable activity-cable goal sequence, or chain, similar to the daisychain's estimate of future cable activity for every cable activity chain within a gearbox.

The reward signal is normalized such that it falls on an interval from zero (the lowest reward ever experienced) to one (the highest). This way, Becca is insensitive to the range of reward values covered by the world. The chains that consistently result in higher reward are identified and exploited.

The hub gathers cable activities from the drivetrain. Using the current set of cable activities, it chooses a cable goal at each time step. It then associates changes in reward with each activity-goal pair. This is where the hub differs from daisychain. The transition chains in daisychain are from cable activity to next cable activity, rather than from cable activity to next cable goal. Like daisychain, the hub does this incrementally, adjusting the estimate by a small amount each time. It also creates a running sum of each chain's activity history, which slowly decays over time. This calculation is very similar to the one used in daisychain.

$$
\begin{aligned}
c &= ab \\
c &: \text{chain activity} \\
a &: \text{cable activity} \\
b &: \text{cable goal magnitude}
\end{aligned}
$$

Reward estimation

In order to associate transitions with delayed rewards, a reward trace is calculated. It is a summation of reward, decayed over time.

$$
\begin{aligned}
q &= \sum_{t=0}^{m} \frac{p_t}{t+1} \\
q &: \text{full reward trace} \\
t &: \text{time steps into the future} \\
m &: \text{trace length} = 10 \\
p &: \text{current reward} \\
p_t &: \text{reward } t \text{ time steps} \\
&\ \text{into the future}
\end{aligned}
$$

The expected reward trace associated with each chain is updated using the instantaneous and cumulative chain activities.

$$
\begin{aligned}
r &= r + (q - r)ck \\
r &: \text{expected reward trace} \\
q &: \text{observed reward trace} \\
c &: \text{chain activity} \\
k &: \text{reward learning rate} = .1
\end{aligned}
$$

Optimism-driven exploration

Becca is mildly optimistic. It assumes that any goal it hasn't pursued is likely to have some small reward. The lowest reward an agent has ever encountered is assigned a value of 0 and the highest is assigned a value of 1. Untried goals are assumed to have a value of .5. This motivates Becca to try goals that it hasn't tried before, until it learns that they aren't worth repeating.

Hub goal selection

The estimated reward value for all the chains is weighted by the square of current cable activities. The winning cable goal is the one with the highest weighted average value. The hub passes that goal to the arborkey.

$$
\begin{aligned}
w &= \text{argmax} \frac{\sum rc^2}{\sum c^2} \\
w &: \text{goal cable} \\
r &: \text{estimated reward value at this time step} \\
c &: \text{cable activity}
\end{aligned}
$$

Of course the final goal selection is performed by the arborkey. Whenever the arborkey selects a goal, it is passed back to the hub so that the estimated rewards can be updated appropriately

SPINDLE

The spindle selects a single cable activity to attend to. It picks the one with the highest activity, weighted by how long it has been since it was last attended.

$$
\begin{aligned}
s &= \left(1 - \frac{1}{t}\right) c \\
a &= \text{argmax } s \\
s &: \text{salience} \\
a &: \text{attended cable} \\
t &: \text{time steps since attended} \\
c &: \text{cable activity}
\end{aligned}
$$

Intuitively, very intensely active cables are given the strongest preference. Very recently attended cables are suppressed avoid neglecting them.

The spindle handles attention and salience. It has a huge impact on the behavior of the agent. As it stands, it's simplistic, only taking into account cable activity and how recently it was observed. There are a lot of other aspects of attention that experimental psychologists have identified. I hope that I'll get to incorporate these into future versions of BECCA.

MAINSPRING

The mainspring learns the expected reward associated with cable-goal chains in a manner very similar to that of the hub. However, it learns rewards that occur on a longer timescale. This allows the agent to learn complex and delayed relationships.

The mainspring also learns expected cable saliences based on previously attended cables.

Rather than learning on all currently active cables, the mainspring learns on recently attended cables. The set of recently attended cables is analogous to short term memory in human psychology. Recently attended cable are decayed hyperbolically. That means that their original salience when attended is multiplied by a $1/t$ term. The least energetic cable is dropped from the set at each time step.

$$
\begin{aligned}
d &= \frac{s}{t} \\
d &: \text{decayed salience} \\
s &: \text{original salience} \\
t &: \text{time steps since attended}
\end{aligned}
$$

The reward trace is calculated exactly as in the hub. The estimated reward for each chain is then estimated as in the hub.

$$\begin{aligned} c &= a^2b \\ c &: \text{chain activity} \\ a &: \text{decayed salience} \\ b &: \text{cable goal magnitude} \end{aligned}$$

$$\begin{aligned} r &= r + (q - r)ck \\ r &: \text{expected reward trace} \\ q &: \text{observed reward trace} \\ c &: \text{chain activity} \\ k &: \text{reward learning rate} = .1 \end{aligned}$$

$$\begin{aligned} s &= \frac{\sum rd}{\sum d} \\ s &: \text{overall estimated reward value} \\ r &: \text{estimated reward value by cable} \\ d &: \text{decayed salience} \end{aligned}$$

This is a way to evaluate potential goals over longer time scales than the hub can alone.

ARBORKEY

The function of the the arborkey is to make the final decision about which goals to issue and when. The hub passes its best goal candidate to the arborkey at each time step. The arborkey reevaluates it using the mainspring's longer term reward estimate. Then the arborkey notes the goal, its expected reward, and when it was put forward as a candidate.

A decayed expected reward is calculated for each goal in the candidate list. The expected reward is a function of the expected reward and how recently they were observed.

$$\begin{aligned} d &= \frac{r}{1 + wt} \\ d &: \text{decayed expected reward} \\ r &: \text{expected reward value} \\ w &: \text{reward decay rate} = .5 \\ t &: \text{time steps since the candidate was put forward} \end{aligned}$$

A restlessness term is added to the decayed expected rewards to calculate the goal values. If the highest goal value is higher than reward currently being experienced, it is chosen as the goal and executed. When this happens, the list of goal candidates is wiped clean and the process begins again.

$$\begin{aligned} v &= d + yt \\ v &: \text{goal value} \\ d &: \text{decayed expected reward} \\ y &: \text{action propensity} = .16 \\ t &: \text{time steps since a goal was executed} \end{aligned}$$

If the list of goal candidates grows to be more than 25 candidates long, the candidate with the lowest decayed expected reward is dropped from the list.

SUMMARY

One of the agent's greatest strengths is that it doesn't start off knowing or believing anything about the world it's connected to. There are only three arrows that connect the two: the sensors and reward flowing in and the actions flowing out. It was designed this way so that BECCA could be used for a brain in a robot of any type, physical or virtual, localized or distributed, commercial or experimental, rigid or flexible, mobile or fixed, walking, rolling, or articulated. For that matter it can be used to learn and control any system that has sensors and can take commands: stock trading, HVAC, retail shipping and receiving, or a self-driving automobile.