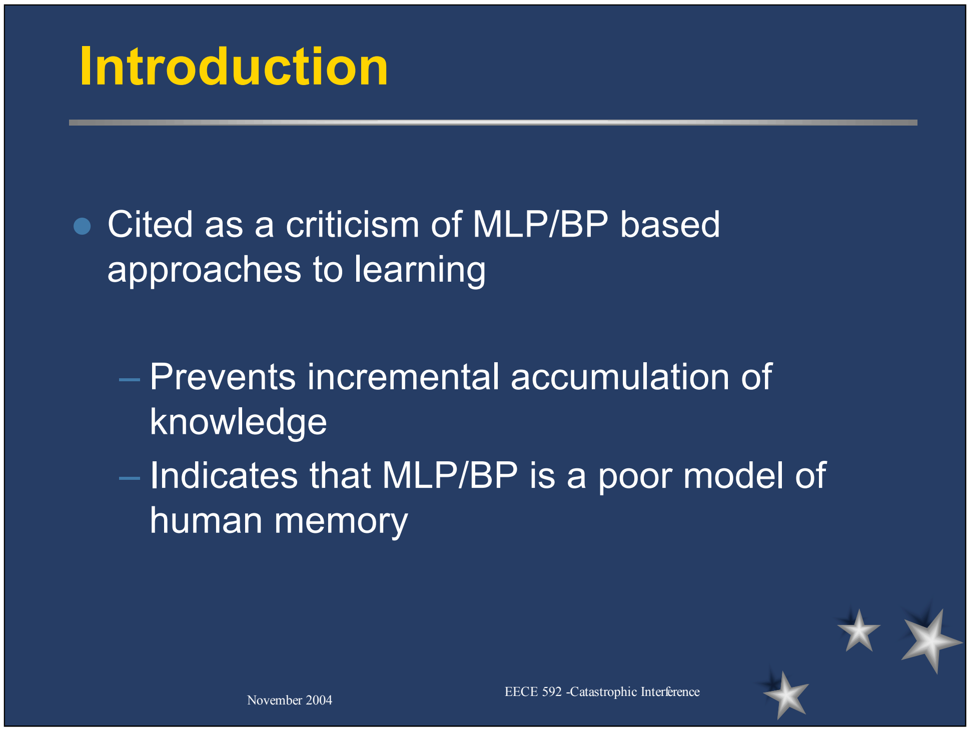
Catastrophic Interference (CI)

When new training disrupts existing memory



Introduction

- Cited as a criticism of MLP/BP based approaches to learning
  - Prevents incremental accumulation of knowledge
  - Indicates that MLP/BP is a poor model of human memory

November 2004

EECE 592 -Catastrophic Interference

Disruption of previous learning by later learning

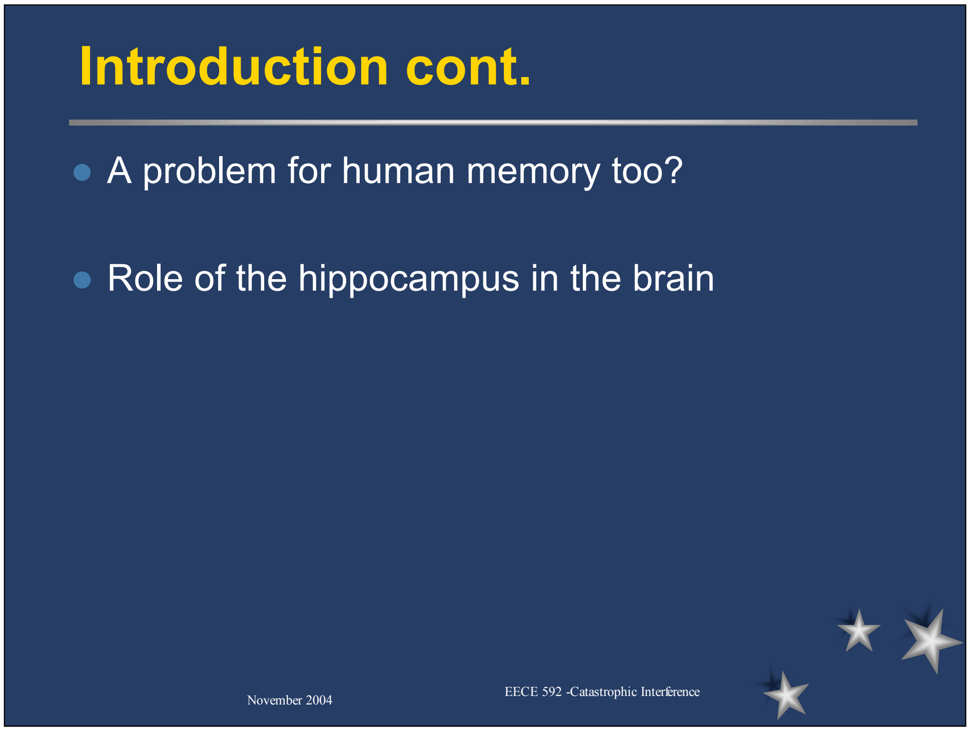
Catastrophic interference refers to the phenomenon that occurs when later training disrupts results of previous training and is characterized by the inability to incrementally learn sets of training patterns. CI is readily observed in studies of backpropagation. This phenomenon is also referred to as sequential learning and sometimes life long learning.

Criticism of MLPs and BP

CI is sighted as a major criticism of backpropagation like learning paradigms and is a serious set back for its acceptance as a plausible model of biological learning. After all, humans do learn by reinforcement too, many teachers will use examples as a way of describing a problem. However, human knowledge can accumulate incrementally, something that backpropagation is wholly incapable of.

Severe problem for BP

Given that the BP algorithm is slow and that its performance diminishes with the number of samples in the training set, it becomes clear that training on very large problems will be difficult. Without any ability to incrementally accumulate knowledge, we are restricted to compiling large, cumbersome training sets as means to defining a task.



Introduction cont.

- A problem for human memory too?
- Role of the hippocampus in the brain

November 2004

EECE 592 -Catastrophic Interference

Incremental accumulation of knowledge

To successfully accumulate knowledge in increments, memory for new patterns must not interfere or disrupt existing memory patterns. Obviously, the internal representations developed by BP are not able to assimilate new memories into existing weights without destroying previously trained patterns.

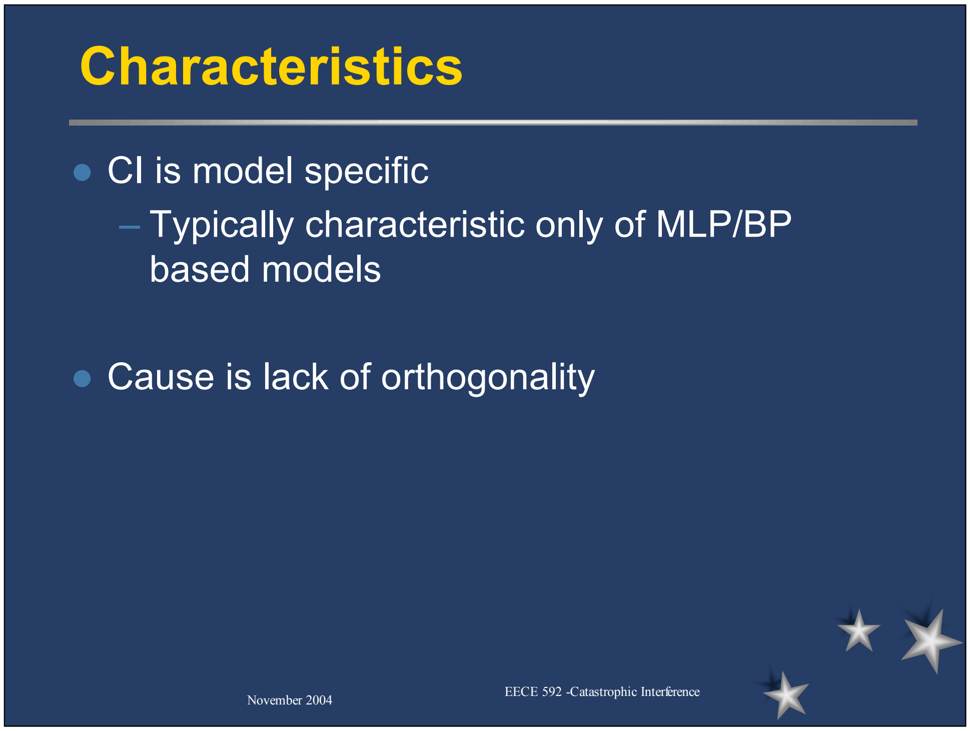
MLP as a model of human memory

The human brain, it seems, has solved this problem by storing new patterns in the hippocampus, not the neocortex. This minimizes interference with other memories. The new information is re-instated in certain situations, most notably during sleep.

This theory is consistent with physiological evidence relating to bilateral removal of the hippocampus, which is known to profoundly affect the ability to rapidly learn new memories.

see McClelland J., McNaugthon B., & O'Reilly R., (1994) "Why There are Complementary Learning Systems in the Hippocampus and Neocortex : Insights from the Success and Failures of Connectionist Models of Learning and Memory." CMU Tech Report. PDP.CNS.91.1

Clearly then, the inability for MLP models to effect sequential learning is a severe set back for such models as a basis for human learning.



Characteristics

- CI is model specific
  - Typically characteristic only of MLP/BP based models
- Cause is lack of orthogonality

November 2004

EECE 592 -Catastrophic Interference

Model specific

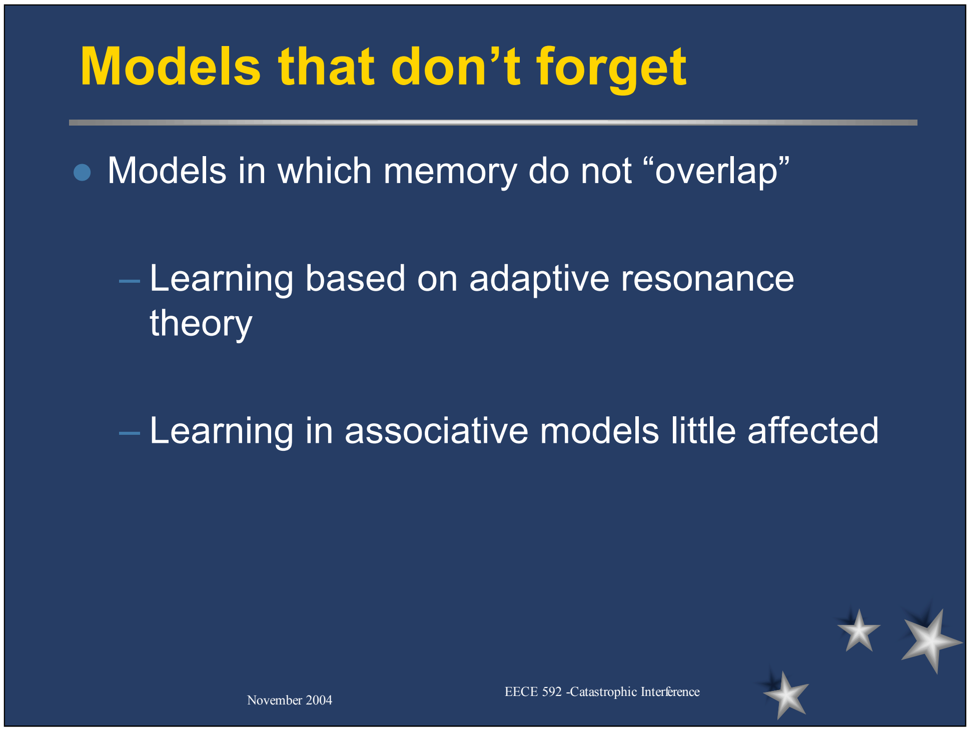
Catastrophic interference is characteristic of backpropagation like learning algorithms. The distributed nature of the representations involved is what gives an MLP the ability to generalize, however it's the very same reason that prevents sequential learning.

Learning algorithms that do not develop such distributed representations, would not be expected to be affected so much by CI. In constructive algorithms for example, units are added for each specific task learned and then the unit is "frozen" preserving its behaviour. (Cascade-Correlation is one such algorithm).

Orthogonality

Mutually orthogonal patterns are those for which the scalar or inner product is zero and for patterns for which this can be satisfied, there is said to be no cross-talk, that is, the weights encoding the patterns will not be shared.

Approaches based on the orthogonality of training patterns would be expected to show some ability to handle catastrophic learning. Whereas the representations generated by BP, bear no significance to the orthogonality of the trained samples. Associative nets on the other hand develop weights which reflect the amount of *overlap* between the patterns.



Models that don't forget

- Models in which memory do not "overlap"
  - Learning based on adaptive resonance theory
  - Learning in associative models little affected

November 2004

EECE 592 -Catastrophic Interference

"Overlap"

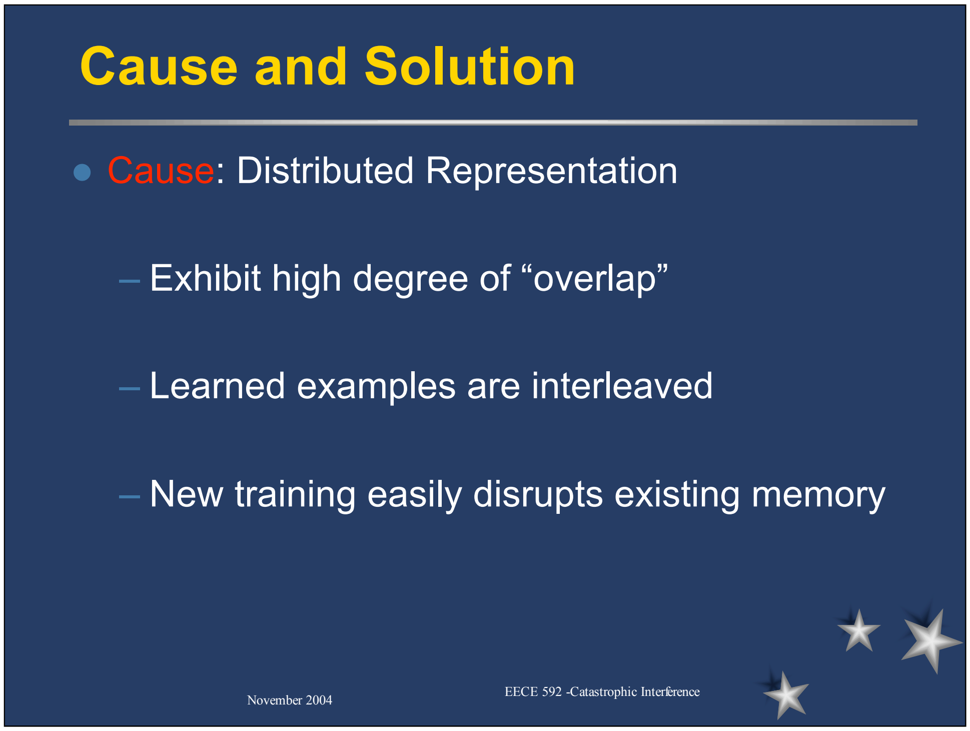
If the learned patterns in a memory overlap heavily, then the introduction of additional patterns will tend to disrupt previous memories, as is the case with distributed representations. If however, the amount of overlap expected between existing patterns and additional patterns can be reduced, some improvements should be observed. Two approaches are ART theory and autoassociators.

ART models

ART models are quite explicit in their aim to assimilate new knowledge in discrete memory chunks, although lack the generalization powers of distributed representations.

Associative models

Learning in autoassociators such as the Hopfield net is based on orthogonality between patterns. It's well known that the number of patterns that a net can reasonably store is related to the orthogonality and dimensionality of the input. However, when trying to store a random set of patterns, it is unlikely that they will be all mutually orthogonal. Theoretically, it could be possible to devise a to way incorporate additional patterns into a trained Hopfield net, assuming that they did not overlap with existing patterns. There doesn't however appear to be much research in this area.



Cause and Solution

- Cause: Distributed Representation
  - Exhibit high degree of “overlap”
  - Learned examples are interleaved
  - New training easily disrupts existing memory

November 2004

EECE 592 -Catastrophic Interference

CI is caused by “distributed representations”

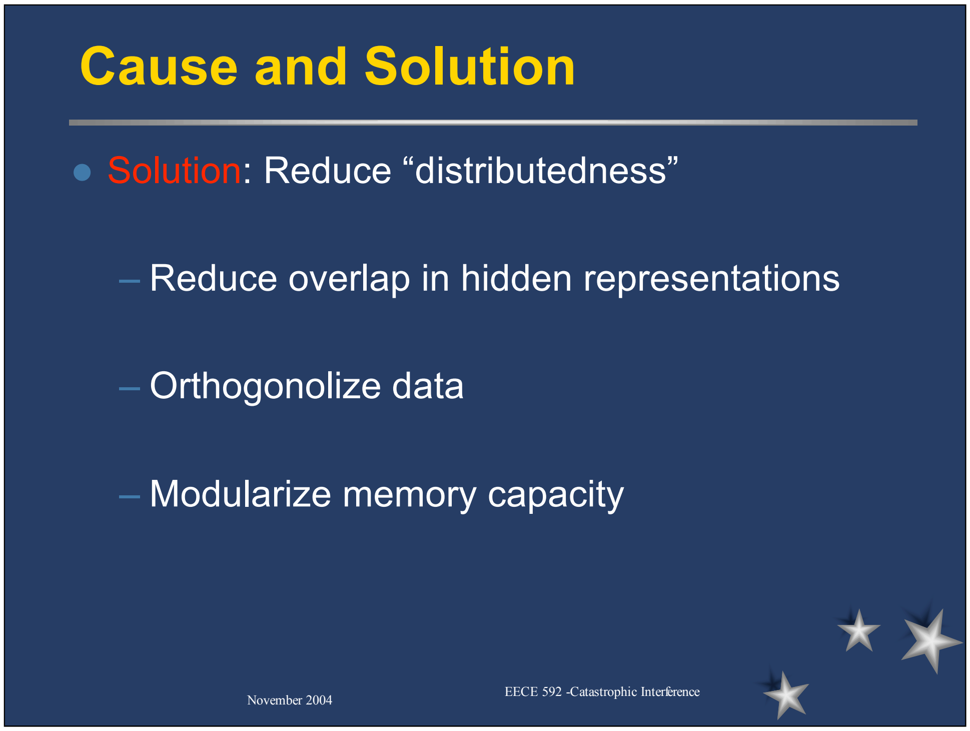
CI occurs as a result of attempting to develop distributed representations. Learning of new patterns needs to use those weights that participate in representing previously learned patterns. Much investigation to overcome CI is directed towards reducing the extent of distributedness. We can say that there is a tradeoff between distributedness and interference (as said earlier, no overlaps mean no CI). Note however, that in a biological system, distributed representations are essential to achieve fault tolerance!

Learned examples are interleaved

Distributed representations exhibit a high degree of overlap. Weights on a single unit will be involved in the memory of all patterns trained.

New examples disrupt existing memory

Learning of new patterns needs to use those weights that participate in representing previously learned patterns. Consequently those weights will be disrupted if new patterns are presented in a separate training session.



Cause and Solution

- Solution: Reduce “distributedness”
  - Reduce overlap in hidden representations
  - Orthogonolize data
  - Modularize memory capacity

November 2004

EECE 592 -Catastrophic Interference

Solution - reduce “distributedness”

Approaches to preventing CI are all based on reducing the distributedness of internal representations.

Reduce overlap in hidden representations

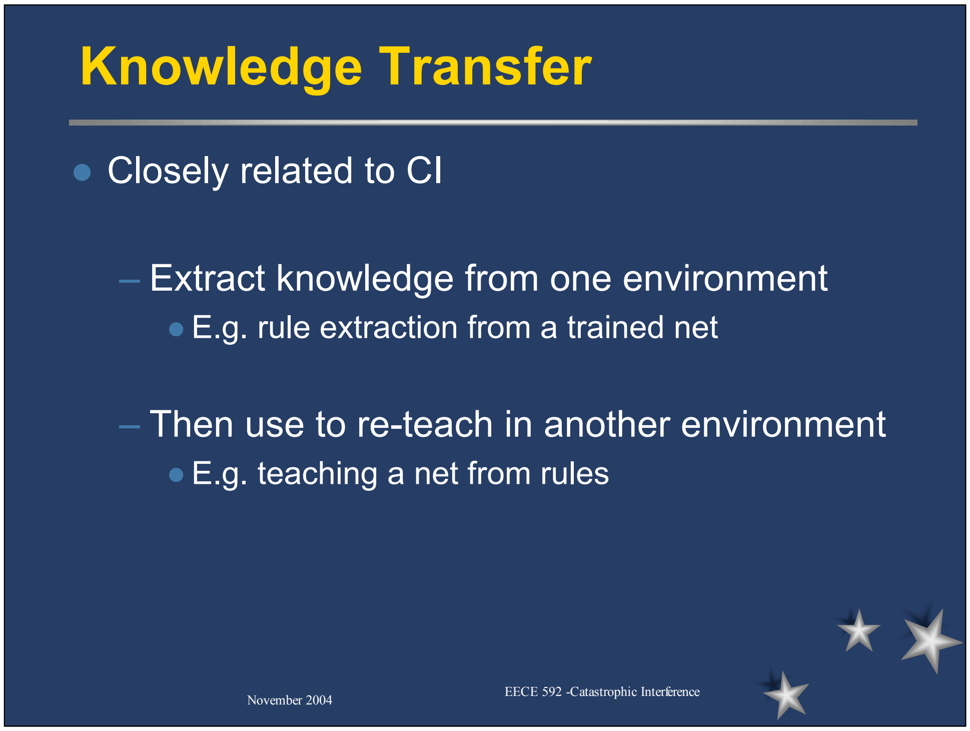
E.g. constructive algorithms. (e.g. Cascade-Correlation will be addressed later).

Orthogonalization

The reduction of cross-talk in associative memories. Useful only when the training model is able to take advantage of orthogonality as in the Hopfield model.

Modularization

Introduce some form of modularization so that different underlying functions are handled in different modules (reducing overlaps among differing tasks). This may not only solve the problem of CI, but also facilitate acquiring new knowledge (positive transfer). Furthermore, this idea is consistent with the general principle of functional localization in the brain.



Knowledge Transfer

- Closely related to CI
  - Extract knowledge from one environment
    - E.g. rule extraction from a trained net
  - Then use to re-teach in another environment
    - E.g. teaching a net from rules

November 2004

EECE 592 -Catastrophic Interference

Knowledge transfer

This is related to CI and refers to the ability to use knowledge learned from one task to perform other tasks. It is a form of generalization.

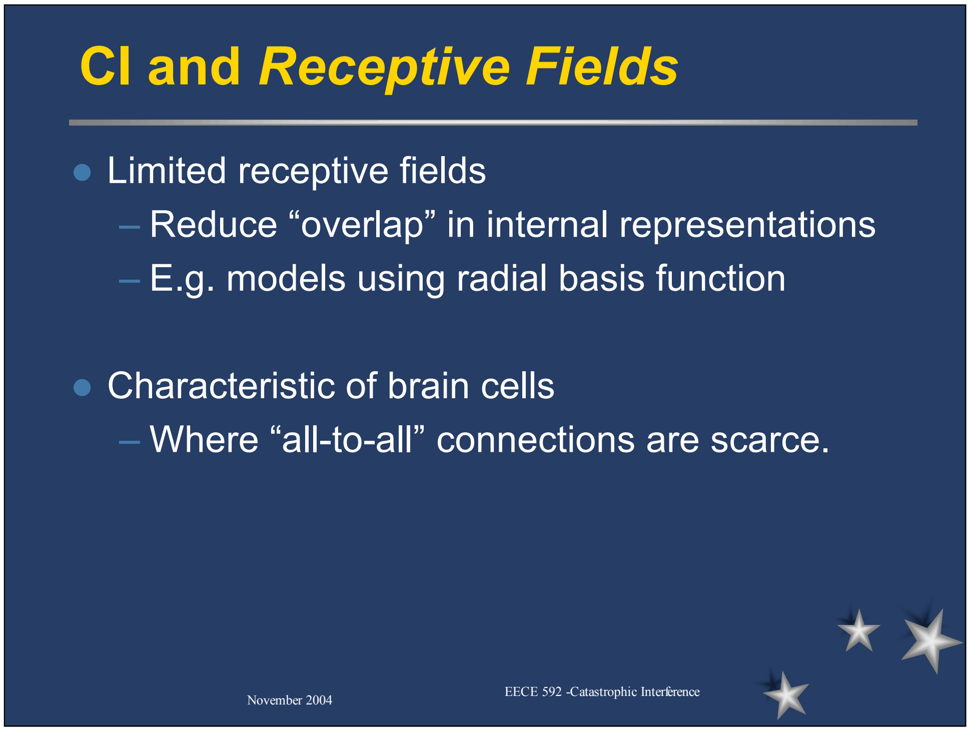
Humans somehow successfully manage to transfer big chunks of knowledge across learning tasks. If we face a new learning task, much of the "training data" which we use for generalization actually stems from other tasks, which we might have faced previously in our lifetime.

For example, once one has learned that the shape of the nose does matter and facial expressions do not matter for the identification of a person, one can transfer this knowledge to new faces and generalize much more accurately from less training examples.

It's not obvious how knowledge transfer can be effected - some method of representing the knowledge is first required.

Knowledge extraction

One approach has been suggested based on the extraction of rules from a trained net. The rules then become the vehicle used to transfer knowledge from one environment to another. This may not solve the interference problem, but surely handles the transfer problem to a certain extent. (It can also deal with the interference problem, if extracted rules are used to train the NN, interspersed with current data.)



CI and *Receptive Fields*

- Limited receptive fields
  - Reduce “overlap” in internal representations
  - E.g. models using radial basis function
- Characteristic of brain cells
  - Where “all-to-all” connections are scarce.

November 2004

EECE 592 -Catastrophic Interference

CI and Receptive Fields

It’s already been mentioned that one approach to reducing CI is to reduce the distributedness of internal representations. Radial basis models fall into this type.

Limited receptive fields

By limiting the receptive field of hidden units, learned representations are less likely to be distributed throughout the net. Instead, the representations will be based on a series of feature detectors trained to perform more specific tasks. This is the thinking behind radial basis models.

Characteristic of brain cells

Limited receptive fields are characteristic of brain cells where “all-to-all” connections are scarce if not at all.