Modular Neural Networks

Lecture 10

Modular Neural Networks

- **Different models** of neural networks combined into a **single system**. Each single network is made into a module that can be freely intermixed with modules of other types in that system.

Definition

- A neural network is said to be modular if the computation performed by the network can be decomposed into two or more modules (subsystems) that operate on distinct inputs without communicating with each other.
- The outputs of the modules are mediated by an integrating unit that is not permitted to feed information back to the modules. In particular, the integrating unit decided both (1) how the modules are combined to form the final output of the system, and (2) which modules should learn which training patterns.

3

Issues Leading to Modular Networks

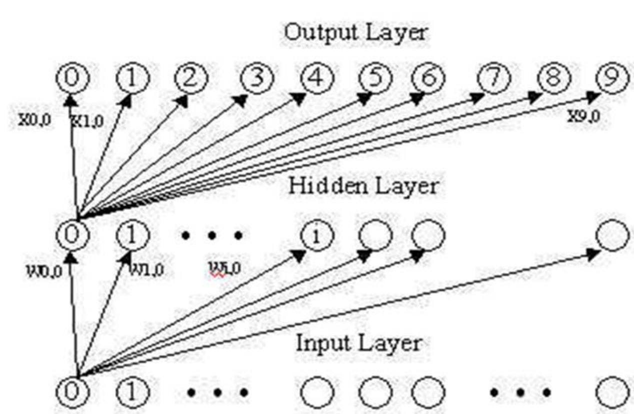
- Reducing Model Complexity
- Incorporating Knowledge
- Data Fusion and Prediction Averaging
- Combination of Techniques
- Learning Different Tasks Simultaneously
- Robustness and Incrementality

4

Problems in Neural Network Modeling:

- The selection of the appropriate number of hidden units
- Inefficiencies of Back-Propagation: Slow Learning, Moving Target problem

Problems in Neural Network Modeling:



Selection of hidden units

Problems in Neural Network Modeling:

Slow Learning:

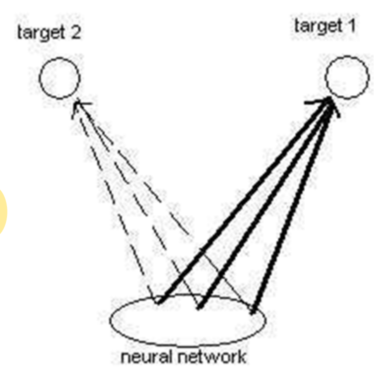
When training a network with back-propagation, all input weights into the hidden units must be re-adjusted to minimize the residual error.

7

Problems in Neural Network Modeling:

Moving Target Problem:

Each unit within the network is trying to evolve into a feature detector but input problems are changing constantly. This causes all hidden units to be in a chaotic state, and it takes a long time to settle down.



8

Problems in Neural Network Modeling:

Moving Target Problem:

The Herd Effect: Suppose we have a number of hidden units to solve **two tasks.** Each **unit can not communicate with one another**, so they must decide independently which task to tackle. If one task generates a larger error signal, then all units tend to solve this task and ignore the other. Once it has been solved, then all units moves to the second task, but the first problem will re-appear.

9

How do you solve the problem?

- Cascade Correlation
- Committee Machines

10

Cascade Correlation (CC): Characteristics

- Supervised learning algorithm - evaluated on its performance via external source
- A network that determines its own size and topology
- - starts with input/output layer
- - builds a minimal multi-layer network by creating its own hidden layer

Cascade Correlation (CC): Characteristics

- Recruits new units according to the residual approximation error
- Trains and adds hidden units one by one to tackle new tasks, hence "*Cascade*"
- The residual error "*Correlation*" between the new units and its output is maximized
- Input weights going into the new hidden unit become frozen (fixed)

Cascade Correlation (CC): Characteristics

- CC combines two ideas:
- Cascade architecture: hidden units added one at a time and is frozen
- Learning algorithm: trains and installs new hidden units

13

The Algorithm

1. CC starts with a minimal network consisting only of an input and an output layer. Both layers are fully connected.
2. Train all the connections ending at an output unit with a usual learning algorithm until the error of the net no longer decreases.
3. Generate the so-called candidate units. Every candidate unit is connected with all input units and with all existing hidden units. Between the pool of candidate units and the output units there are no weights.

14

The Algorithm

4. Try to maximize the correlation between the activation of the candidate units and the residual error of the net by training all the links leading to a candidate unit. Learning takes place with an ordinary learning algorithm. The training is stopped when the correlation scores no longer improves.
5. Choose the candidate unit with the maximum correlation, freeze its incoming weights and add it to the net.

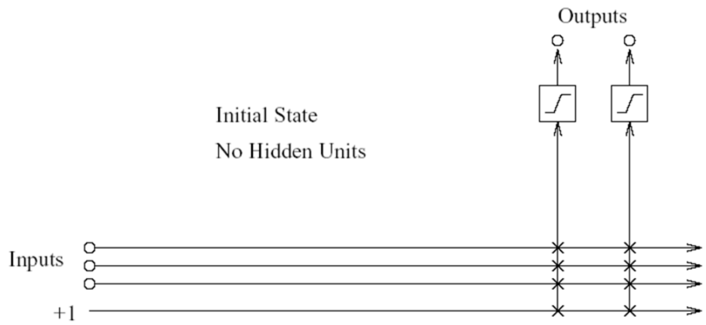
15

The Algorithm

6. To change the candidate unit into a hidden unit, generate links between the selected unit and all the output units. Since the weights leading to the new hidden unit are frozen, a new permanent feature detector is obtained. Loop back to step 2.
7. This algorithm is repeated until the overall error of the net falls below a given value

16

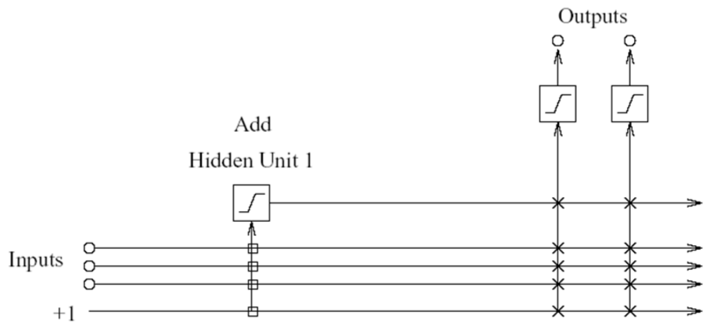
Cascade Correlation (CC): Diagram



The vertical lines sum all incoming activation. Boxed connections are frozen, X connections are trained repeatedly.

17

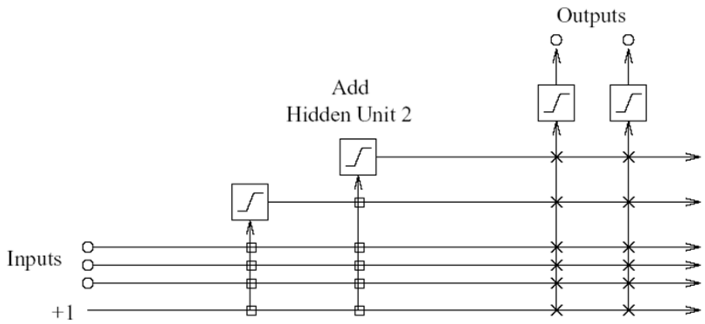
Cascade Correlation (CC): Diagram



The vertical lines sum all incoming activation. Boxed connections are frozen, X connections are trained repeatedly.

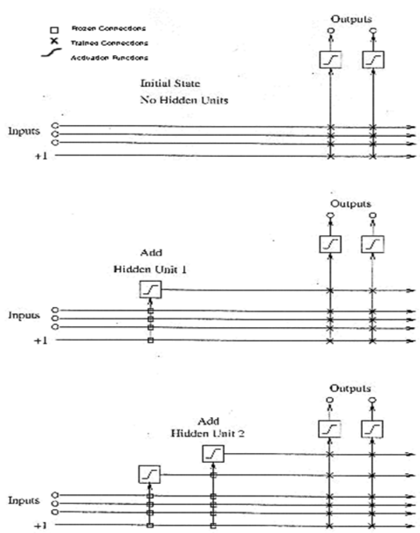
18

Cascade Correlation (CC): Diagram



The vertical lines sum all incoming activation. Boxed connections are frozen, X connections are trained repeatedly.

Cascade Correlation (CC): Diagram



Example: Speech Recognition

The difficulties with speech recognition:

- Deciphering different phonetics sounds
- Everyone has a different voice!

Example: Speech Recognition

A simple example:

Designing a network which can classify speech data into one of 10 different phonemes.

Example: Speech Recognition

- Train 10 hidden units separately, put them together and train the output unit one by one
- Adding new phoneme: train new hidden units for this phoneme and add to the network, and then retrain the output layer

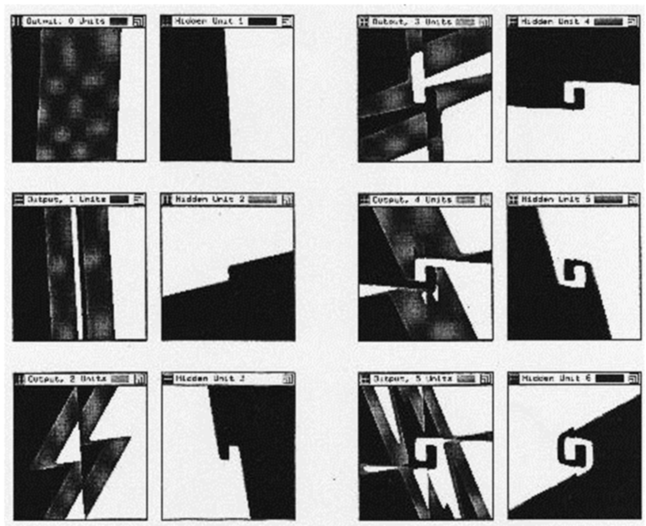
23

Example: Two Spirals Problem

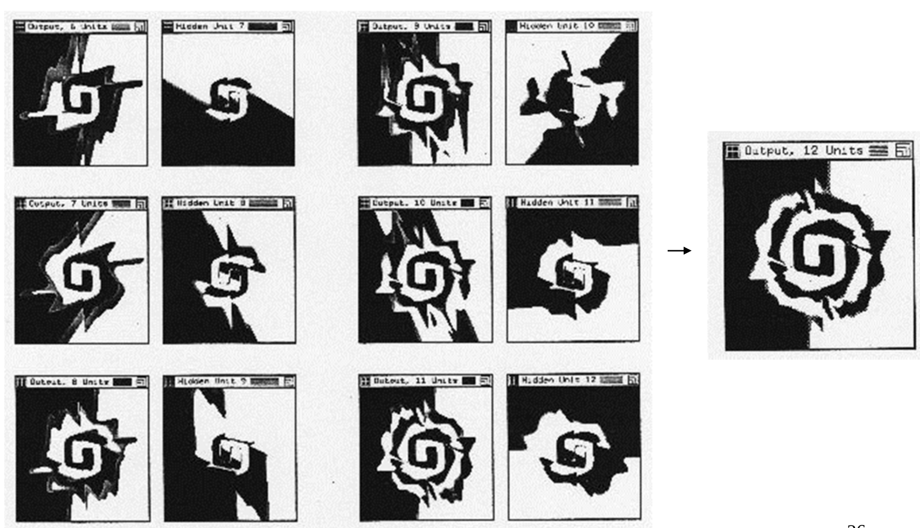
A primary benchmark for the back-propagation algorithms because it is an extremely hard problem to solve.

24

Example: Two Spirals Problem



Example: Two Spirals Problem



Cascade Correlation (CC)

Advantages:

- It learns at least 10 times faster than standard Back-propagation Algorithms.
- Incremental Learning
- Transparent
- Creates a structured network

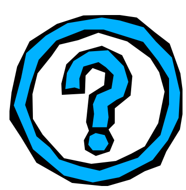
Disadvantages:

- Can lead to specialization of just the training sets.

27

References

- The Cascade Correlation Learning Architecture. Scott Fahlman and Christian Lebiere. - reports- **http://archive.adm.cs.cmu.edu/anon/1990/CMU-CS-90-100.ps**
- A Tutorial on Cascade-correlation. Thomas R. Shultz **http://www.psych.mcgill.ca/perpg/fac/shultz/cdp/Cascade-correlation_tutorials.htm**



28

Committee Machines

- Universal Approximators
- Static Structures
- Dynamic Structures

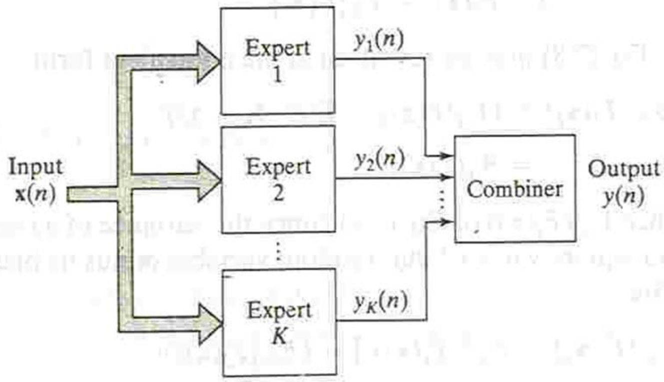
29

Static Structures

- The response pf several predictors (experts) are combined by means of a mechanism that does not involve the input signal, hence the designation “static”.
- **Ensemble Averaging:** Outputs of different predictors are linearly combined to produce an overall output
- **Boosting:** Weak algorithm is converted into one that achieves arbitrarily high accuracy.

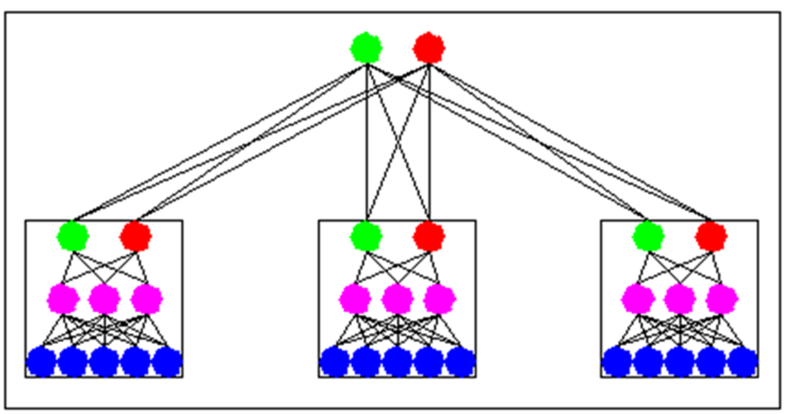
30

Ensemble Averaging



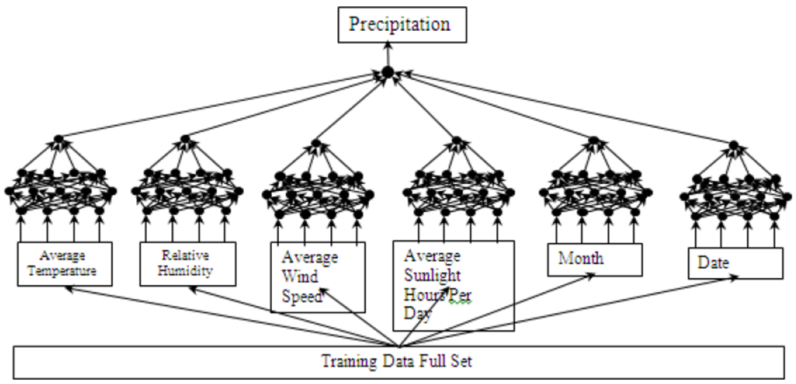
31

Example



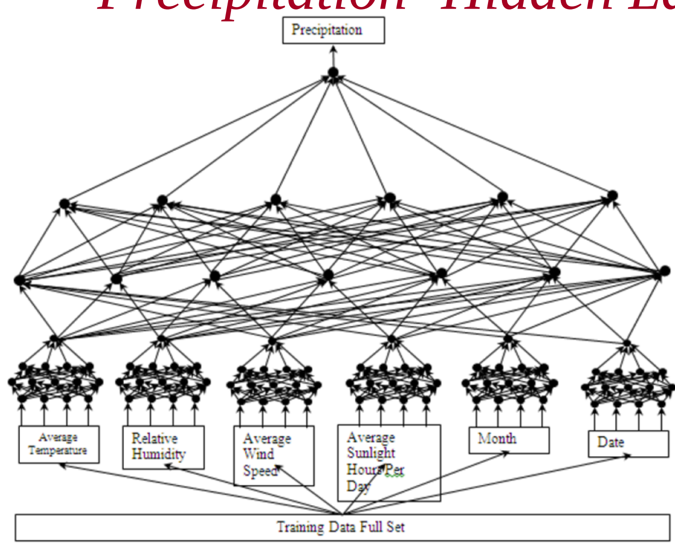
32

Example - Precipitation



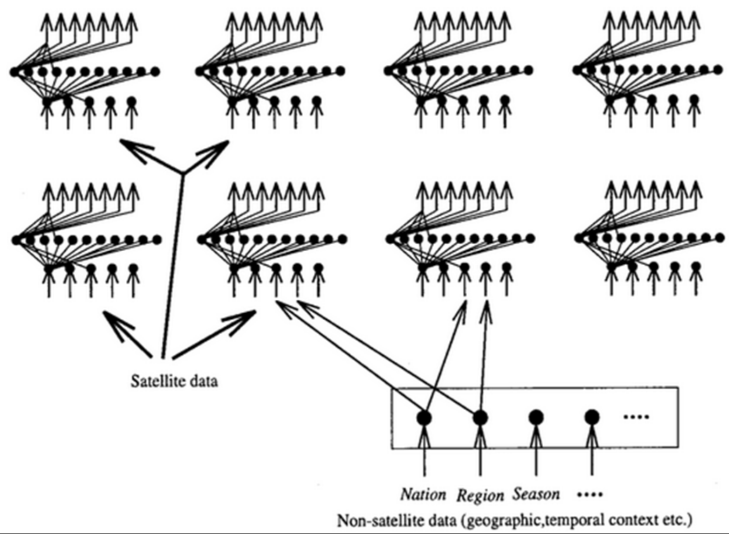
33

Precipitation- Hidden Layer



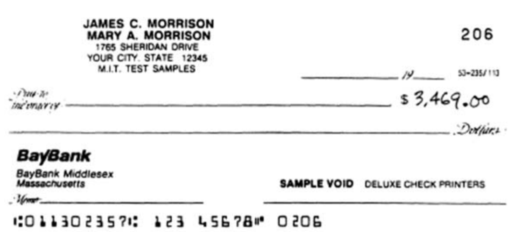
34

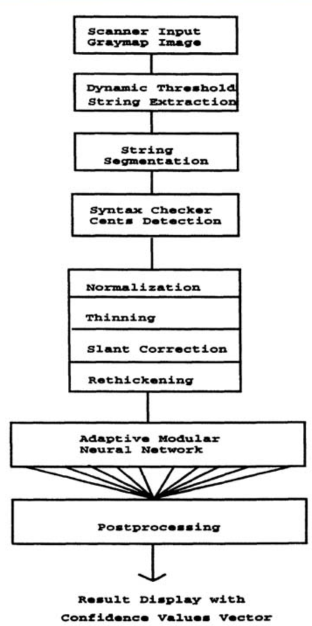
Remote Sensing Example



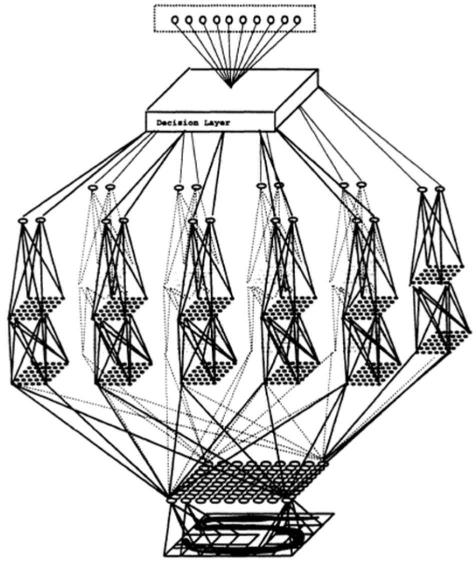
35

Example: Cheque Data

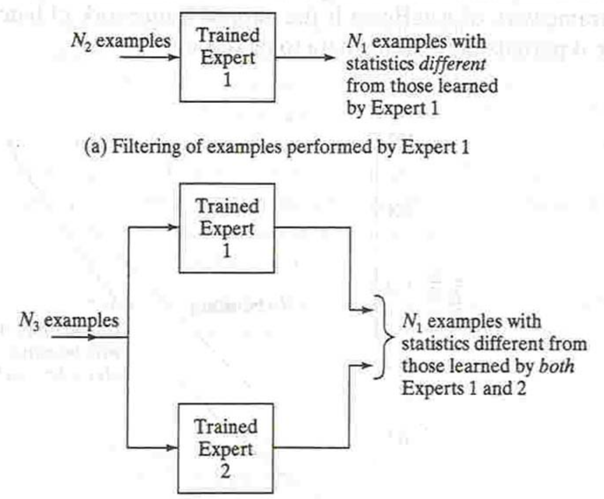




36


Decision Layer

Boosting

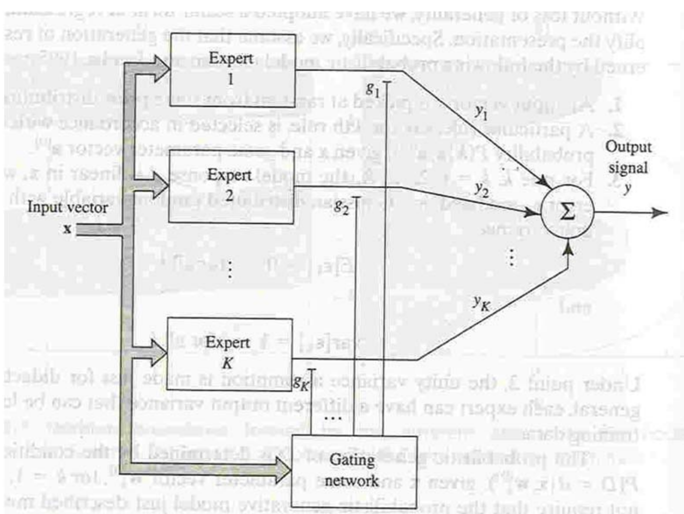

N_2 examples
Trained Expert 1
N_1 examples with statistics *different* from those learned by Expert 1
(a) Filtering of examples performed by Expert 1
Trained Expert 1
N_3 examples
N_1 examples with statistics different from those learned by *both* Experts 1 and 2
Trained Expert 2

Dynamic Structures

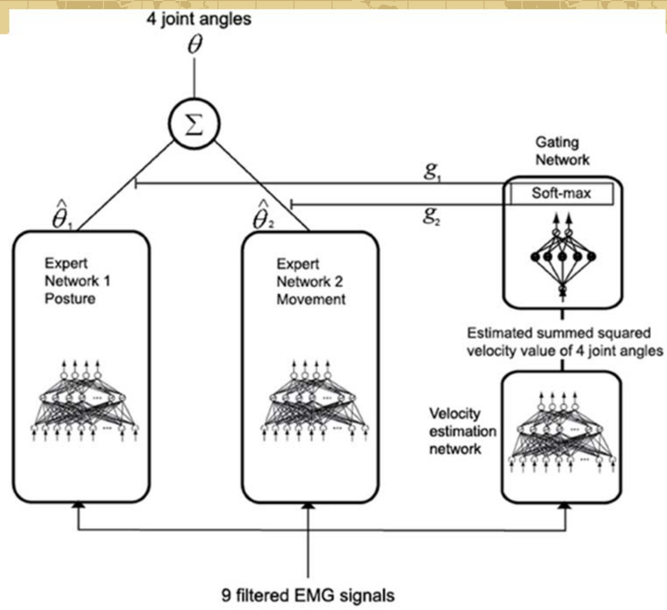
- The input signal is directly involved in actuating the mechanism that integrates the outputs of the individual experts into an overall output, hence the designation "dynamic".
- **Mixture of experts-** Individual response of experts are nonlinearly combined by means of single gating network
- **Hierarchical mixture of experts -** Individual response of experts are nonlinearly combined by means of several gating networks arranged in hierarchical fashion

39

Mixture of Experts

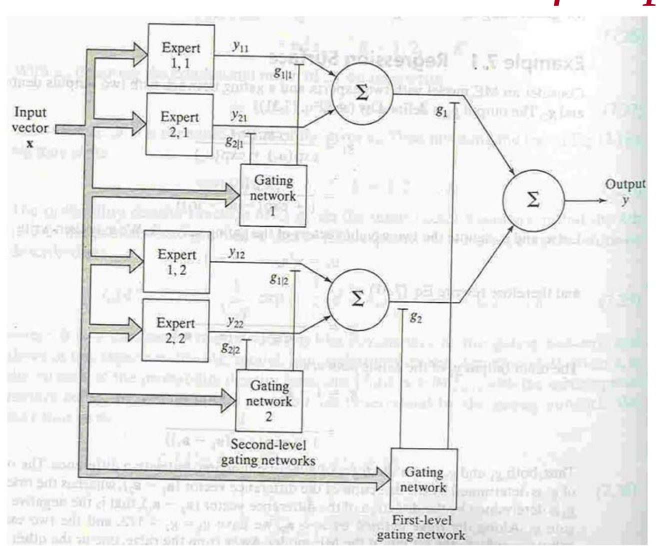


40



41

Hierarchical Mixture of Experts



42