Wake-sleep algorithm

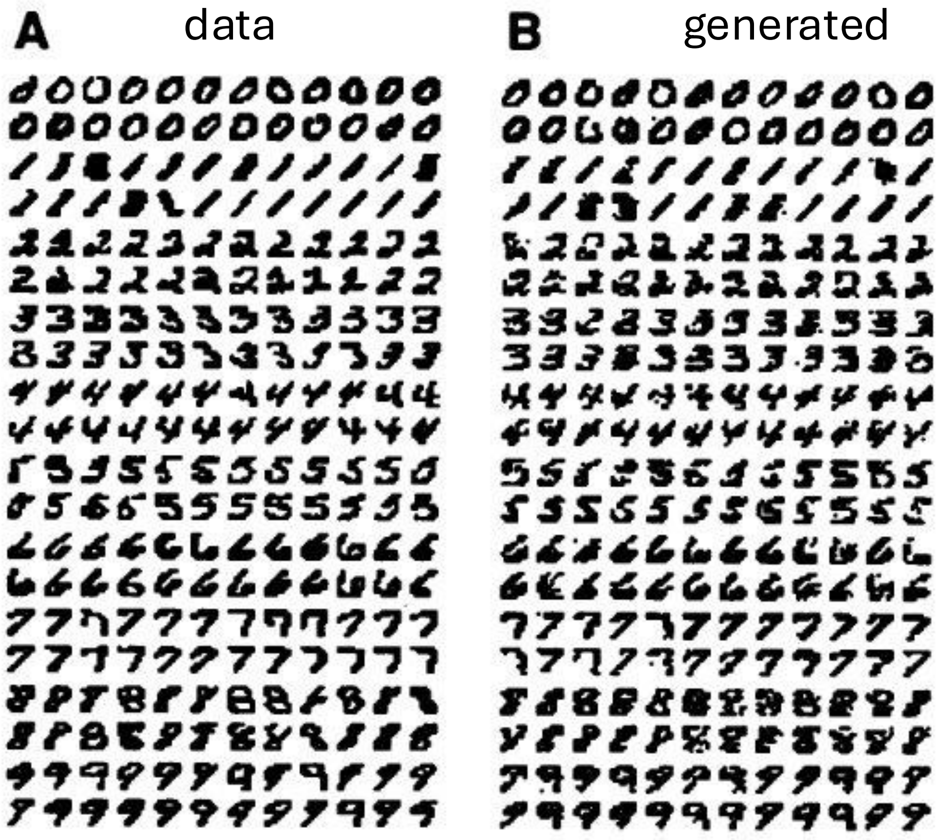
Hinton et al., 1995, Science

Objective

- Unsupervised learning of data

- Generative model

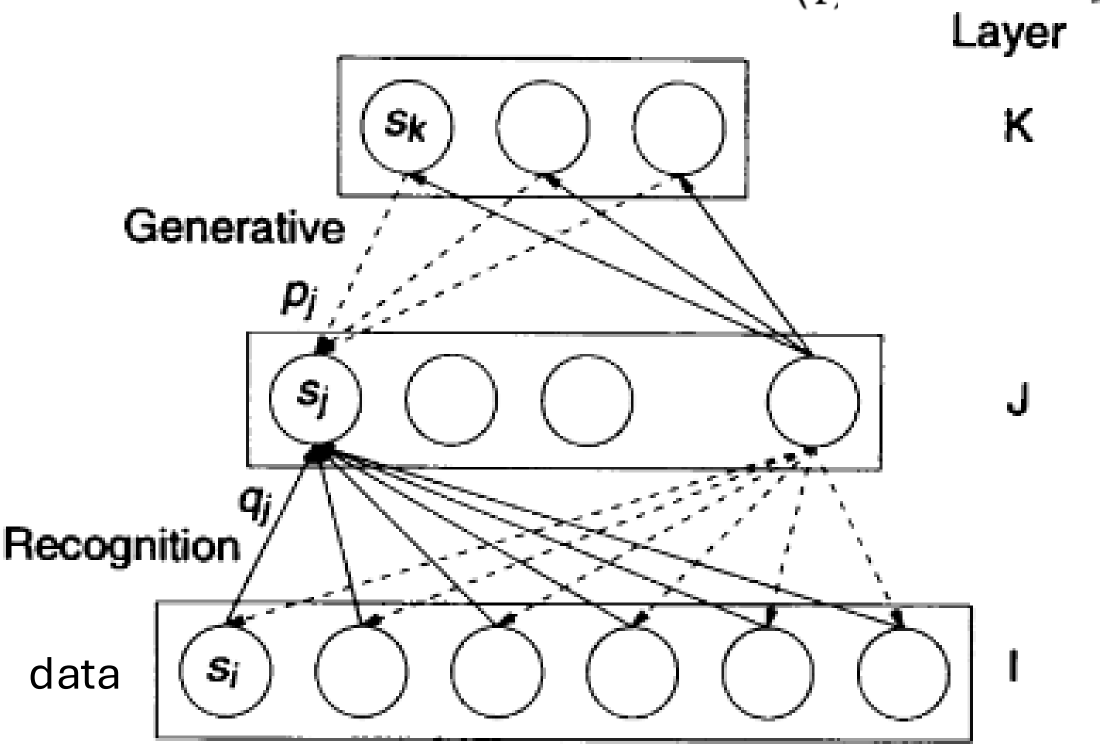
- Local updates (no back-prop)



ASK! (I might be wrong)
Hopping between paper & presentation

Core idea

$$\text{Prob}(s_v = 1) = \frac{1}{1 + \exp(-b_v - \sum_u s_u w_{uv})} \quad (1)$$



data

Sigmoid believe network (stochastic, binary states)

- Alternating update steps (wake-sleep)
  - Wake: use recognition, update generation
  - Sleep: use generation, update recognition
- Sample new data starting from layer K
- Optimize with information-theory
  - Minimize the information required to "transmit" data
    - Representation 'cost' (knowing generative weights)
    - Difference to real input

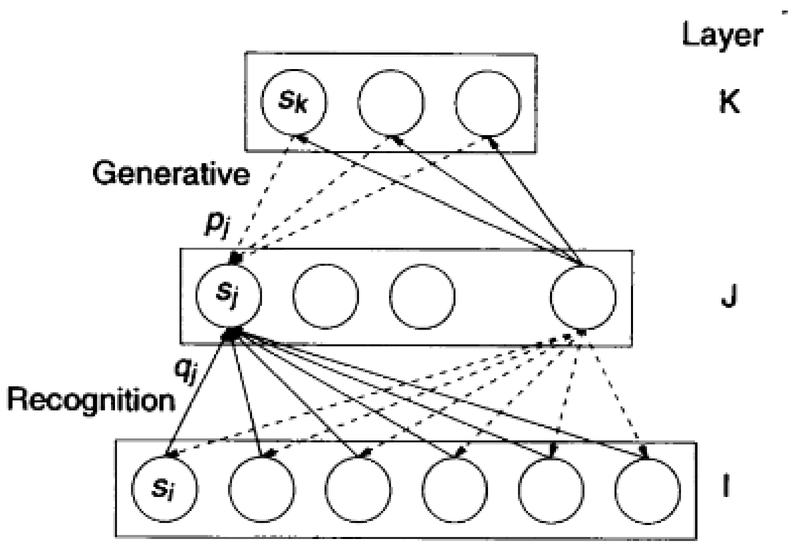
Wake

$$\text{Prob}(s_v = 1) = \frac{1}{1 + \exp(-b_v - \sum_u s_u w_{uv})} \qquad (1)$$

$$C(s_j^\alpha) = -s_j^\alpha \log p_j^\alpha - (1 - s_j^\alpha)\log(1 - p_j^\alpha) \qquad (2)$$

$$\begin{aligned} C(\alpha,d) &= C(\alpha) + C(d|\alpha) \\ &= \sum_{\ell\in L} \sum_{j\in\ell} C(s_j^\alpha) + \sum_i C(s_i^d|\alpha) \qquad (3) \end{aligned}$$

1. Infer latents (recognition)



2. Update generative weights

$$\Delta w_{kj} = \epsilon s_k^\alpha (s_j^\alpha - p_j^\alpha) \qquad (4)$$

Sleep

$$\cancel{\begin{aligned} \min_\alpha C(\alpha,d) &= C(\alpha) + C(d|\alpha) \\ &= \sum_{\ell \in L} \sum_{j \in \ell} C(s_j^\alpha) + \sum_i C(s_i^d|\alpha) \qquad (3) \end{aligned}}$$

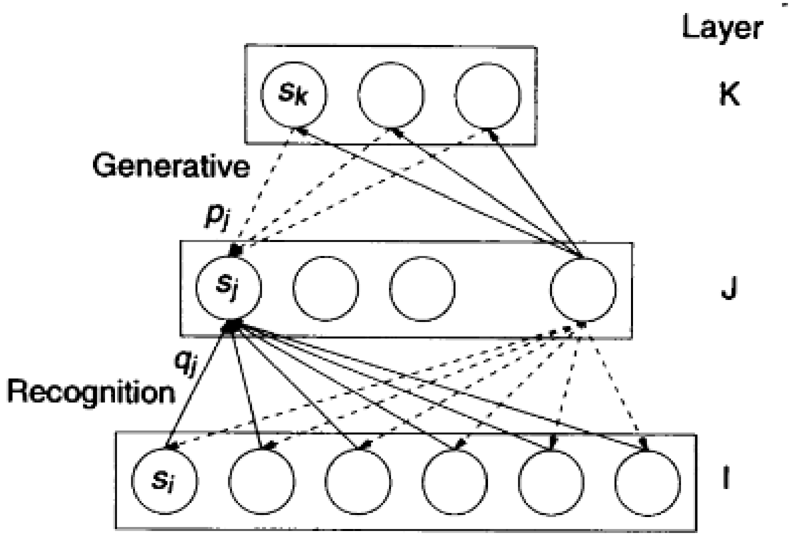
$$C(d) = \sum_\alpha Q(\alpha|d) C(\alpha,d) - \left[-\sum_\alpha Q(\alpha|d) \log Q(\alpha|d) \right] \qquad (5)$$

Entropy

of the system. As in physics, C(d) is minimized when the probabilities of the alternatives are MAGIC IN BETWEEN costs by the Boltzmann distribution (at a temperature of 1)

$$P(\alpha|d) = \frac{\exp[-C(\alpha,d)]}{\sum_\beta \exp[-C(\beta,d)]} \qquad (6)$$

1. Generate data (fantasize)



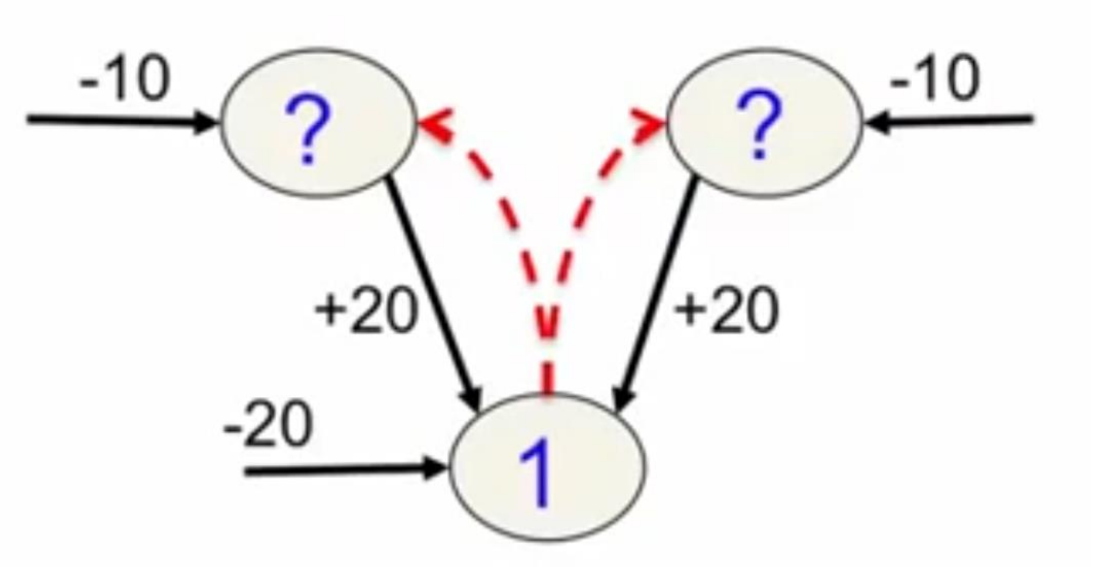
2. Update recognition weights

Optimize recognition weights across*all possible representations*

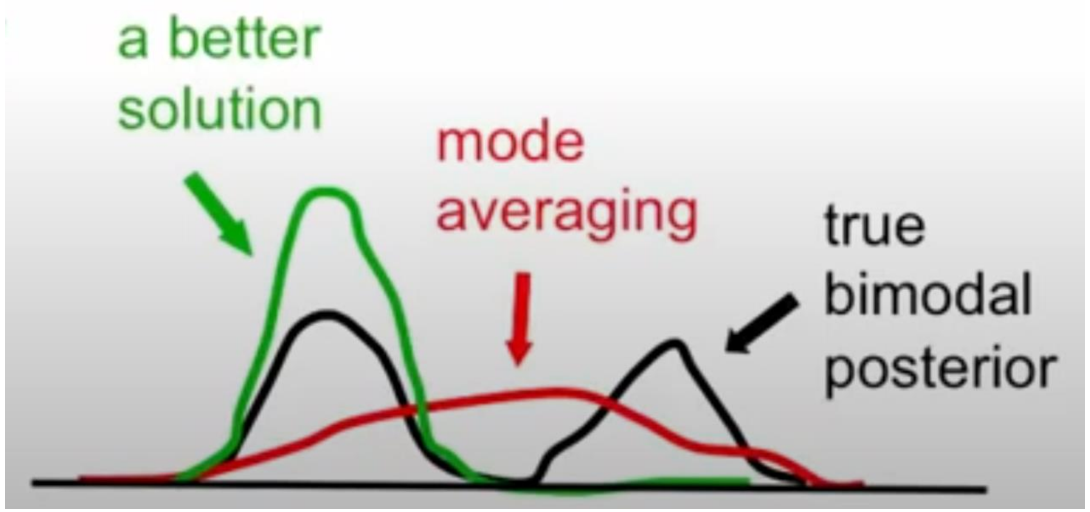
$$\Delta w_{jk} = \epsilon s_j^\gamma (s_k^\gamma - q_k^\gamma)$$

Caveats: Mode averaging

Sleep (generative) phase



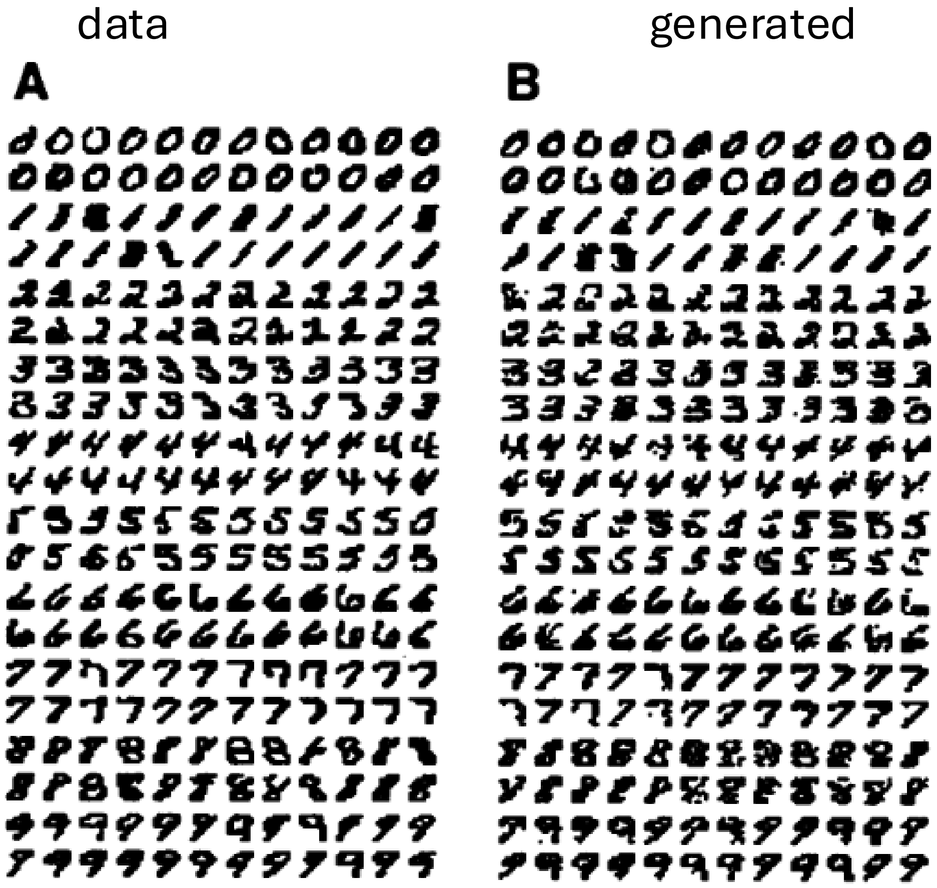
$$\Delta w_{jk} = \epsilon s_j^\gamma (s_k^\gamma - q_k^\gamma)$$



Not a fatal flaw, because wake-phase training partially avoids such situations

Take-home

- Simple idea -> effective data generation
- Mathematically grounded
- Many ideas of today's generational networks are 30 years old


data
generated
A
B