

Energy-Based Models for Sequence Labeling

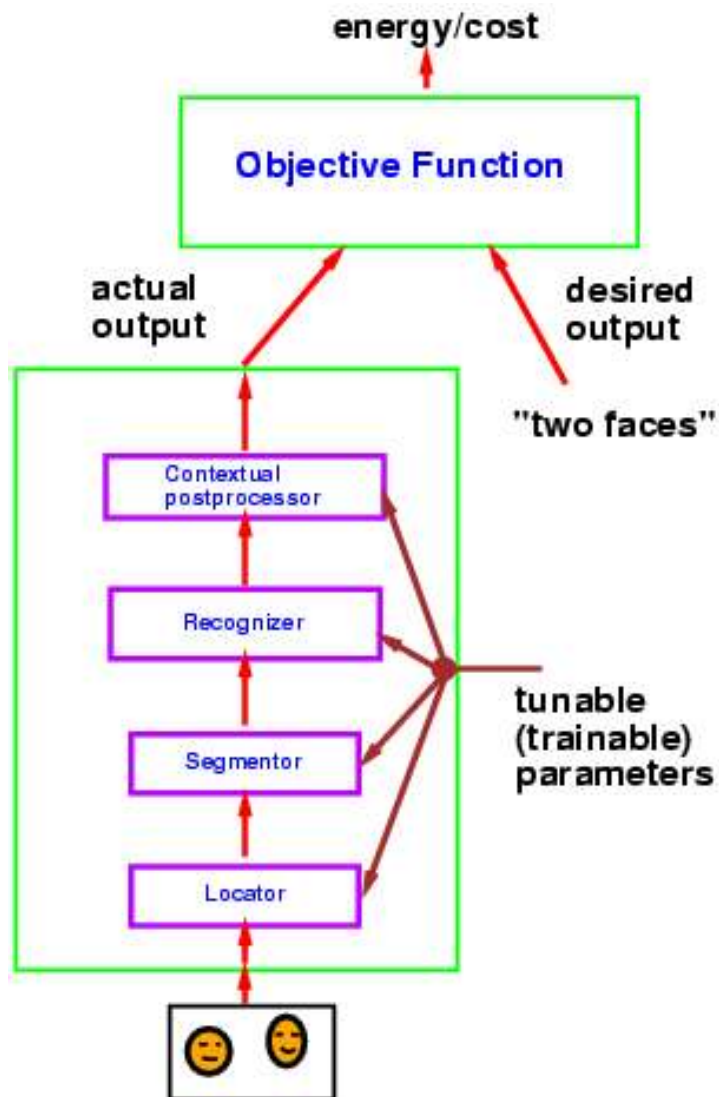
Yann LeCun (Courant Institute, NYU)

Leon Bottou (NEC Labs),

Yoshua Bengio (Montreal),

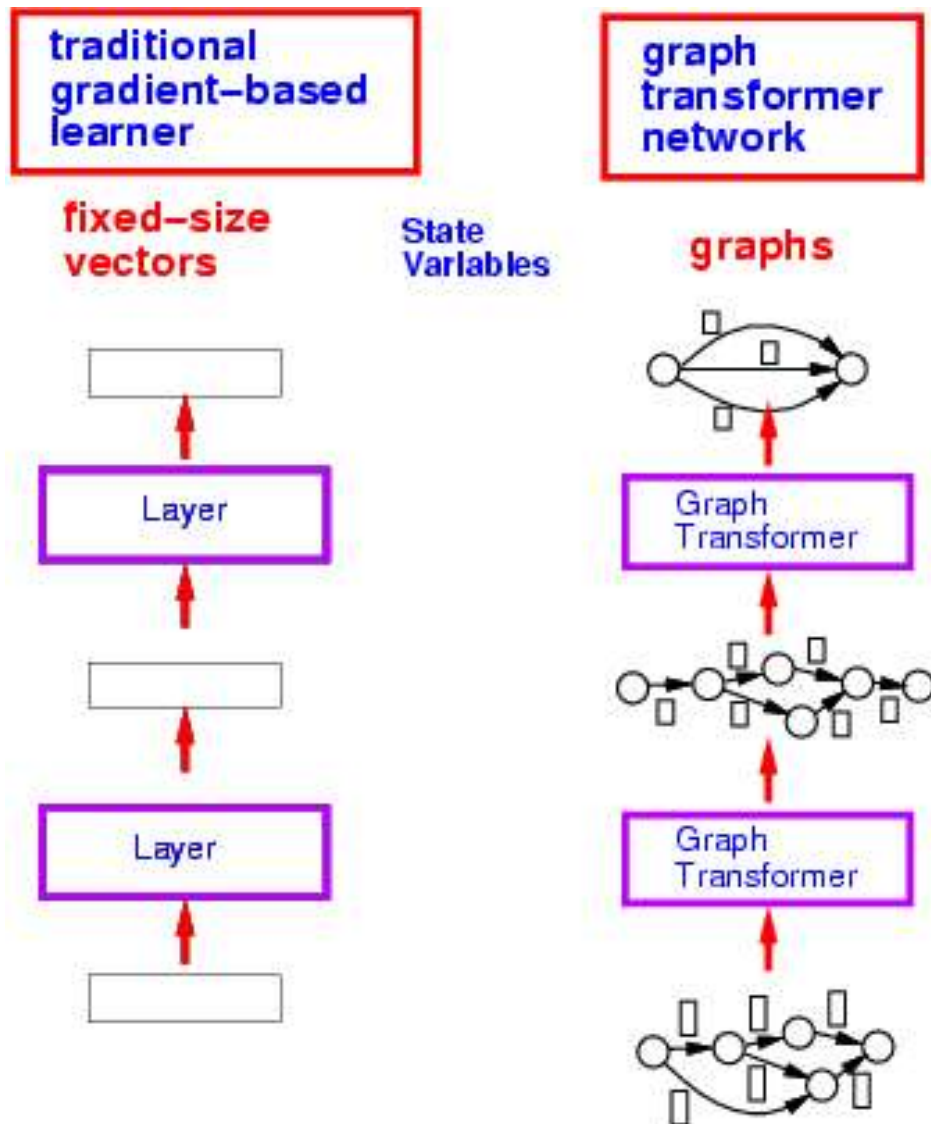
Patrick Haffner (AT&T)

End-to-End Learning.



- Making every single module in the system trainable.
- Every module is trained simultaneously so as to optimize a global loss function.

Using Graphs instead of Vectors.



- Whereas traditional learning machines manipulate **fixed-size vectors**, Graph Transformer Networks **manipulate graphs**.

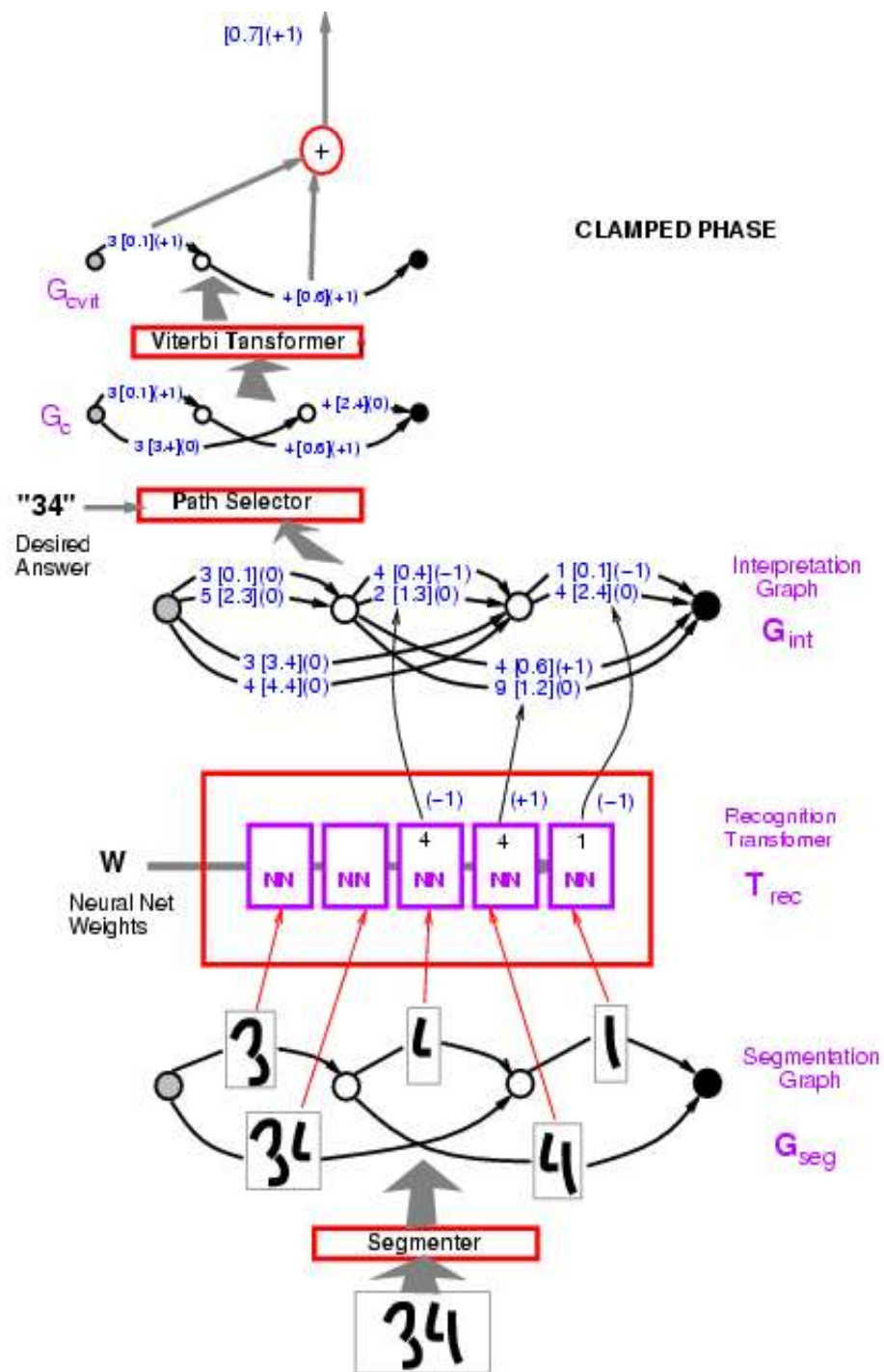
Graph Transformer Networks

Variables:

- ▶ X: input image
- ▶ Z: path in the interpretation graph/segmentation
- ▶ Y: sequence of labels on a path

Loss function: computing the energy of the desired answer:

$$E(W, Y, X)$$



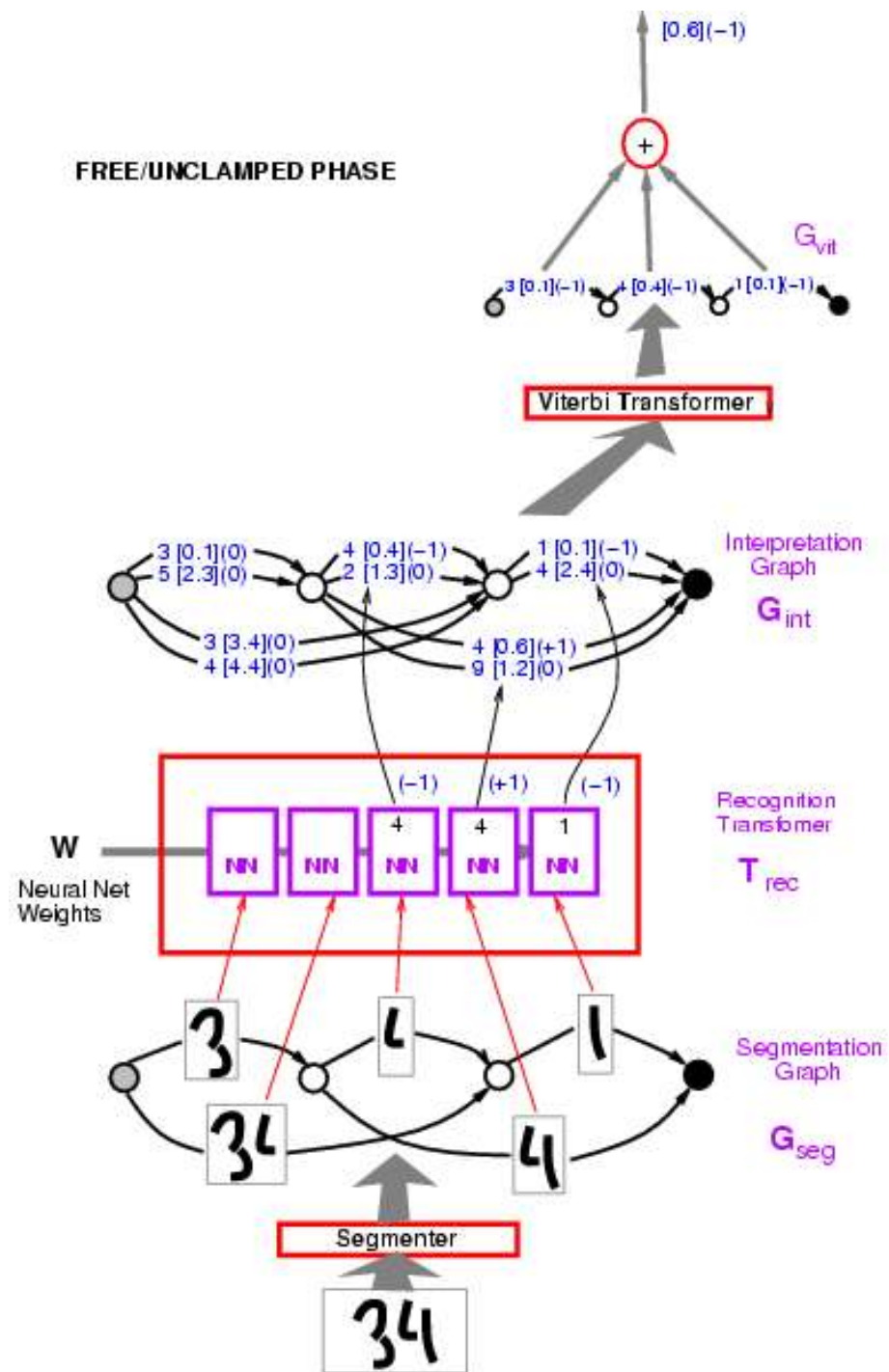
Graph Transformer Networks

Variables:

- ▶ X: input image
- ▶ Z: path in the interpretation graph/segmentation
- ▶ Y: sequence of labels on a path

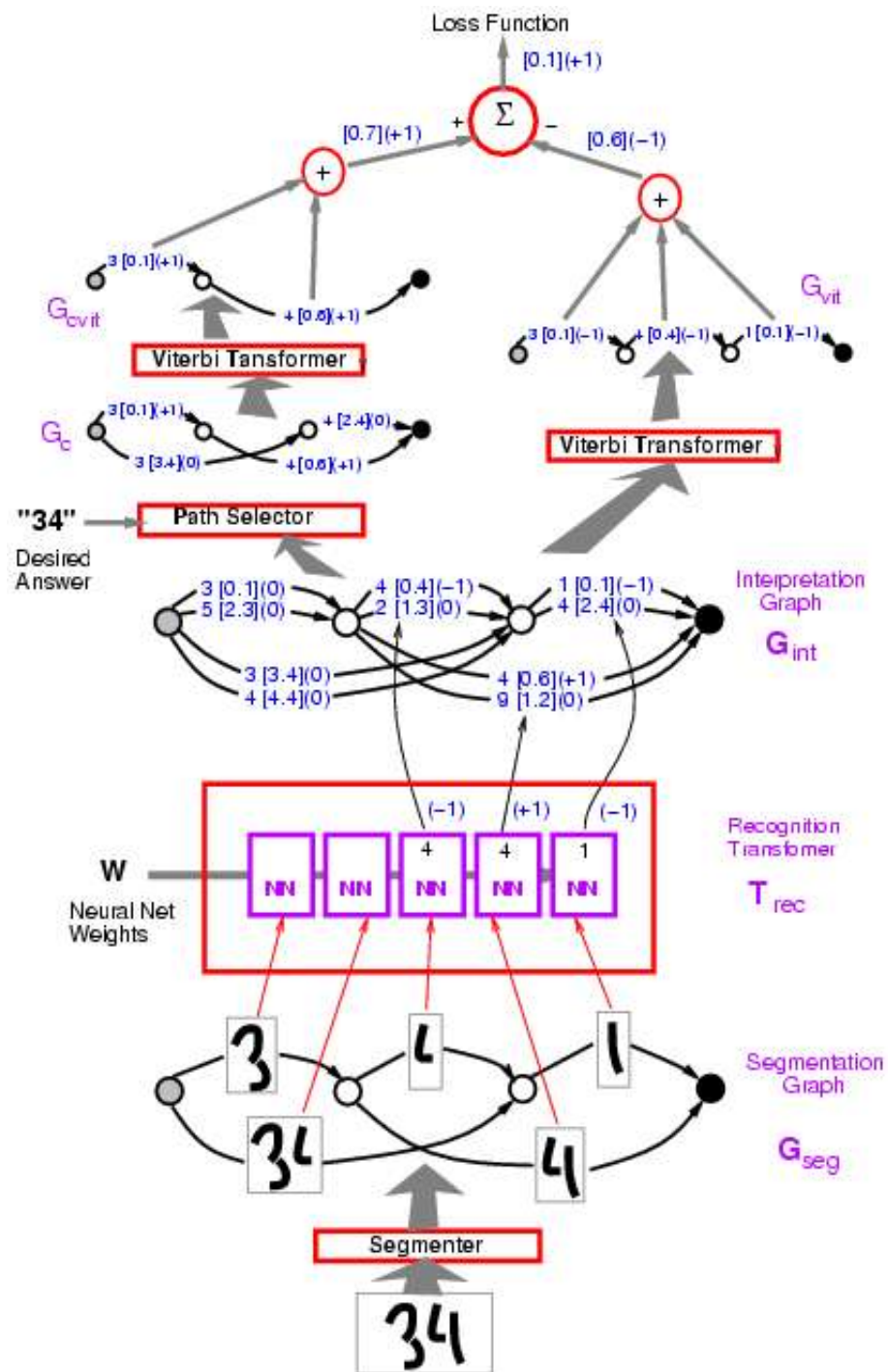
Loss function: computing the contrastive term:

$$E(W, \check{Y}, X)$$



Graph Transformer Networks

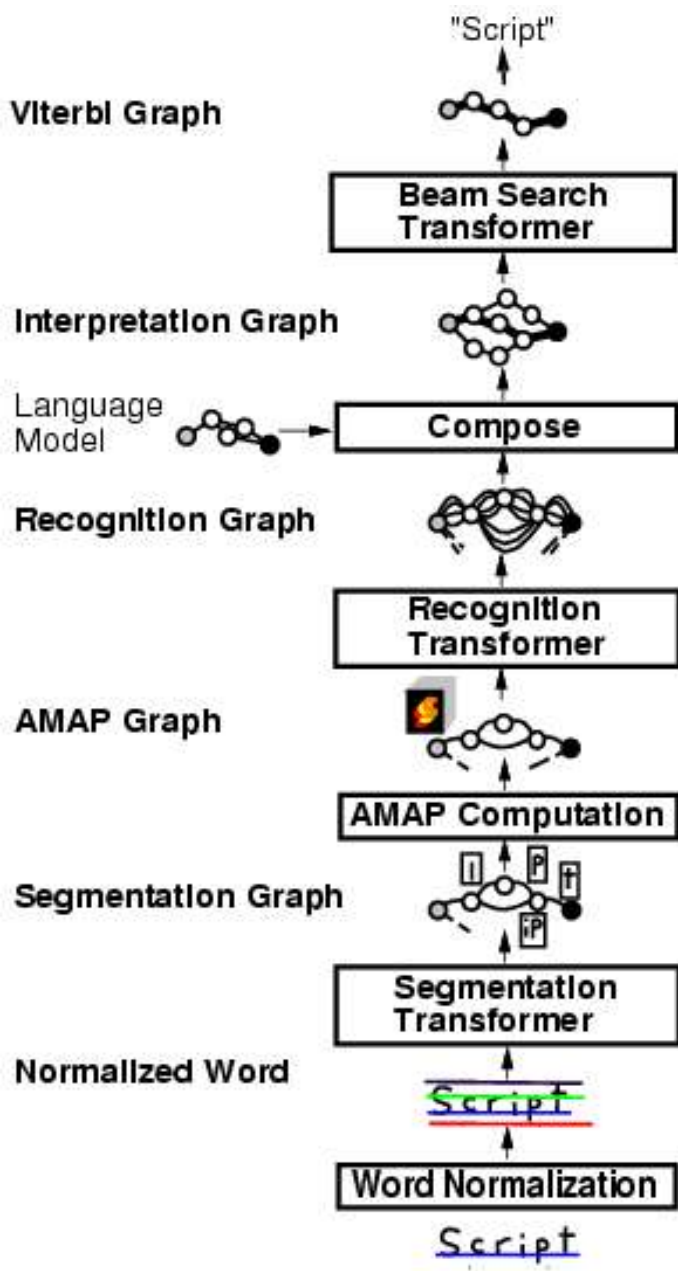
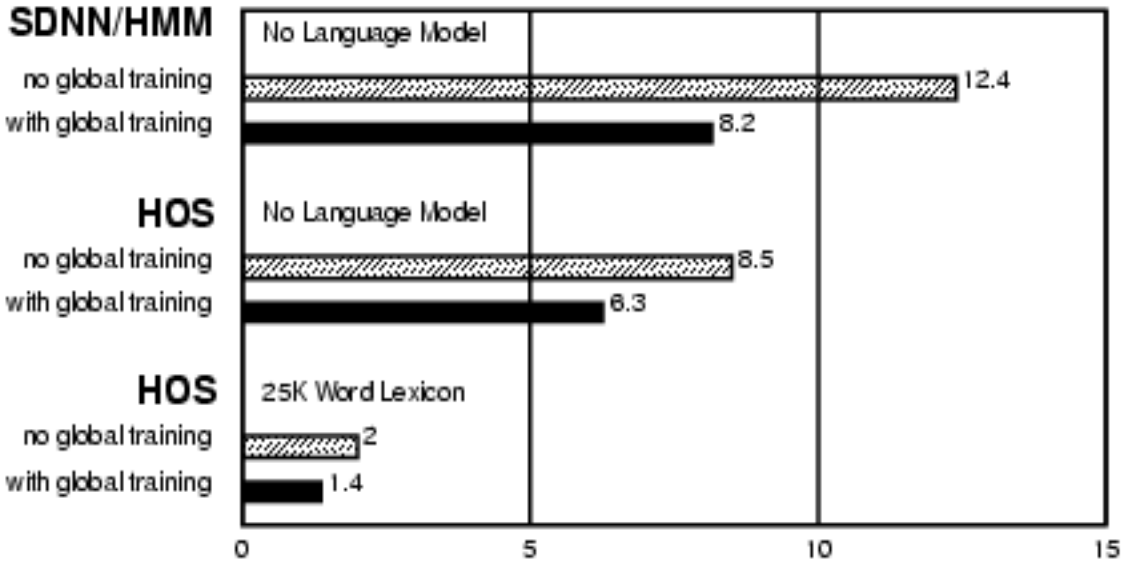
- Example: Perceptron loss
- Loss = Energy of desired answer – Energy of best answer.
- ▶ (no margin)



Global Training Helps

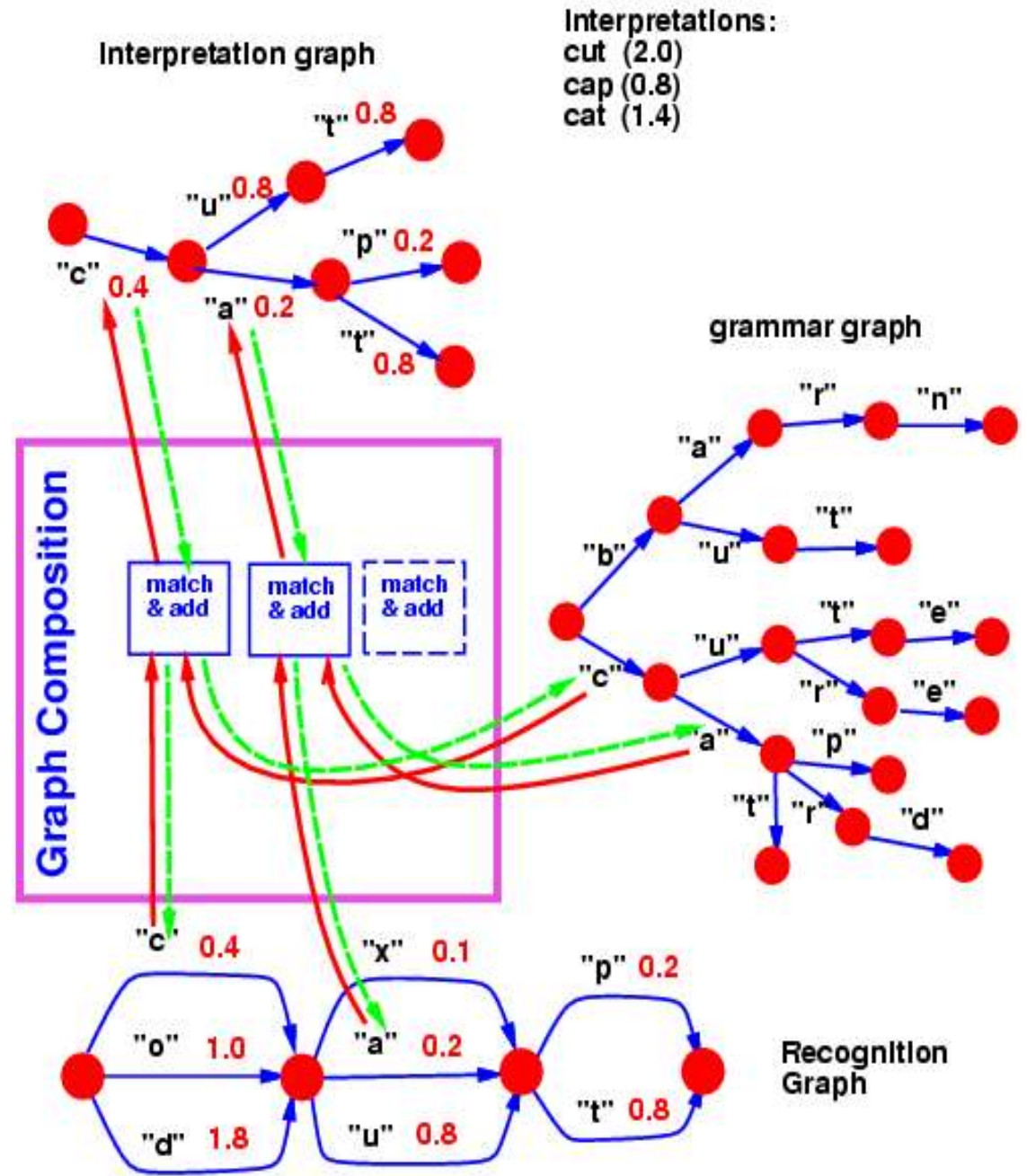
Pen-based handwriting recognition (for tablet computer)

▶ [Bengio&LeCun 1995]



Graph Composition, Transducers.

- The composition of two graphs can be computed, the same way the dot product between two vectors can be computed.
- General theory: semi-ring algebra on weighted finite-state transducers and acceptors.



Check Reader

- Graph transformer network trained to read **check amounts**.
- Trained globally with **Negative-Log-Likelihood loss**.
- 50%** percent correct, **49%** reject, **1%** error (detectable later in the process).
- Fielded in **1996**
- Processes an estimated **10%** of **all the checks written in the US**.

