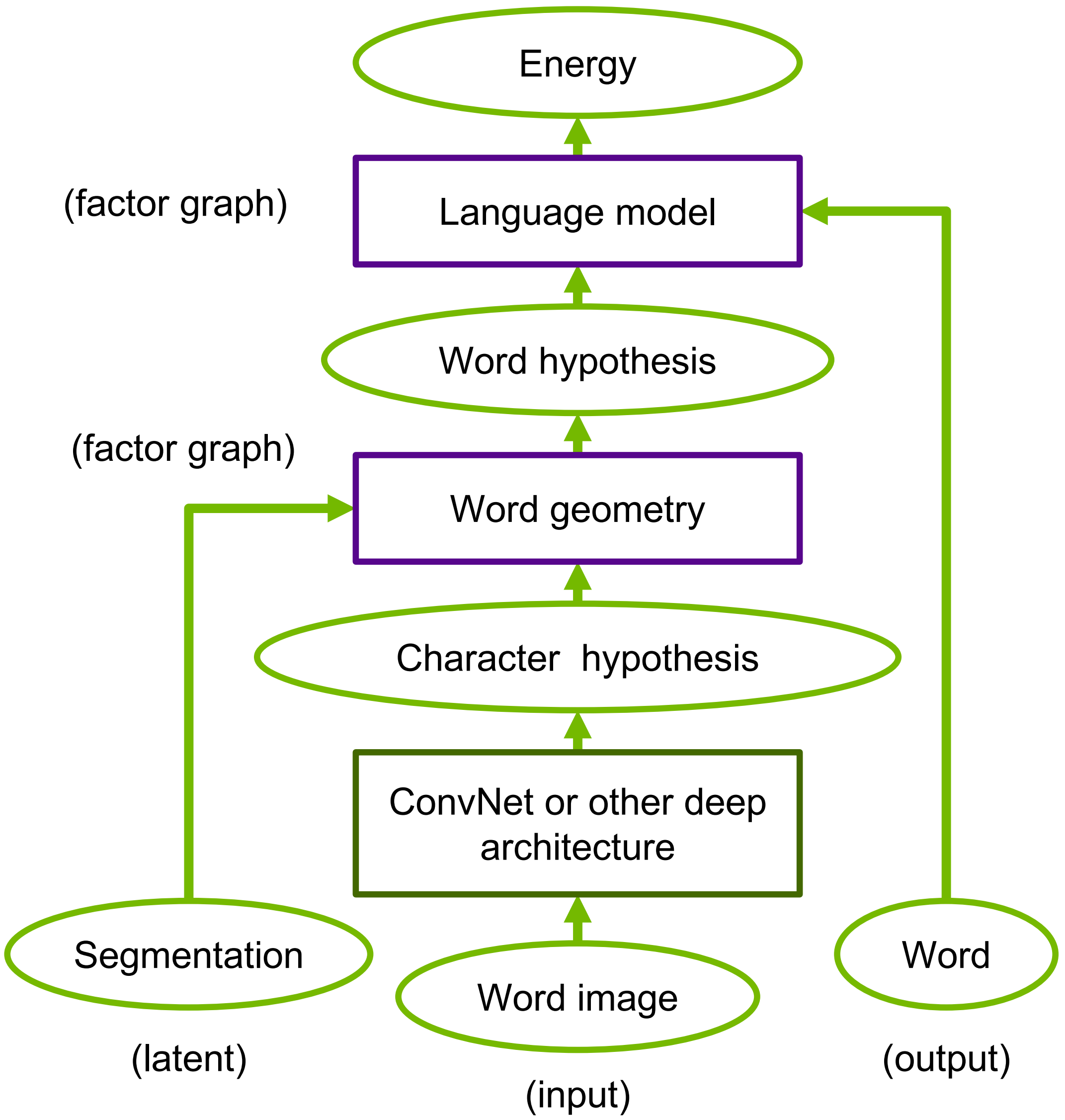
Structural Prediction and Natural Language Processing

Jay Urbain, PhD
Credits: NYU, nVidia DLI

End-to-end learning – Word-level discriminative training



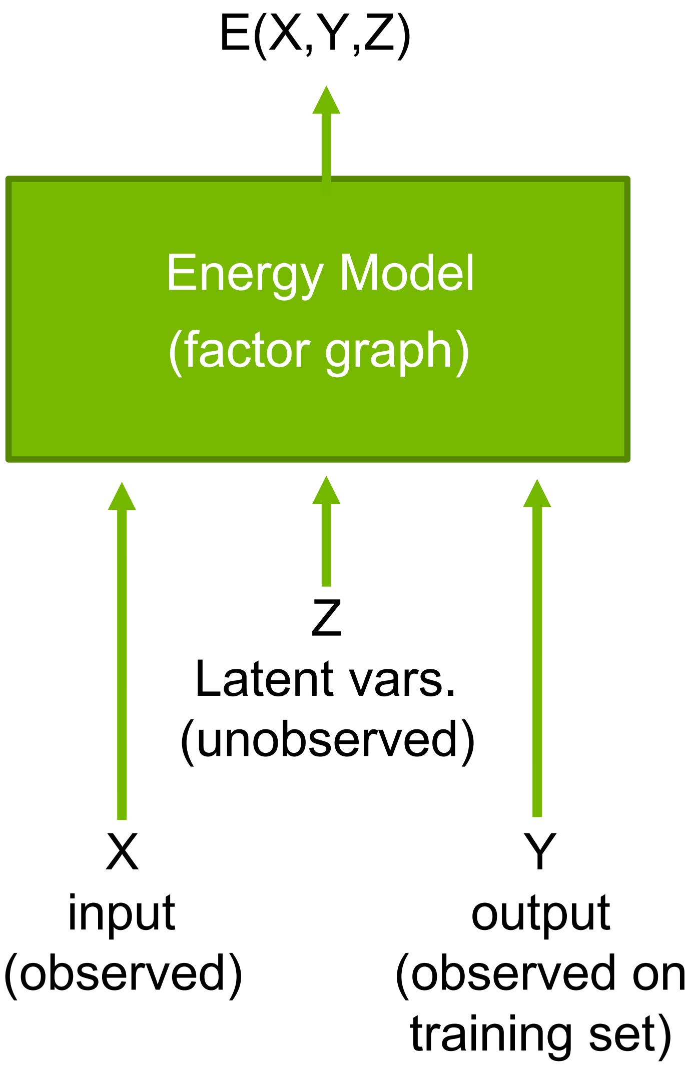
- Making every single module in the system trainable.
- Every module is trained simultaneously so as to optimize a global loss function.
- Includes the feature extractor, the recognizer, and the contextual post-processor (graphical model)
- Problem: back-propagating gradients through the graphical model.

Integrating Deep Learning and structured prediction

- Deep Learning systems can be assembled into factor graphs (graphical models)
  - Energy function measures the "compatibility" between variables.
  - X: inputs (observed variables)
  - Z: latent variables (never observed)
  - Y: outputs (observed on training set)
- Inference is energy minimization (MAP) or free energy minimization (marginalization) over Z and Y given an X

- Y* = Argmin E(X,Y,Z)

- Energy function can embed whole deep learning systems (e.g. ConvNets).



E(X,Y,Z)

Energy Model

(factor graph)

X
input
(observed)

Z
Latent vars.
(unobserved)

Y
output
(observed on
training set)

Latent variable models

- The energy includes “hidden” variables Z whose value is never given to us
  - We can minimize the energy over those latent variables
  - We can also “marginalize” the energy over the latent variables

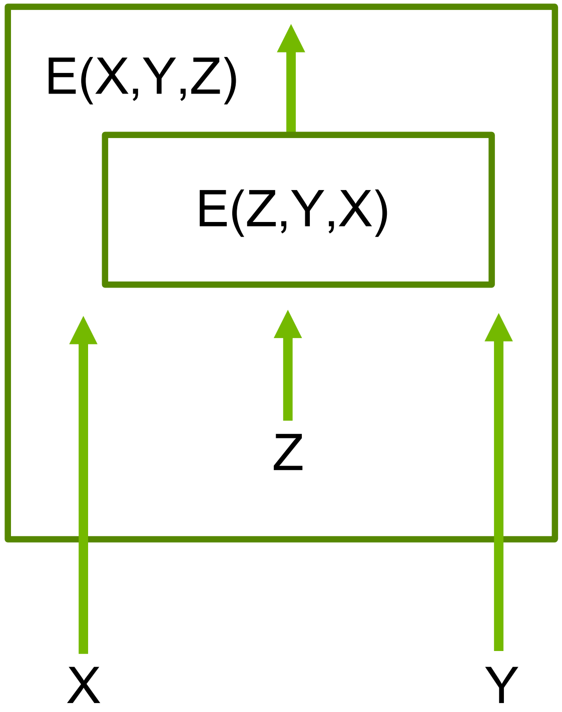
- Minimization over latent variables:

$$E(Y,X) = \min_{Z \in \mathcal{Z}} E(Z,Y,X).$$

- Marginalization over latent variables:

$$E(X,Y) = -\frac{1}{\beta} \log \int_{z \in \mathcal{Z}} e^{-\beta E(z,Y,X)}$$

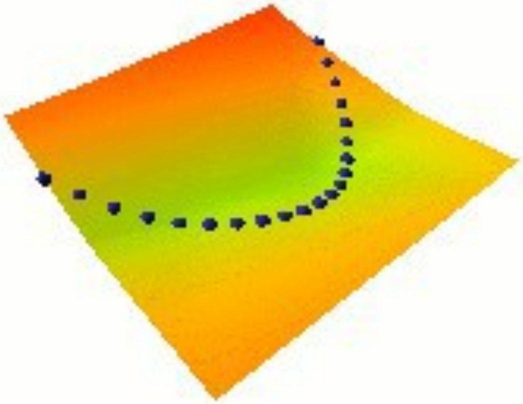
- Estimating this integral may require some approximations (sampling, variational methods, maximum a posteriori....)



Loss functions for EBM

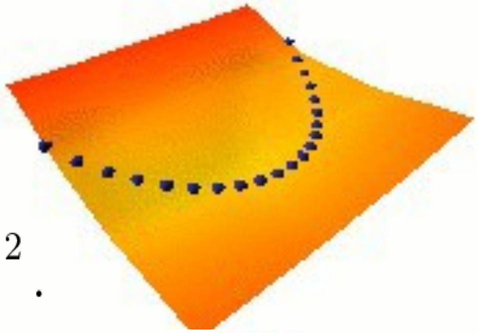
Loss functions

– Energy Loss (pull down on the data points)



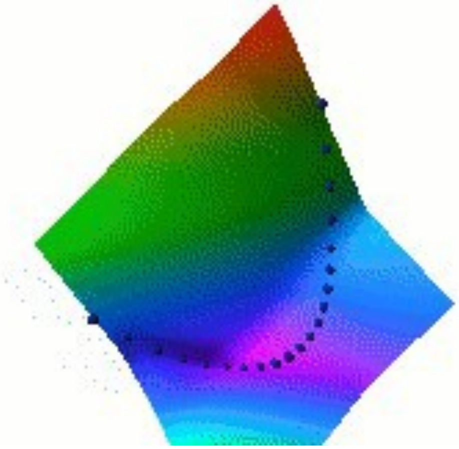
$$L_{energy}(Y^i, E(W, \mathcal{Y}, X^i)) = E(W, Y^i, X^i).$$

– Square-Square Loss



$$L_{\mathrm{sq-sq}}(W, Y^i, X^i) = E(W, Y^i, X^i)^2 + \left(\max(0, m - E(W, \bar{Y}^i, X^i))\right)^2.$$

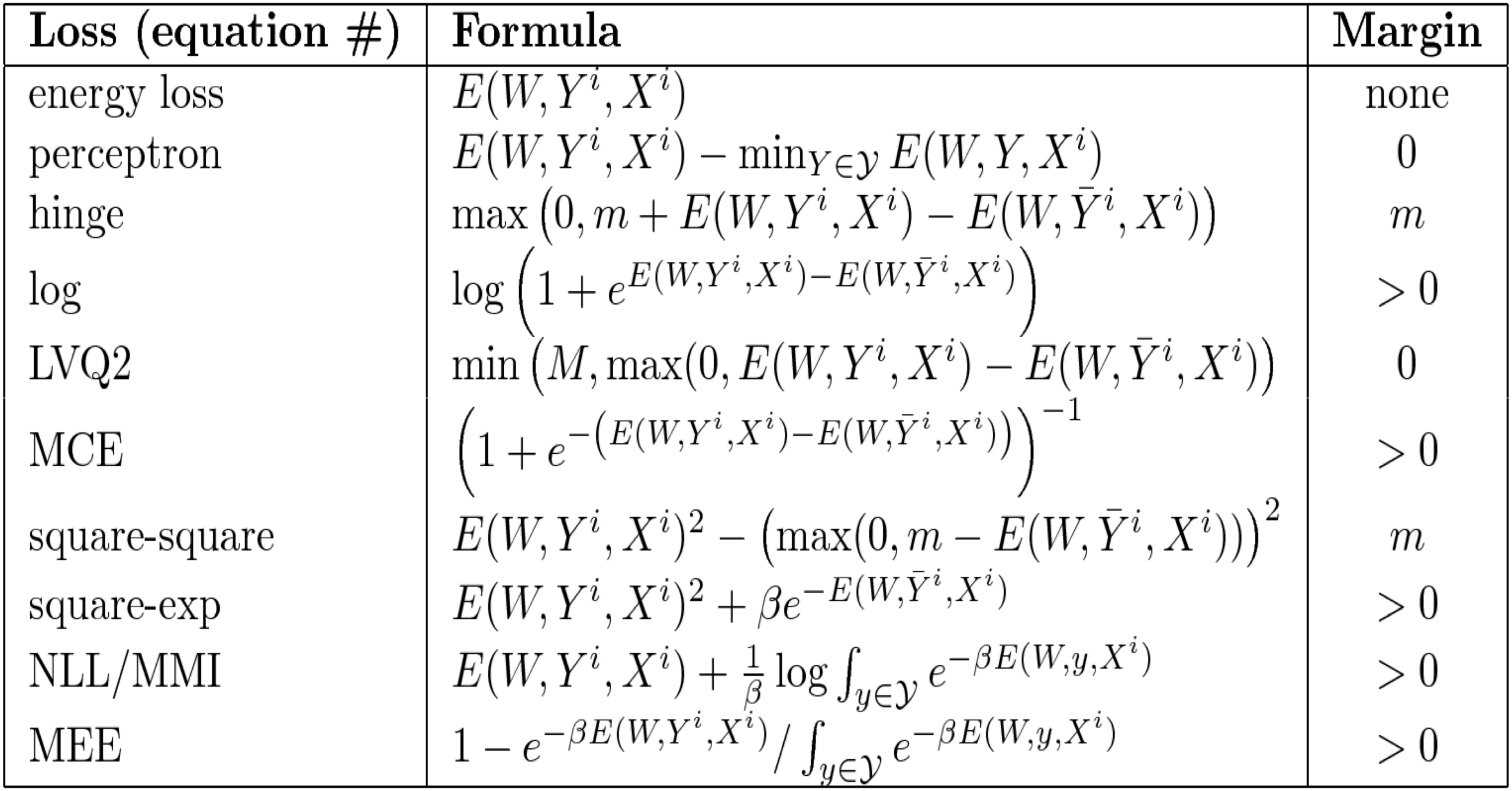
– Negative Log Likelihood



$$\mathcal{L}_{\mathrm{nll}}(W, \mathcal{S}) = \frac{1}{P}\sum_{i=1}^{P}\left(E(W, Y^i, X^i) + \frac{1}{\beta}\log\int_{y\in\mathcal{Y}} e^{-\beta E(W, y, X^i)}\right).$$

Loss Function to train energy-based models

- Good and bad loss functions
- A tutorial on Energy-Based Learning [LeCun et al 2006]



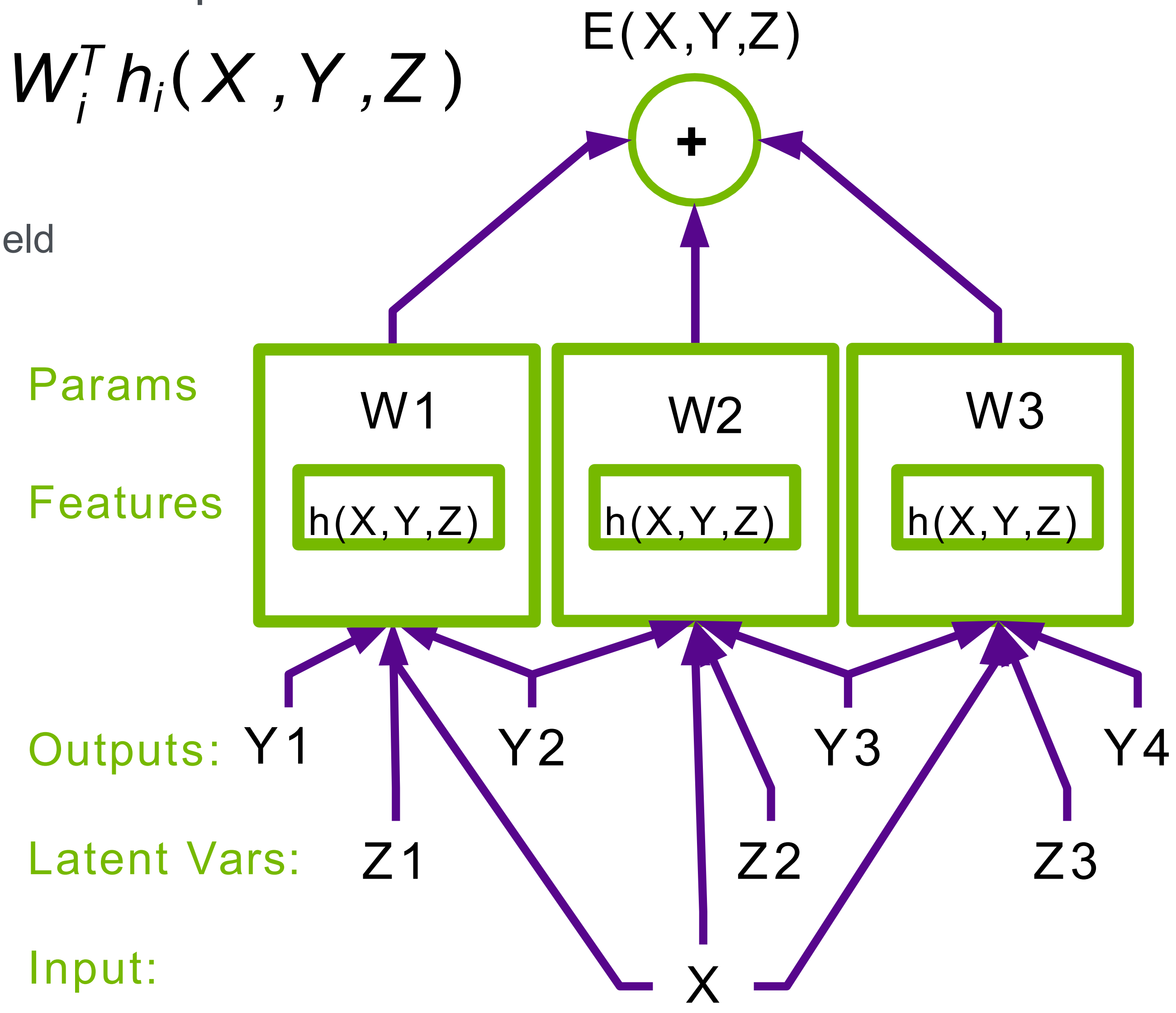
Loss (equation #)	**Formula**	**Margin**
energy loss	$E(W, Y^i, X^i)$	none
perceptron	$E(W, Y^i, X^i) - \min_{Y \in \mathcal{Y}} E(W, Y, X^i)$	0
hinge	$\max\left(0, m + E(W, Y^i, X^i) - E(W, \bar{Y}^i, X^i)\right)$	m
log	$\log\left(1 + e^{E(W,Y^i,X^i) - E(W,\bar{Y}^i,X^i)}\right)$	> 0
LVQ2	$\min\left(M, \max(0, E(W, Y^i, X^i) - E(W, \bar{Y}^i, X^i)\right)$	0
MCE	$\left(1 + e^{-\left(E(W,Y^i,X^i) - E(W,\bar{Y}^i,X^i)\right)}\right)^{-1}$	> 0
square-square	$E(W, Y^i, X^i)^2 - \left(\max(0, m - E(W, \bar{Y}^i, X^i))\right)^2$	m
square-exp	$E(W, Y^i, X^i)^2 + \beta e^{-E(W,\bar{Y}^i,X^i)}$	> 0
NLL/MMI	$E(W, Y^i, X^i) + \frac{1}{\beta} \log \int_{y \in \mathcal{Y}} e^{-\beta E(W,y,X^i)}$	> 0
MEE	$1 - e^{-\beta E(W,Y^i,X^i)} / \int_{y \in \mathcal{Y}} e^{-\beta E(W,y,X^i)}$	> 0

“Shallow” structured prediction

- Energy function is linear in the parameters

$$E(X,Y,Z)=\sum_i W_i^T h_i(X,Y,Z)$$

- with the NLL Loss :
  - Conditional Random Field [Lafferty, McCallum, Pereira 2001]
- with Hinge Loss:
  - Max Margin Markov Nets and Latent SVM [Taskar, Altun, Hofmann...]
- with Perceptron Loss
  - Structured Perceptron [Collins...]

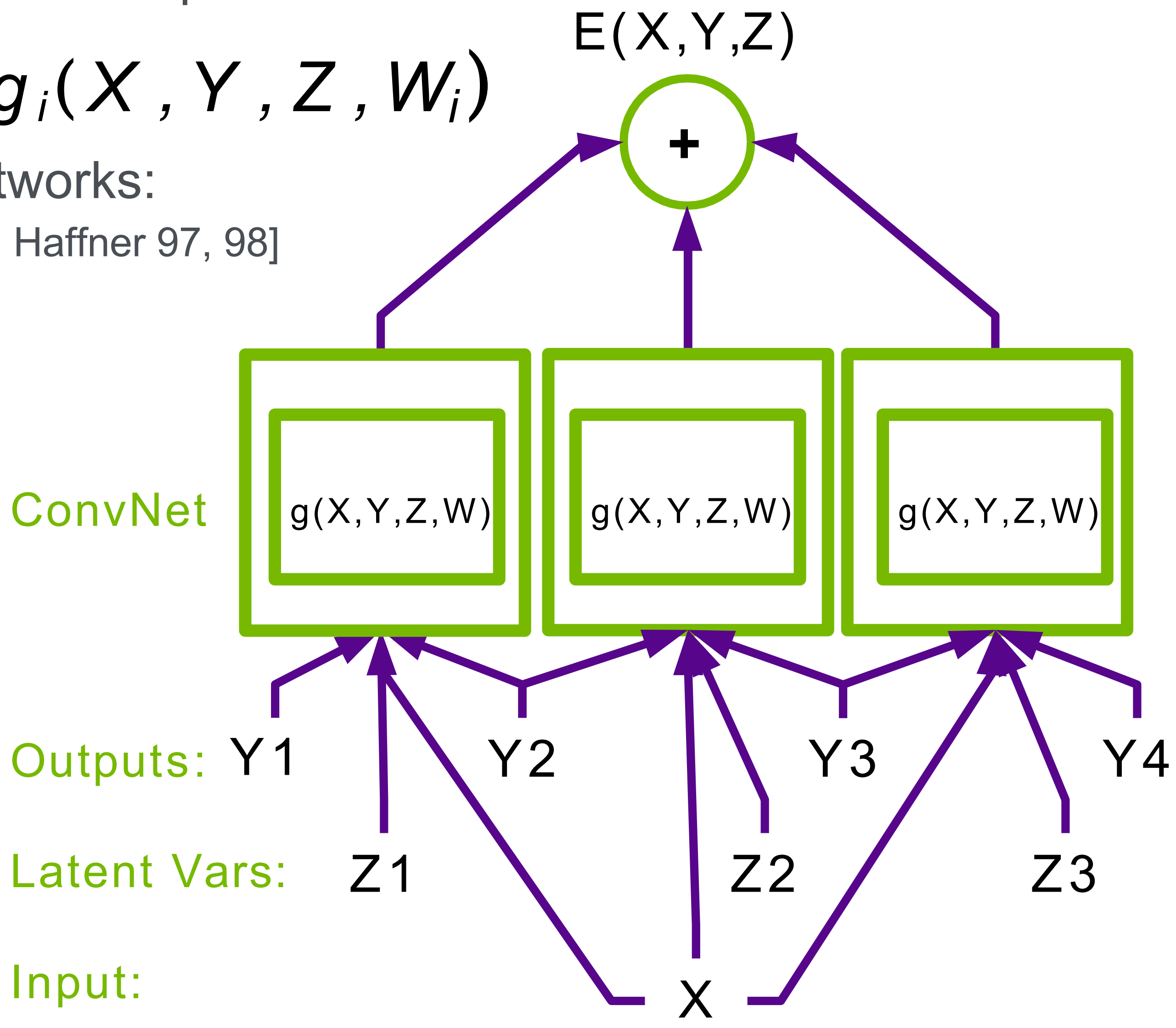


“Shallow” structured prediction

- Energy function is linear in the parameters

$$E(X,Y,Z)=\sum_i g_i(X,Y,Z,W_i)$$

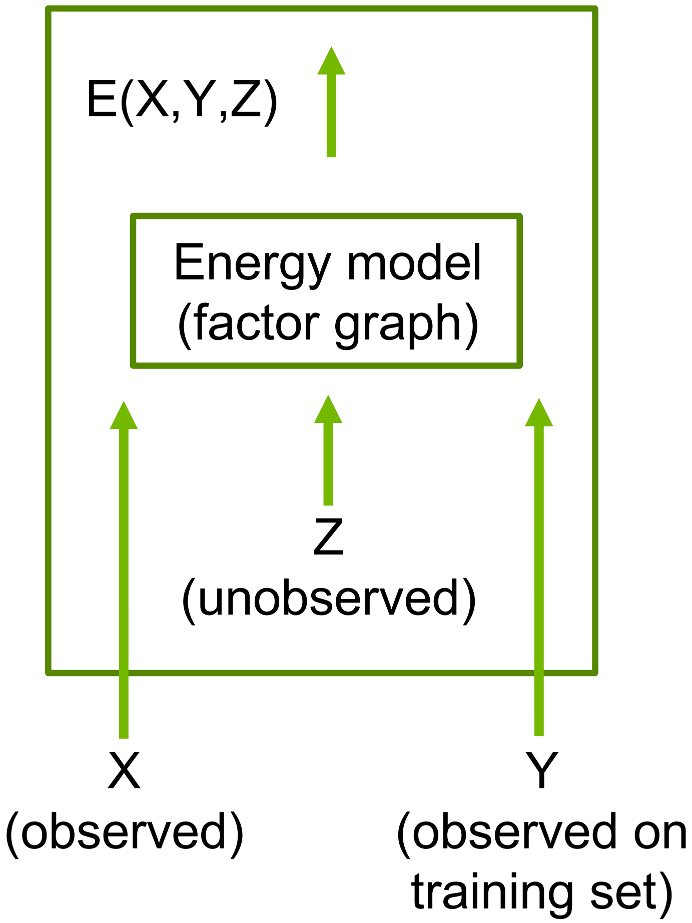
- Graph Transformer Networks:
  - [LeCun, Bottou, Bengio, Haffner 97, 98]
  - NLL loss
  - Perception loss



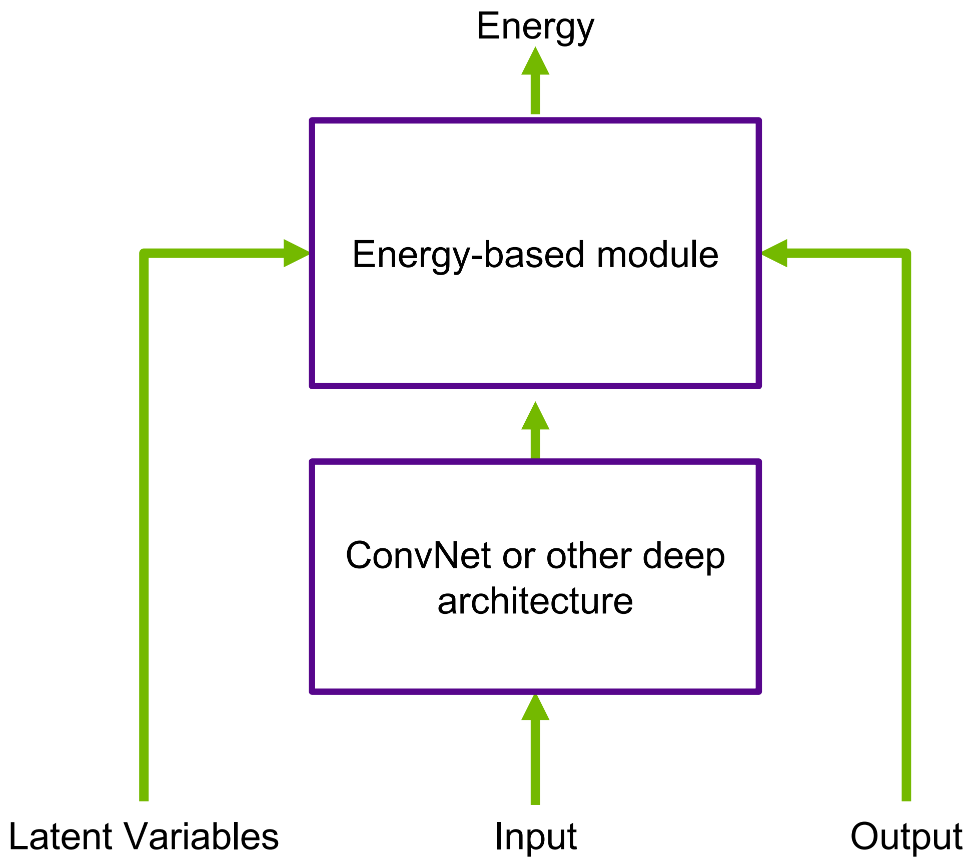
Integrating Deep Learning and structured prediction

- Deep Net + Graphical Model = Factor Graph
  - Energy function is a sum of factors
  - Factors can embed deep architectures X: observed variables (inputs)
  - Z: never observed (latent variables)
  - Y: observed on training set (output variables)
- Inference is energy minimization (MAP) or free energy minimization (marginalization) over Z and Y given an X
  - F(X,Y) = MIN_z E(X,Y,Z)
  - F(X,Y) = -log[SUM_z exp(-E(X,Y,Z))]

F(X,Y) = MIN_z E(X,Y,Z)



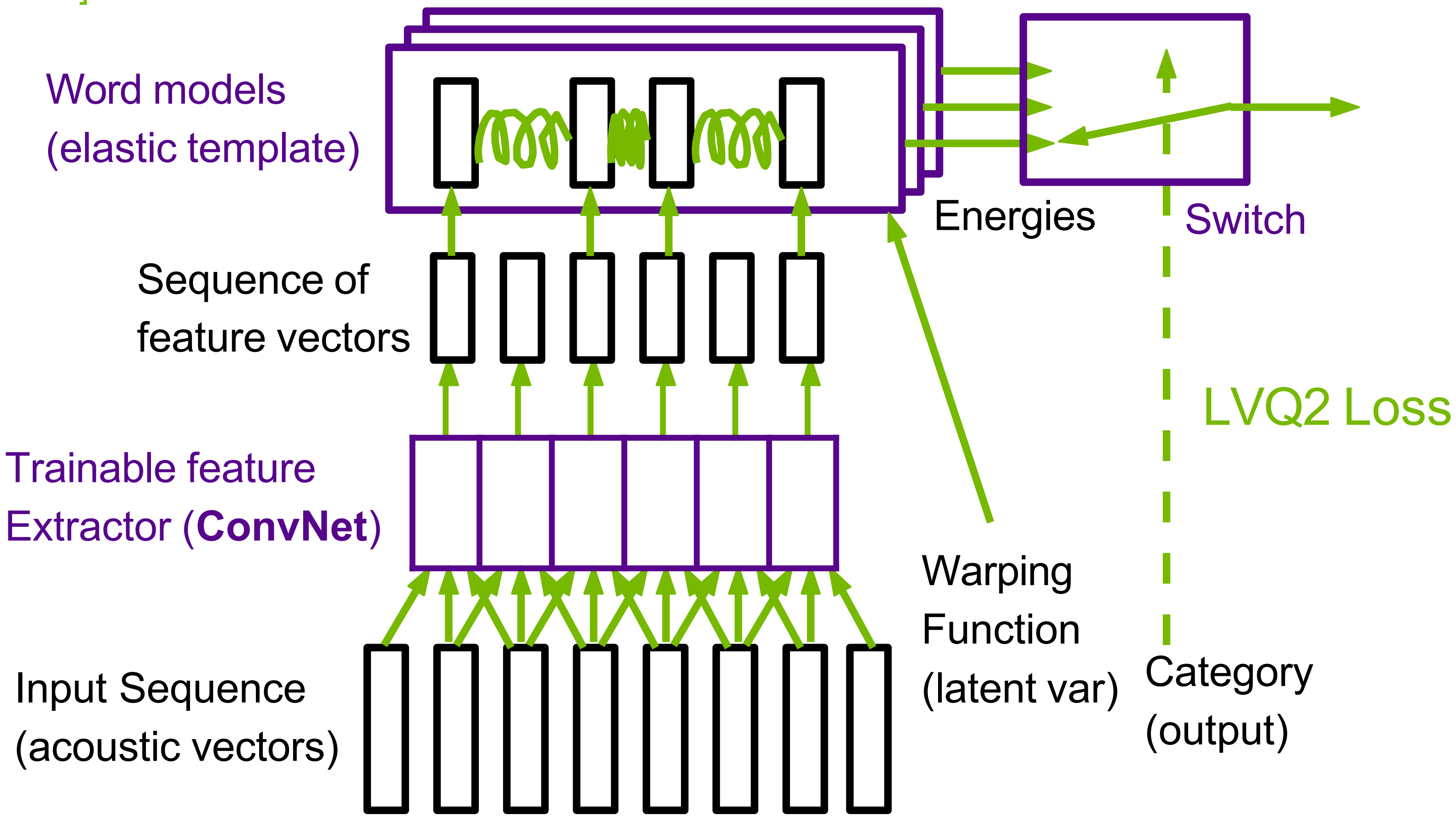
End-to-end learning – Word-level training



- Making every single module in the system trainable.
- Every module is trained simultaneously so as to optimize a global loss function.
- Includes the feature extractor, the recognizer, and the contextual post-processor (graphical model)
- Problem: back-propagating gradients through the graphical model.

Deep structured prediction for speech recognition (1991)

Spoken word recognition with trainable elastic templates and trainable feature extraction [Driancourt&Bottou 1991, Bottou 1991, Driancourt 1994]



Word models
(elastic template)

Energies

Switch

Sequence of
feature vectors

LVQ2 Loss

Trainable feature
Extractor (**ConvNet**)

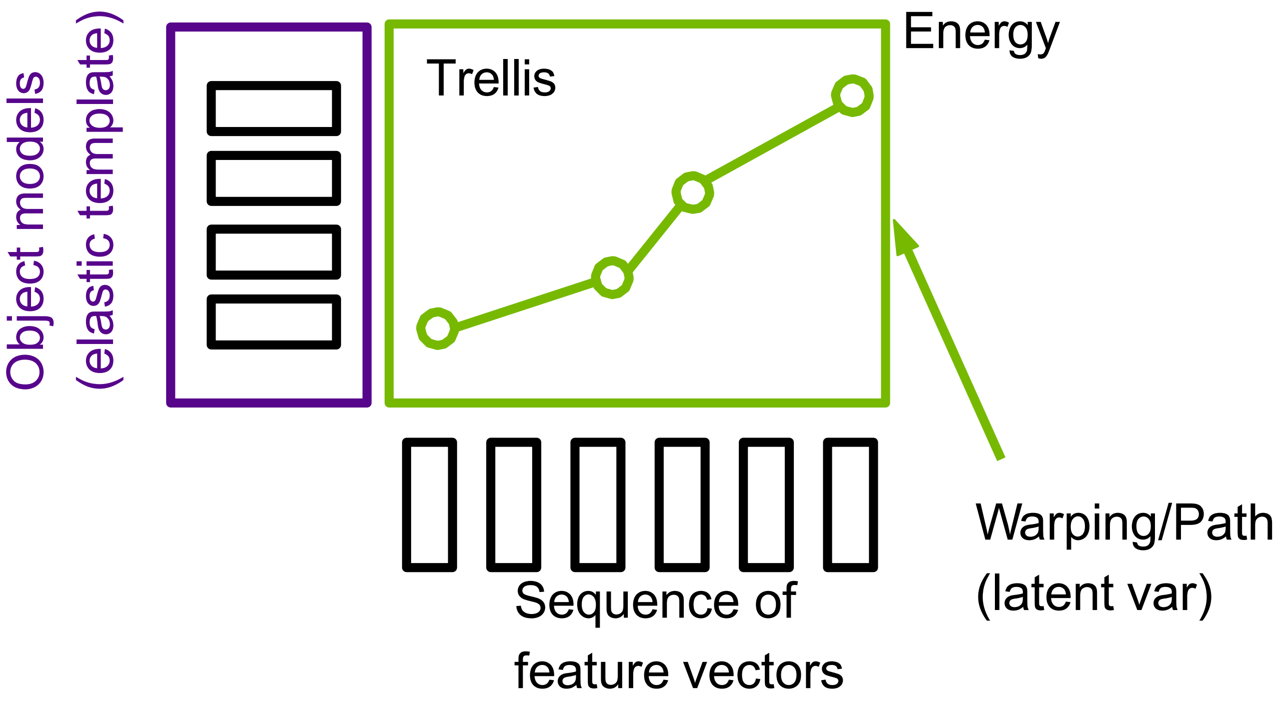
Warping
Function
(latent var)

Category
(output)

Input Sequence
(acoustic vectors)

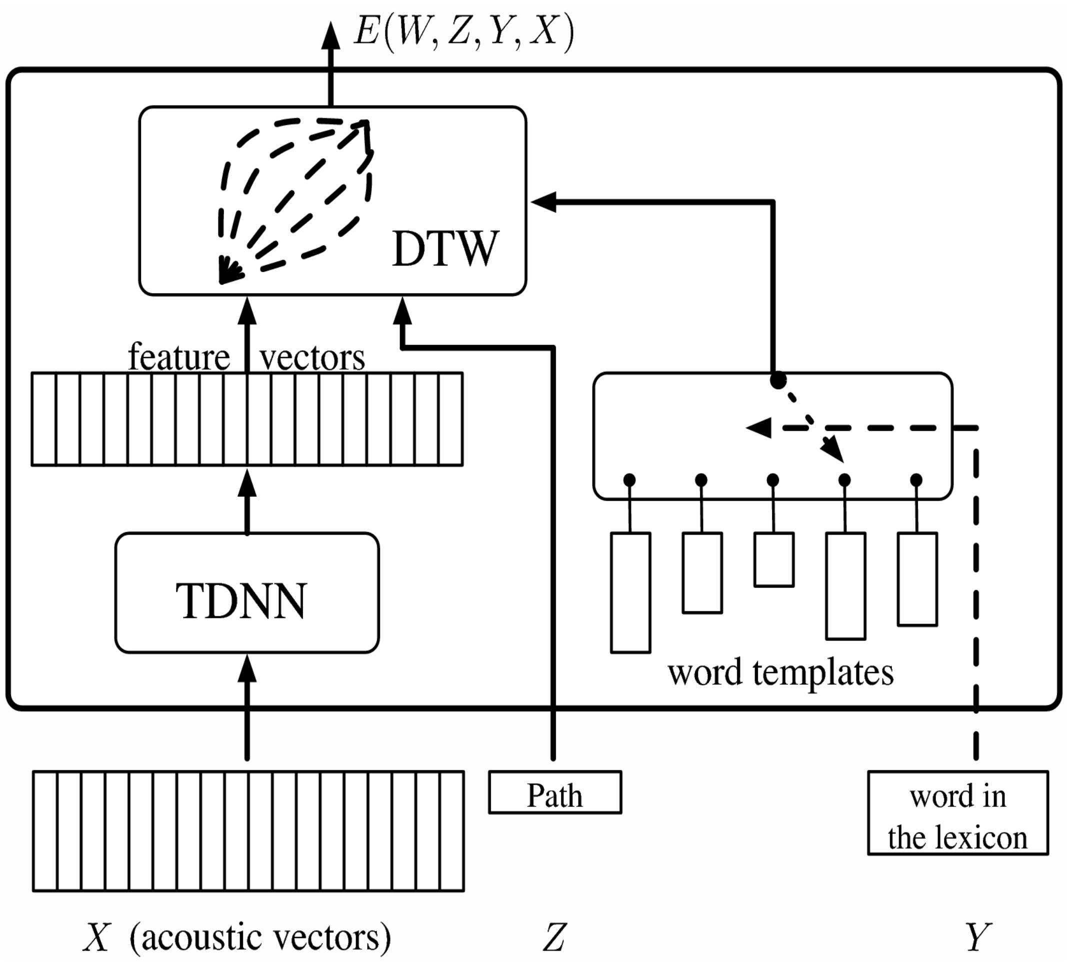
Alignment through dynamic time warping

- Spoken word recognition with trainable elastic templates and trainable feature extraction
  - [Driancourt&Bottou 1991, Bottou 1991, Driancourt 1994]
- Elastic matching using dynamic time warping (Viterbi algorithm on a trellis).

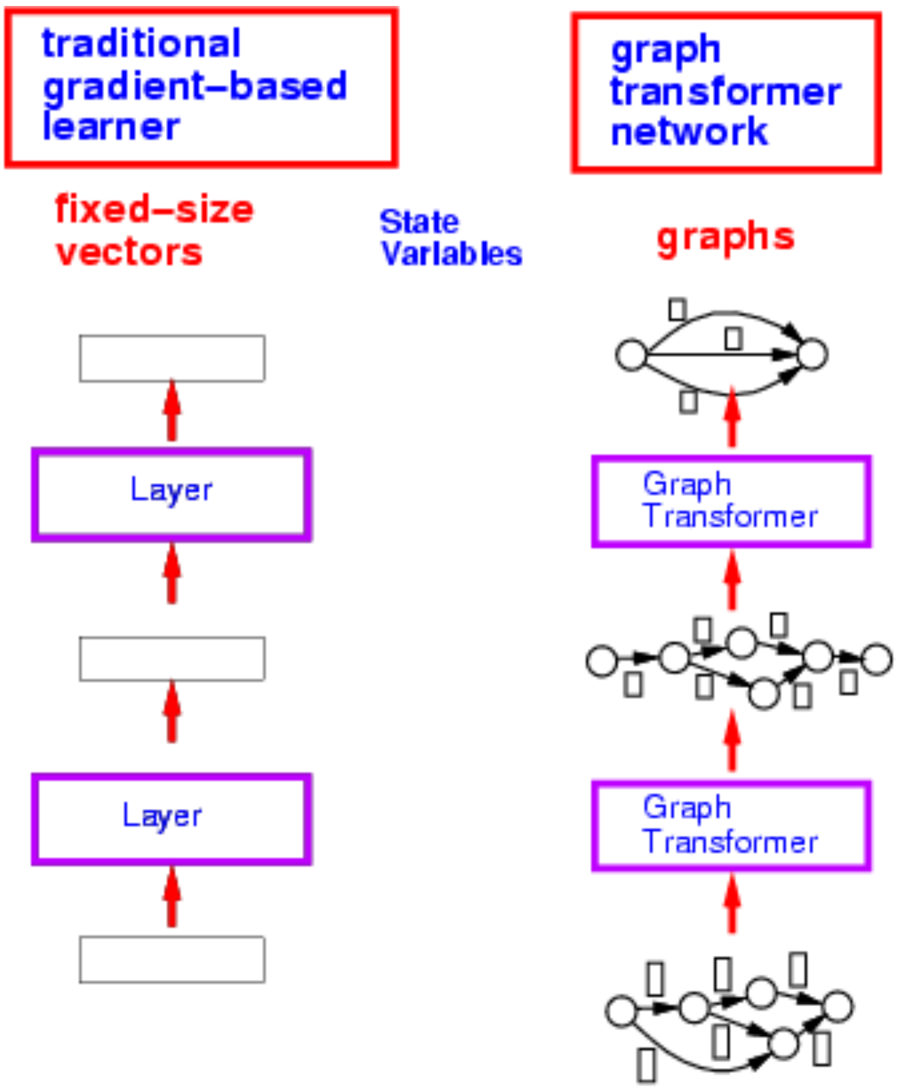


Deep factors / deep graph: ASR with TDNN/DTW

- Trainable Automatic Speech Recognition system with convolutional nets (TDNN) and dynamic time warping (DTW)
- Training the feature extractor as part of the whole process.
- with the LVQ2 Loss :
  - Driancourt and Bottou's speech recognizer (1991)
- with NLL:
  - Bengio's speech recognizer (1992)
  - Haffner's speech recognizer (1993)



Using graphs instead of vectors or arrays



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

Graph transformer networks

Variables:

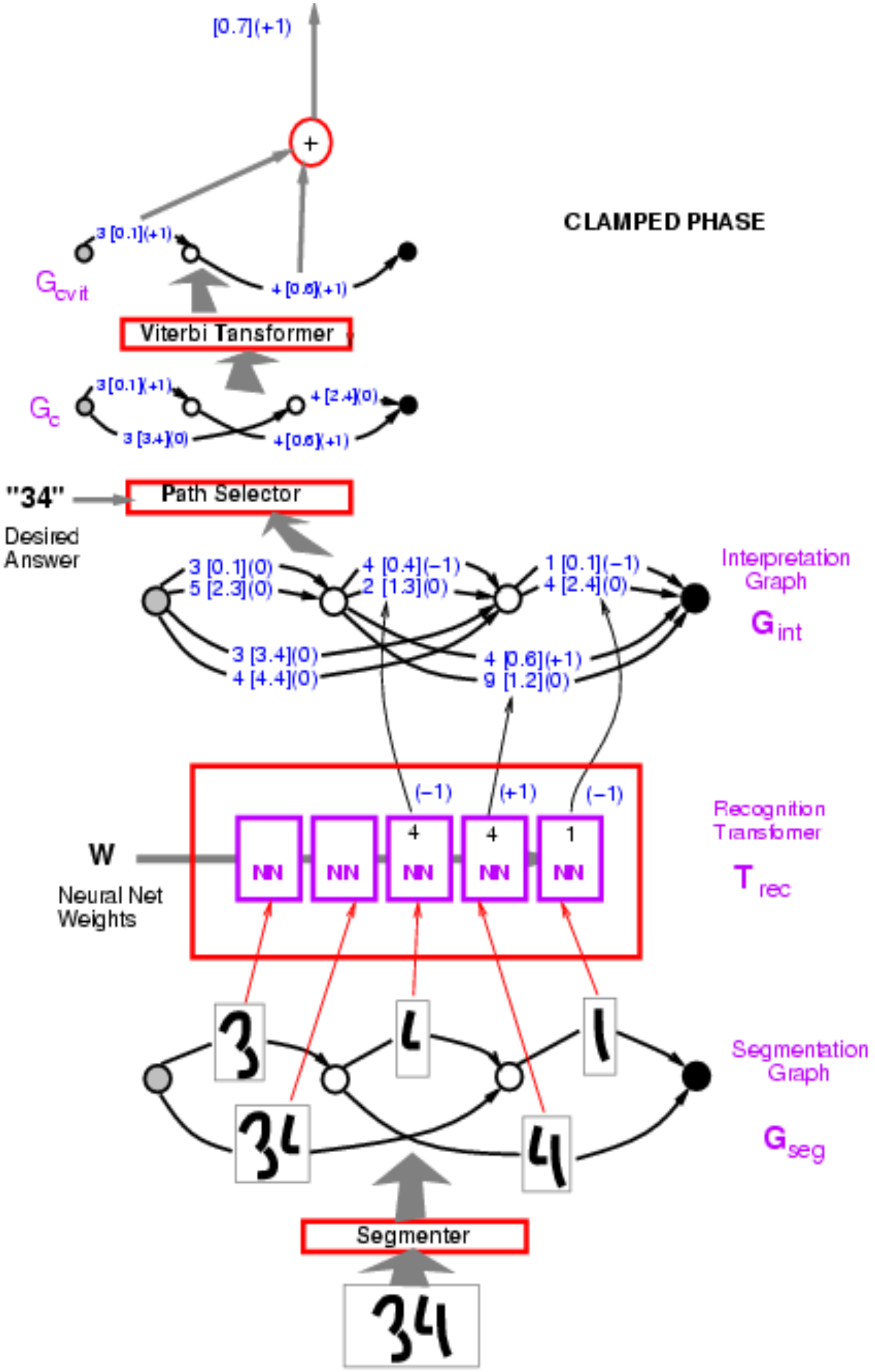
– X: input image

– Z: path in the interpretation graph/segmentation

– Y: sequence of labels on a path

Loss function: computing the energy of the desired answer:

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



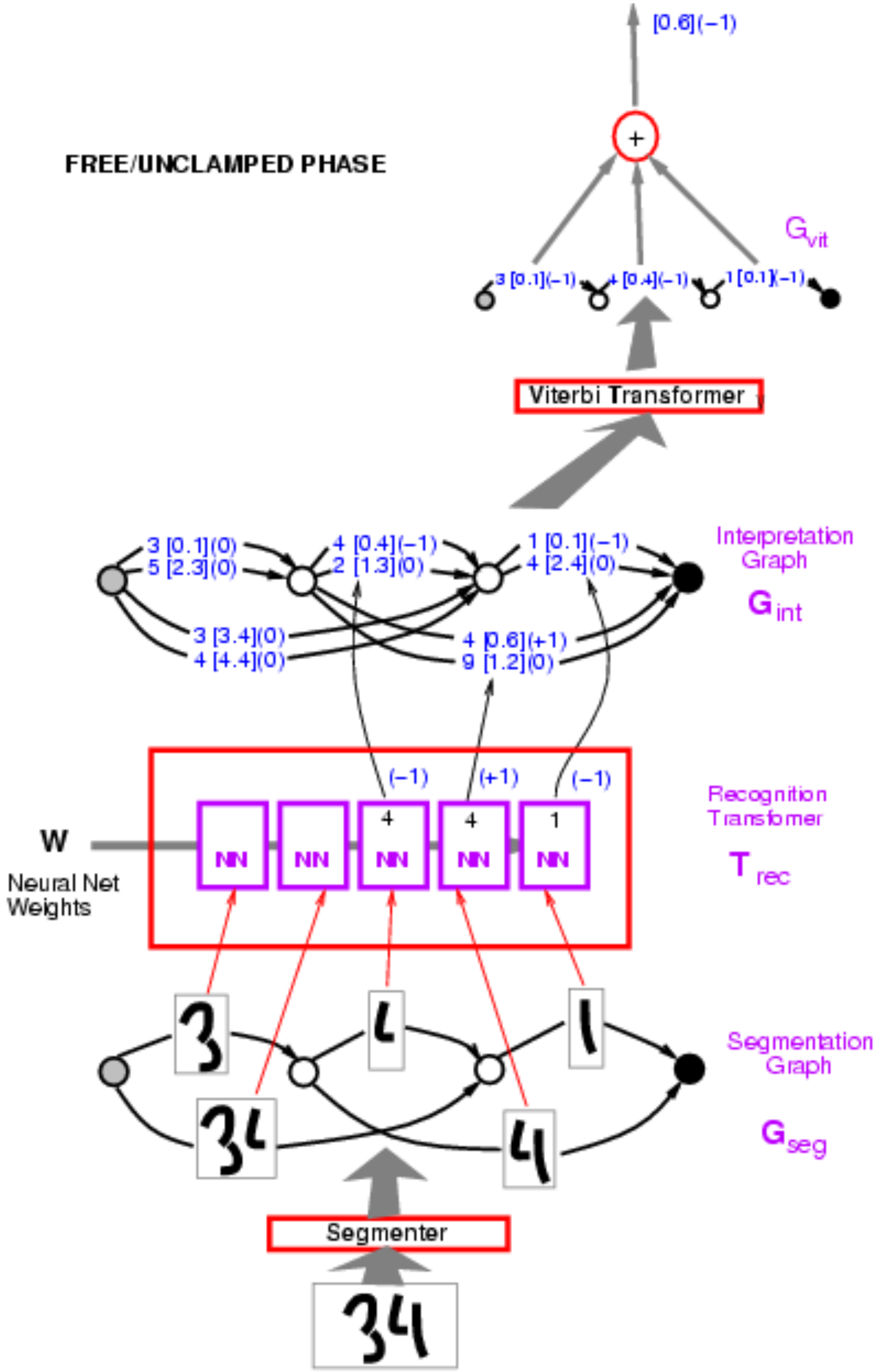
Graph transformer networks

Variables:

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

Loss function: computing the constrastive term:

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

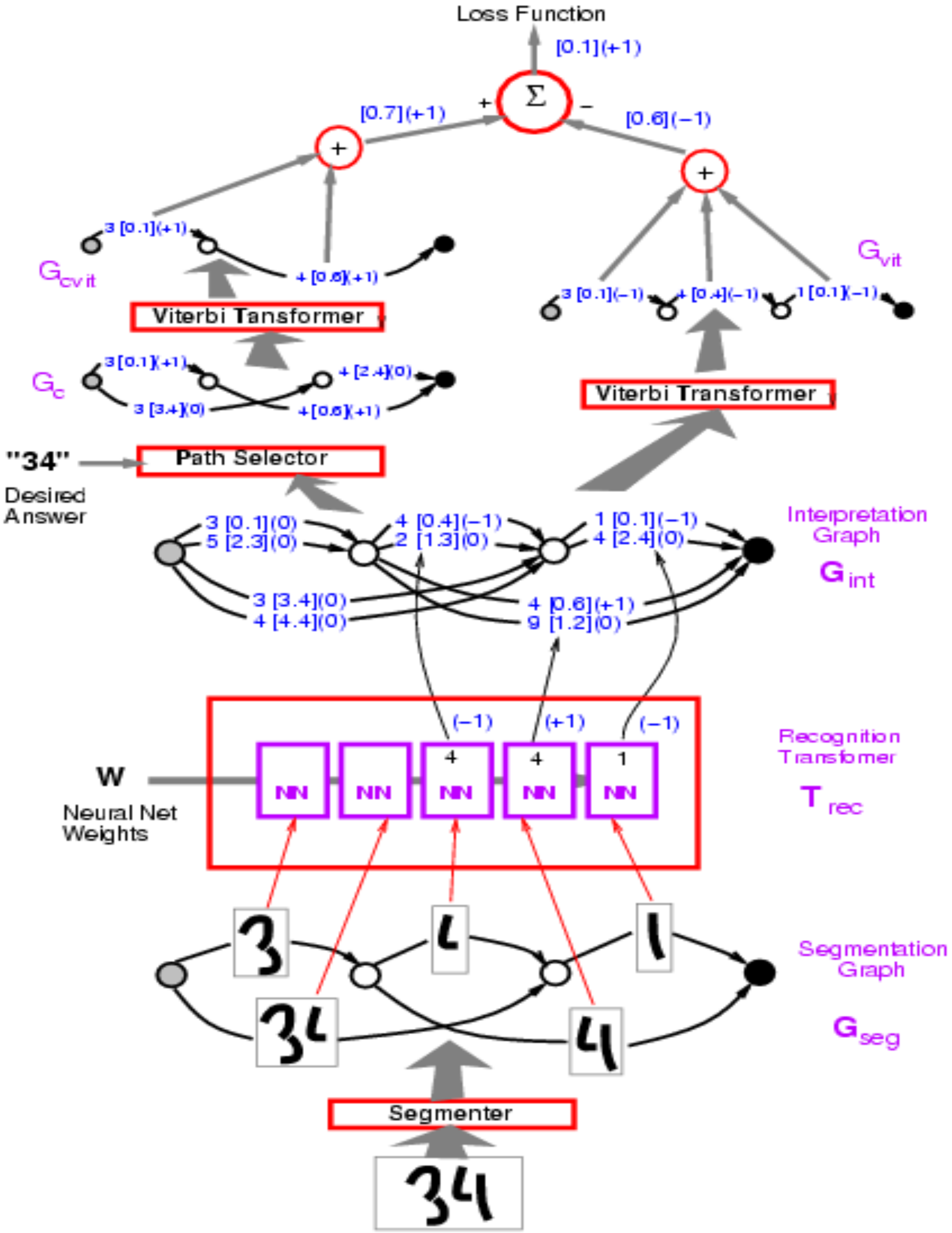


Graph transformer networks

Example: Perceptron loss

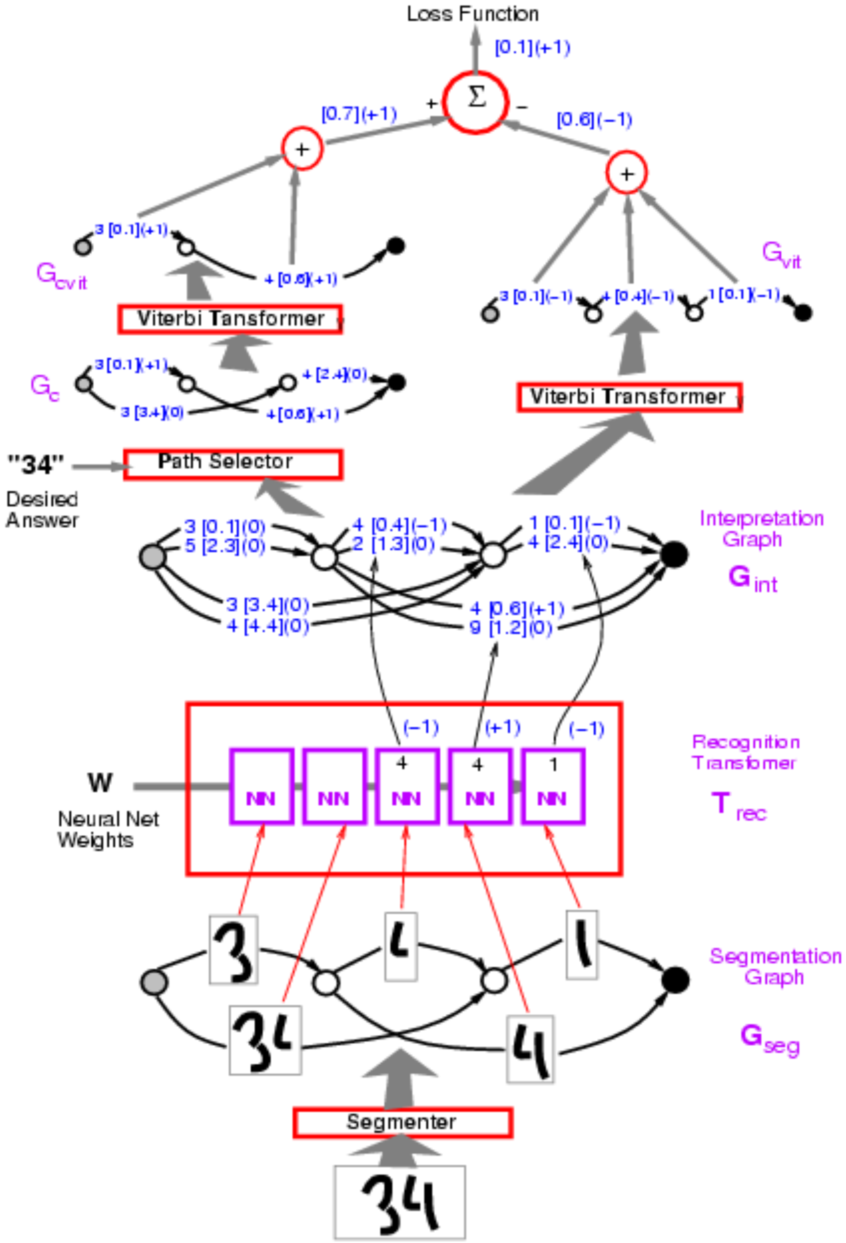
Loss = Energy of desired answer - Energy of best answer.

- (no margin)



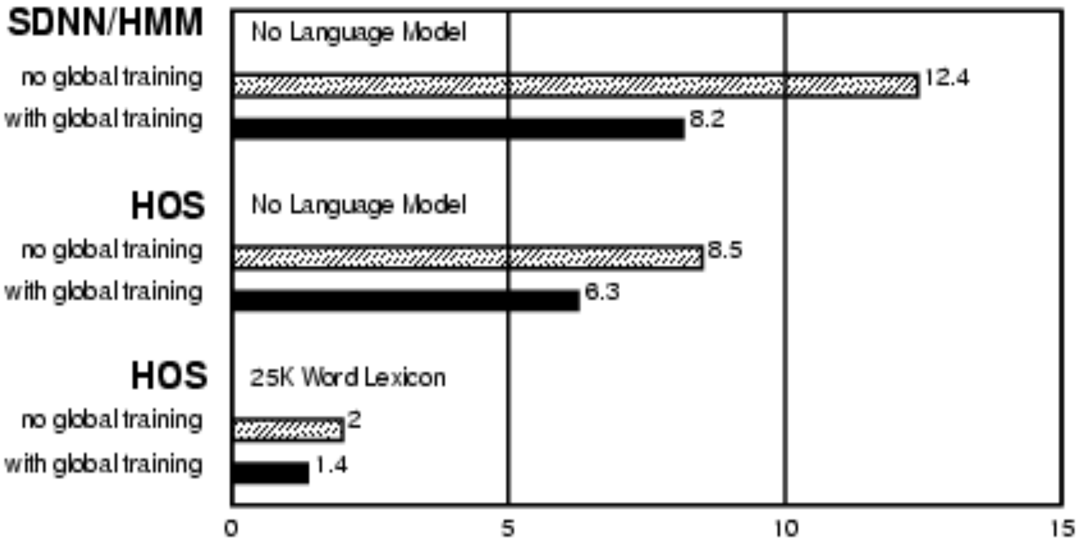
Integrating deep learning and structured prediction

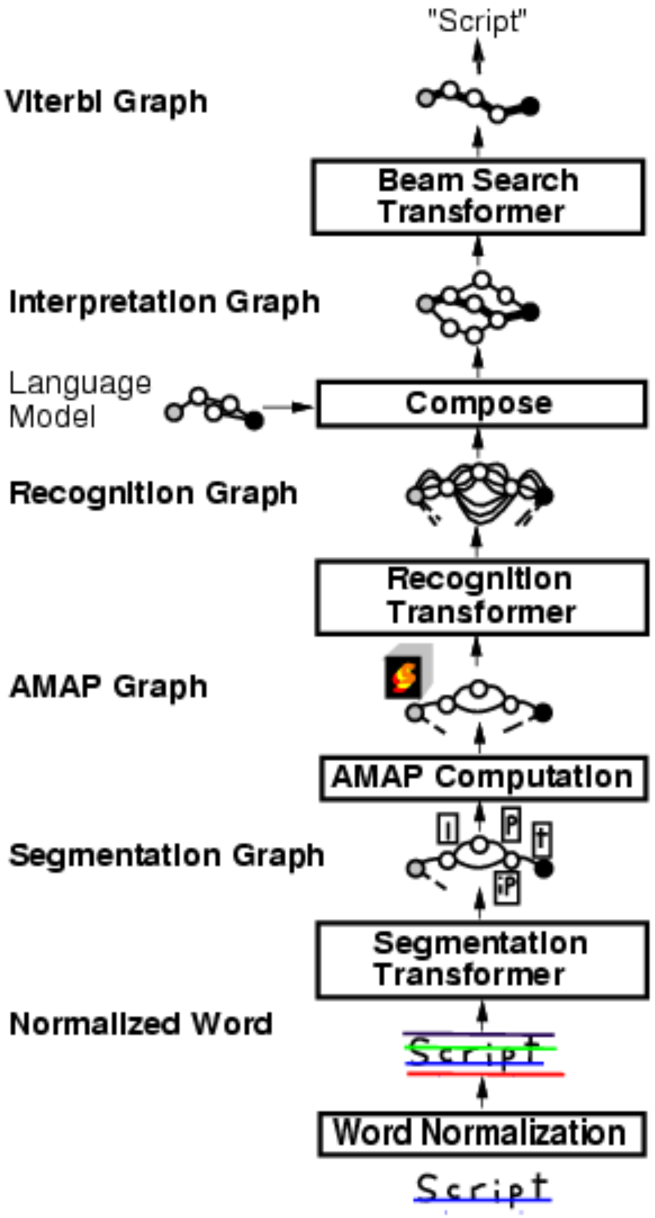
- Structured prediction: when the output is structured: string, graph.....
- Integrating deep learning and structured prediction is an old idea
  - In fact, it predates structured prediction [LeCun, Bottou, Bengio, Haffner 1998]
- Globally-trained convolutional-net + graphical models for handwriting recognition
  - Trained discriminatively at the word level
  - Loss identical to CRF and structured perceptron
  - Compositional movable parts model



Global (word-level) training helps

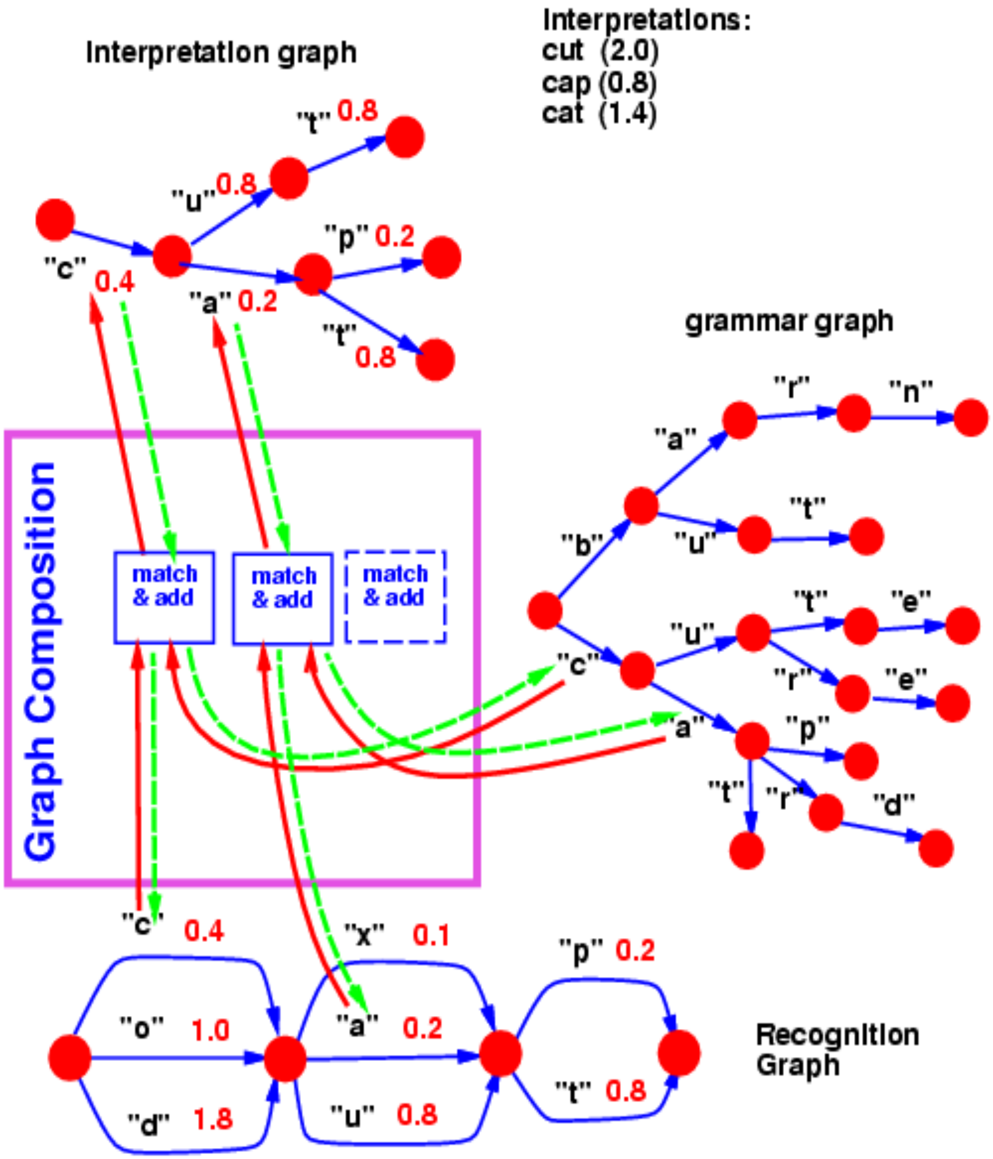
- Pen-based handwriting recognition (for tablet computer)
  - [Bengio&LeCun 1995]





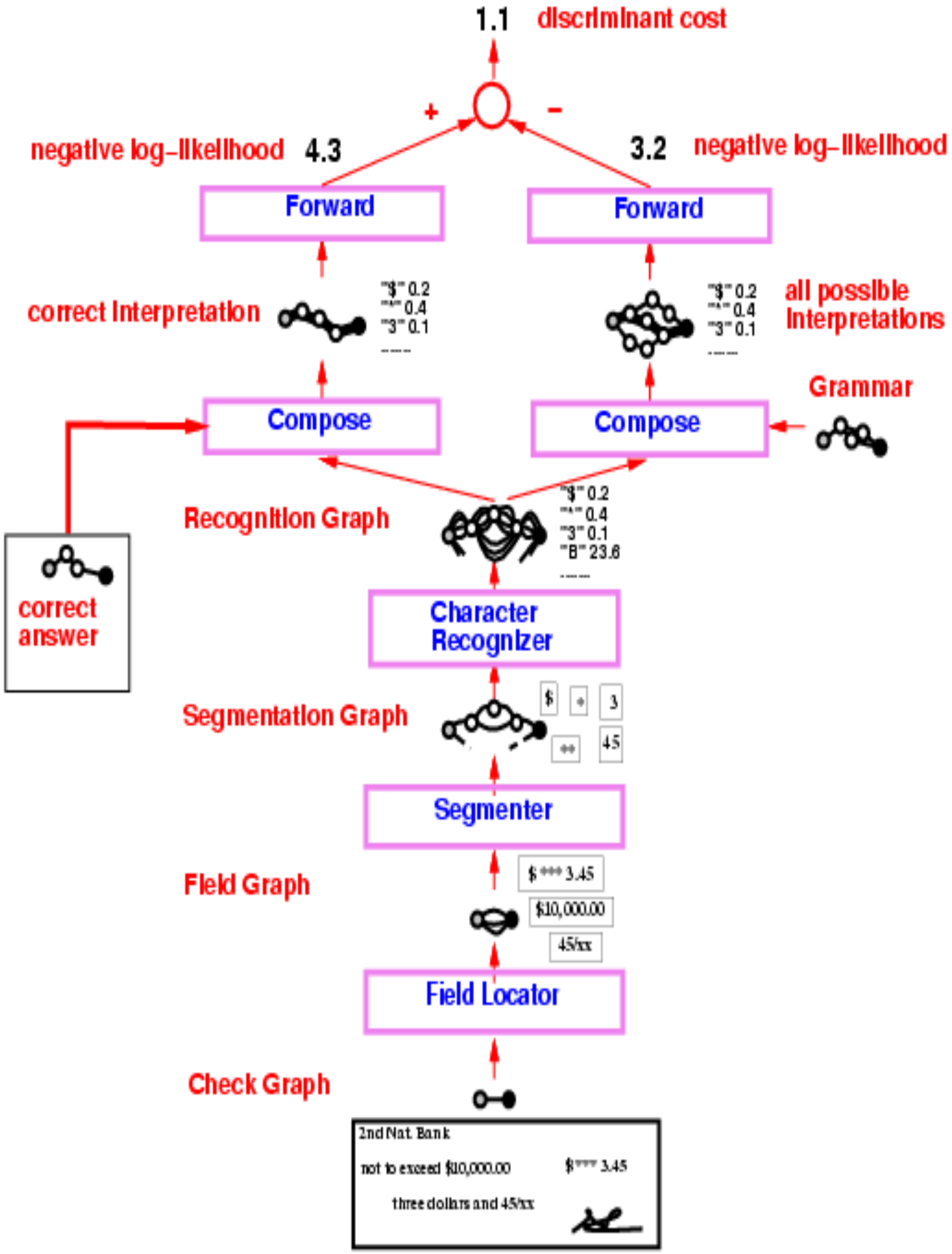
Graph composition, transducers.

- 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, used in many banks in the US and Europe.
- Processes an estimated 10% to 20% of all the checks written in the US.



Deep structured predictions for speech and handwriting

Trainable Speech/Handwriting Recognition systems that integrate Neural Nets (or other “deep” classifiers) with dynamic time warping, Hidden Markov Models, or other graph-based hypothesis representations

Word-level global discriminative training with GMM:

With Minimum Empirical Error loss

- Ljolje and Rabiner (1990)

With MCE

- Juang et al. (1997)

Word-level global discriminative training with ConvNets:

with the LVQ2 Loss:

- Driancourt and Bottou's speech recognizer (1991)

with Neg Log Likelihood (aka MMI):

- Bengio (1992), Haffner (1993), Bourlard (1994)

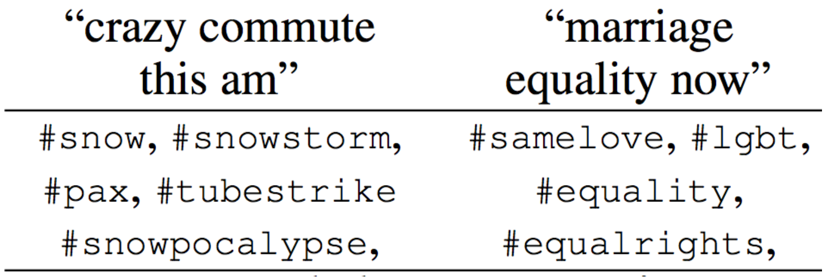
CRF-like Late normalization

- un-normalized HMM
- Bottou pointed out the **label bias problem** (1991)
- Denker and Burges proposed a solution (1995)
- Implemented in (LeCun et al 1998)

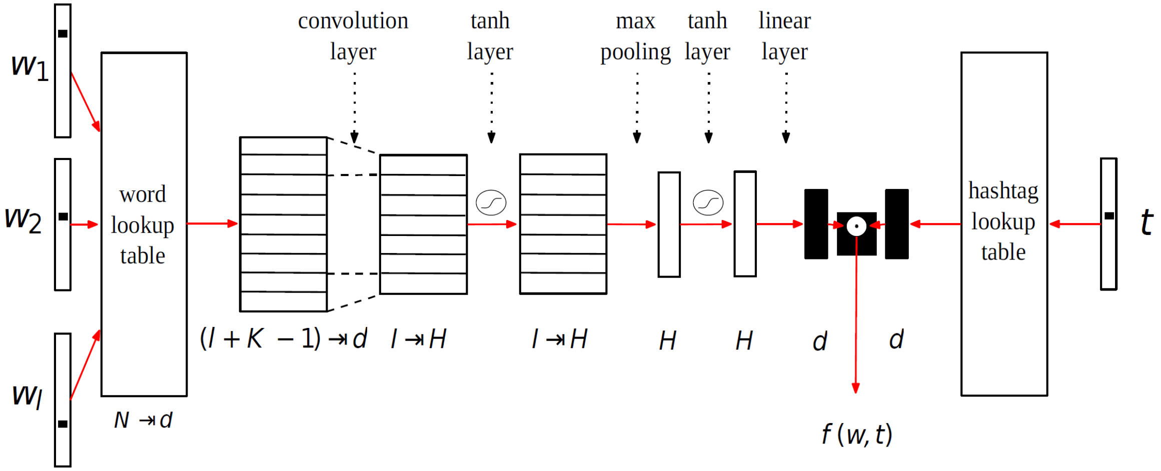
Deep learning in natural language processing research at Facebook

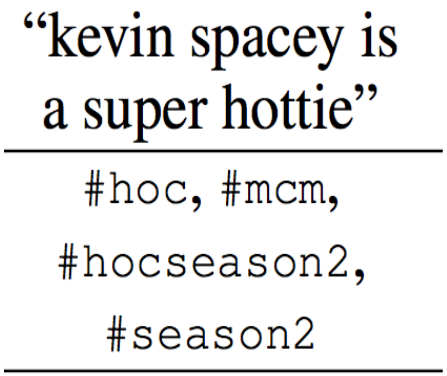
TagSpace: hash tag prediction [Adams et al. 2014]

- We want to learn semantic features for post content.
- Hashtags are a cheap, plentiful labeling of text provided by authors, similar to classical NLP task labels:
- disambiguation (chips #futurism vs. chips #junkfood); identification of named entities (#sf49ers); sentiment (#dislike); and topic annotation (#yoga).
- We train a neural network using hashtags as supervisor training on 5.5 billion words and 100,000 possible hashtags.
- The learned features are useful for other tasks e.g. food



“crazy commute this am”	“marriage equality now”
#snow, #snowstorm,	#samelove, #lgbt,
#pax, #tubestrike	#equality,
#snowpocalypse,	#equalrights,



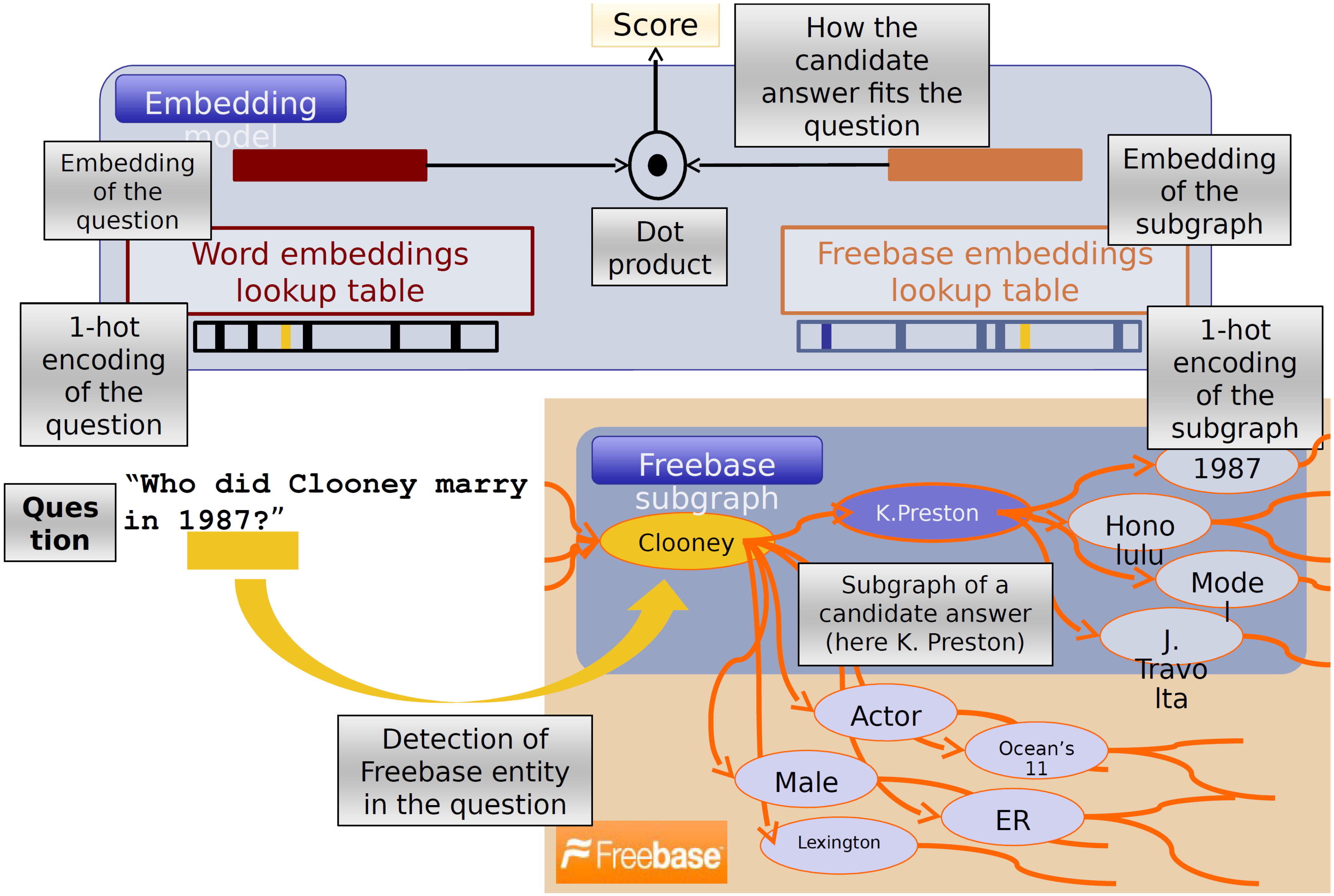


“kevin spacey is a super hottie”
#hoc, #mcm,
#hocseason2,
#season2

Predicting posts and comments with recurrent nets

```
- happy birthday to you ! big 17 ! !
- ready to go on the swim alone ! :)
- the fbi devil hits my folder but really really crazy .
   its not my bro ? ! ?
- happy birthday UNK ! ! i hope you have a great day . .
- i have an urgent candy crush exam . she UNK some awesome service
   on plus june 7th \u0040 UNK camp 2014 center , 8pm central to
   go watch it now ? it 's not hurt . everybody know it 's been
   a long journey . watch it UNK the UNK islands of earth .
   y'all stay safe UNK UNK love ya honey ! thank you so much for
   all the flowers and nurses for me . i will bounce down until
   next temple but please stop following to helping my friend as
   they have no choice like this . \ women are the ones i used to ?
- so proud of jack williams . thanks to all the teachers who have
   served to send respect to every faithfulness and generations
   to us all .
- # UNK UNK insecurity rose so high that our UNK ladies no longer
   have even safe days !
- hey boss ! i was the change of UNK . . we have a party together
   on friday UNK ba ang mga bum ?
```

Question-answering system



How the candidate answer fits the question

Embedding of the question

Embedding of the subgraph

1-hot encoding of the question

1-hot encoding of the subgraph

Question

"Who did Clooney marry in 1987?"

Subgraph of a candidate answer (here K. Preston)

Detection of Freebase entity in the question

Question-answering system

what are bigos?

["stew"] ["stew"]

what are dallas cowboys colors?

[“navy_blue", "royal_blue", "blue", "white", "silver"] ["blue", "navy_blue", "white", "royal_blue", "silver"]

how is egyptian money called?

["egyptian_pound"] ["egyptian_pound"]

what are fun things to do in sacramento ca?

["sacramento_zoo"] ["raging_waters_sacramento", "sutter_s_fort", "b_street_theatre", "sacramento_zoo", "california_state_capitol_museum", ...]

how are john terry's children called?

["georgie_john_terry", "summer_rose_terry"] ["georgie_john_terry", "summer_rose_terry"]

what are the major languages spoken in greece?

["greek_language", "albanian_language"] ["greek_language", "albanian_language"]

what was laura ingalls wilder famous for?

["writer", "author"] ["writer", "journalist", "teacher", "author"]

Question-answering system

who plays sheldon cooper mother on the big bang theory?

["jim_parsons"] ["jim_parsons"]

who does peyton manning play football for?

["denver_broncos"] ["indianapolis_colts", "denver_broncos"]

who did vladimir lenin marry?

["nadezhda_krupskaya"] ["nadezhda_krupskaya"]

where was teddy roosevelt's house?

["new_york_city"] ["manhattan"]

who developed the tcp ip reference model?

["vint_cerf", "robert_e._kahn"] ["computer_scientist", "engineer”]

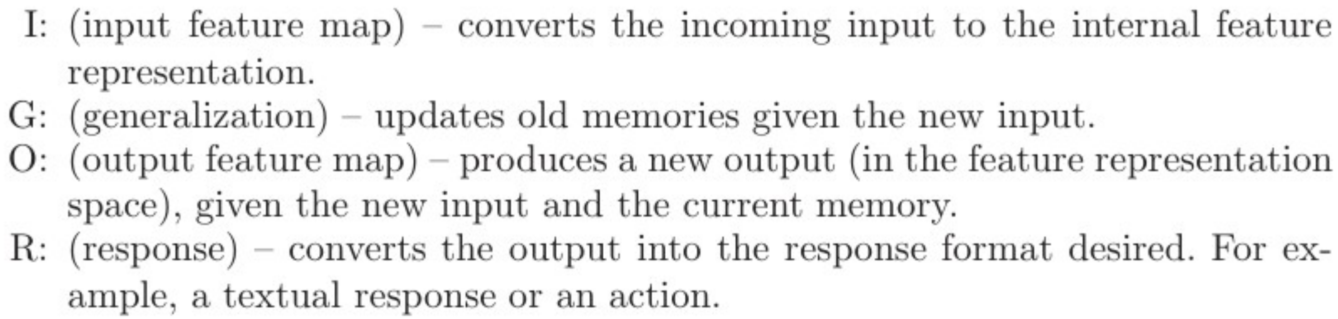
Memory network [Weston, Chopra, Bordes 2014]

Add a short-term memory to a network http://arxiv.org/abs/1410.3916

- Joe went to the garden then Fred picked up the milk; Joe moved to the bathroom and Fred dropped the milk, and then Dan moved to the living room.
  - Where is Dan? A: living room I believe
  - Where is Joe? A: the bathroom
- Fred moved to the bedroom and Joe went to the kitchen then Joe took the milk there and Dan journeyed to the bedroom; Joe discarded the milk.
  - Where is the milk now? A: the milk is in the kitchen
  - Where is Dan now? A: I think he is in the bedroom
- Joe took the milk there, after that Mike travelled to the office, then Joe went to the living room, next Dan went back to the kitchen and Joe travelled to the office.
  - Where is Joe now? A: I think Joe is in the office

Memory network [Weston, Chopra, Bordes 2014]

Add a short-term memory to a network http://arxiv.org/abs/1410.3916

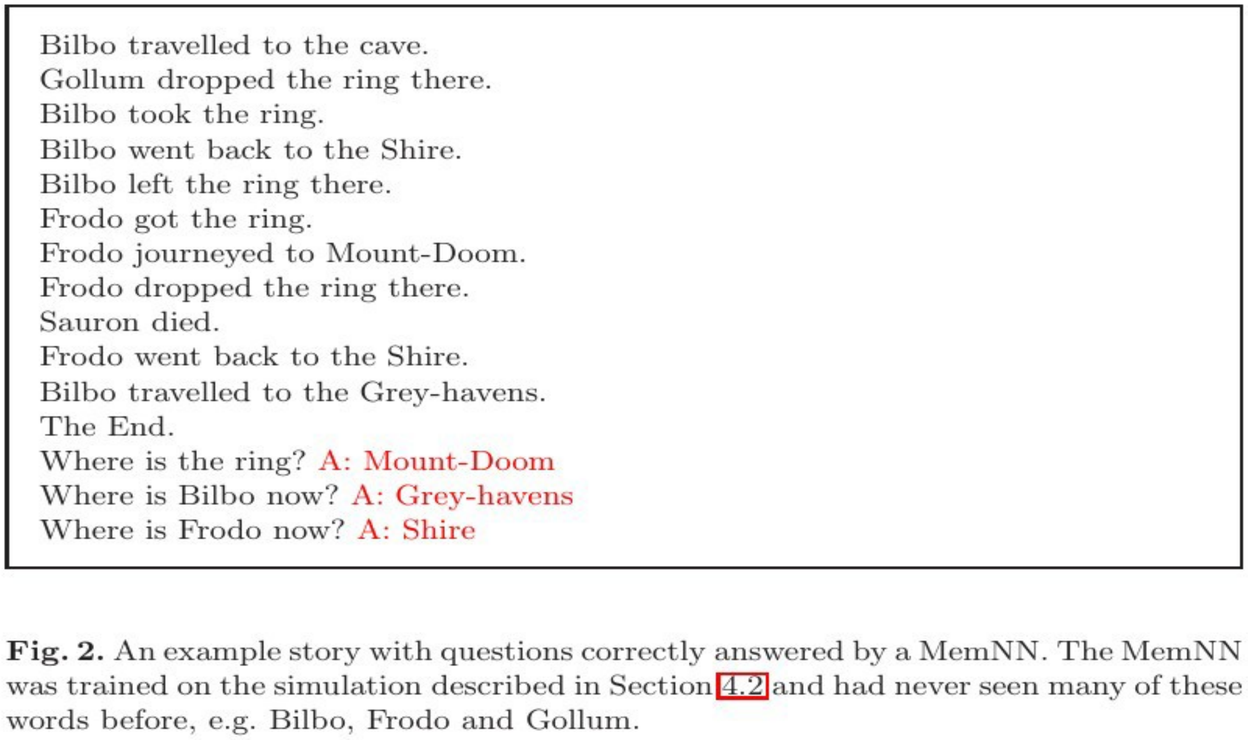


I: (input feature map) – converts the incoming input to the internal feature representation.

G: (generalization) – updates old memories given the new input.

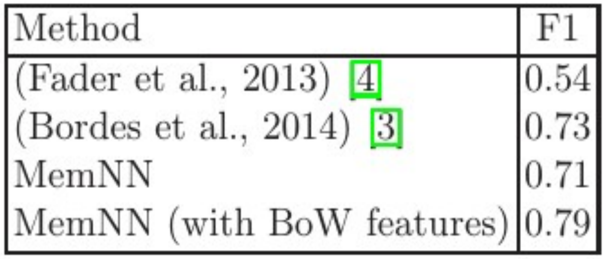
O: (output feature map) – produces a new output (in the feature representation space), given the new input and the current memory.

R: (response) – converts the output into the response format desired. For example, a textual response or an action.



Bilbo travelled to the cave.
Gollum dropped the ring there.
Bilbo took the ring.
Bilbo went back to the Shire.
Bilbo left the ring there.
Frodo got the ring.
Frodo journeyed to Mount-Doom.
Frodo dropped the ring there.
Sauron died.
Frodo went back to the Shire.
Bilbo travelled to the Grey-havens.
The End.
Where is the ring? A: Mount-Doom
Where is Bilbo now? A: Grey-havens
Where is Frodo now? A: Shire

Fig. 2. An example story with questions correctly answered by a MemNN. The MemNN was trained on the simulation described in Section 4.2 and had never seen many of these words before, e.g. Bilbo, Frodo and Gollum.



Method	F1
(Fader et al., 2013) [4]	0.54
(Bordes et al., 2014) [3]	0.73
MemNN	0.71
MemNN (with BoW features)	0.79

Results on question answering task