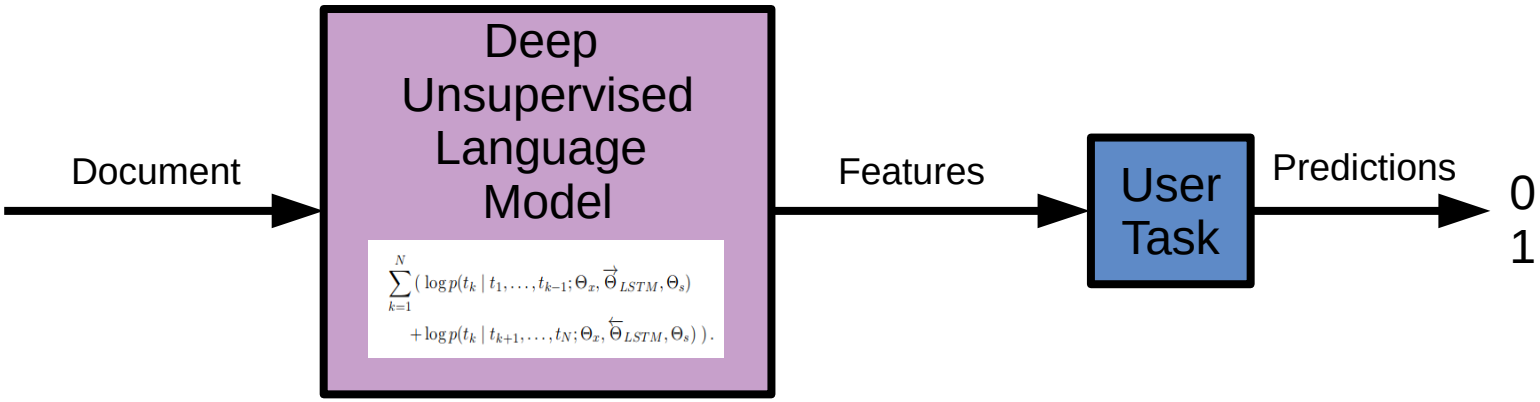
ELMO & BERT

Transfer learning for NLP

UMEÅ UNIVERSITY

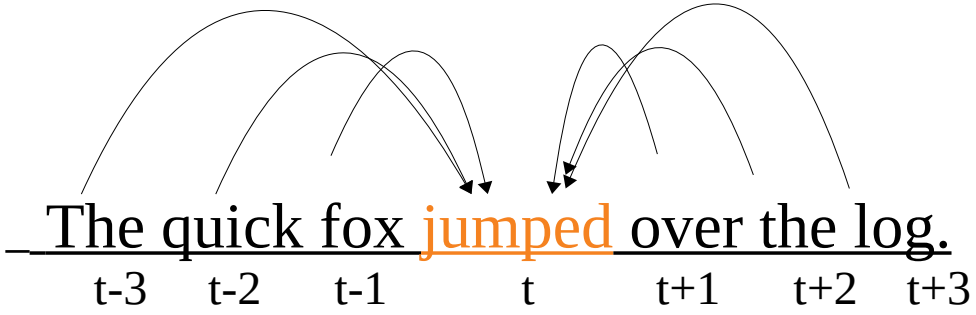
TRANSFER LEARNING



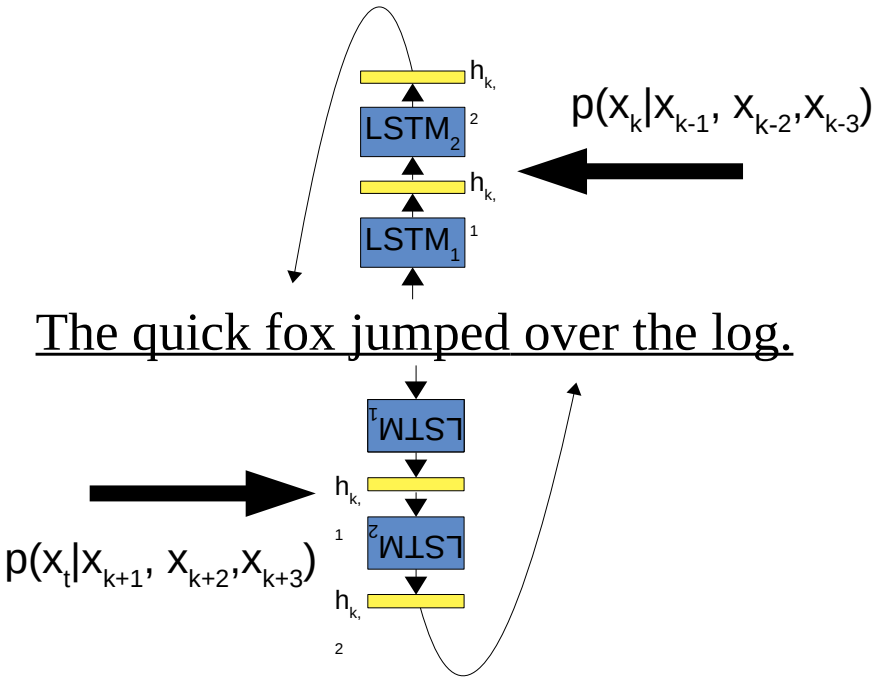
ELMO

Forward language model $p(x_t|x_{t-1}, x_{t-2},x_{t-3})$

Backward language model $p(x_t|x_{t+1}, x_{t+2},x_{t+3})$



ELMO



- At every token x_k

$$R_k = \{\mathbf{x}_k^{LM}, \overrightarrow{\mathbf{h}}_{k,j}^{LM}, \overleftarrow{\mathbf{h}}_{k,j}^{LM} \mid j = 1, \ldots, L\} = \{\mathbf{h}_{k,j}^{LM} \mid j = 0, \ldots, L\},$$

- Loss

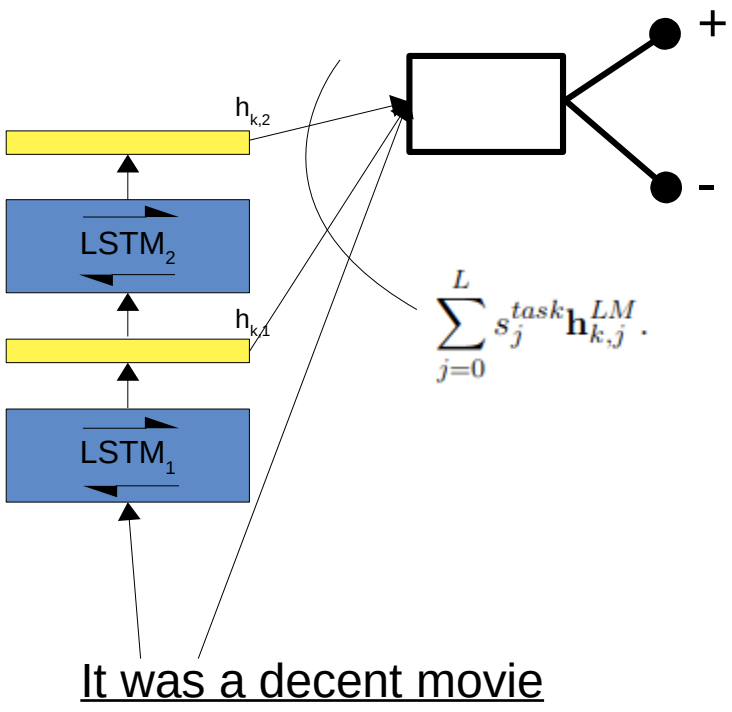
$$\sum_{k=1}^{N} \left(\log p(t_k \mid t_1, \ldots, t_{k-1}; \Theta_x, \overrightarrow{\Theta}_{LSTM}, \Theta_s) + \log p(t_k \mid t_{k+1}, \ldots, t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s) \right).$$

- ELMO can be fine-tuned on different tasks

$$\mathbf{ELMo}_k^{task} = E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM}.$$

  - γ^{task} determines how much a task is important
  - s_j^{task} how layer j is important for *task*

ELMO



Whenever we train a classifier on top of ELMO:

$$\mathbf{ELMo}_k^{task} = E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM} .$$

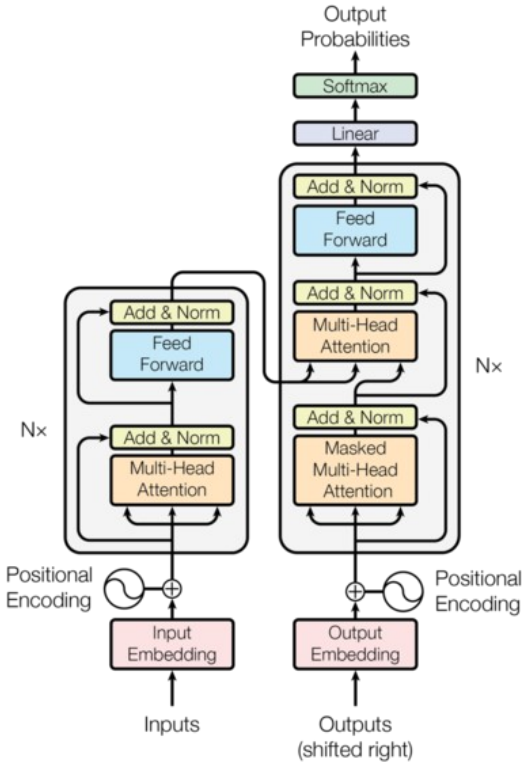
- *task* becomes fixed
- s_j^{task} gets fine-tuned

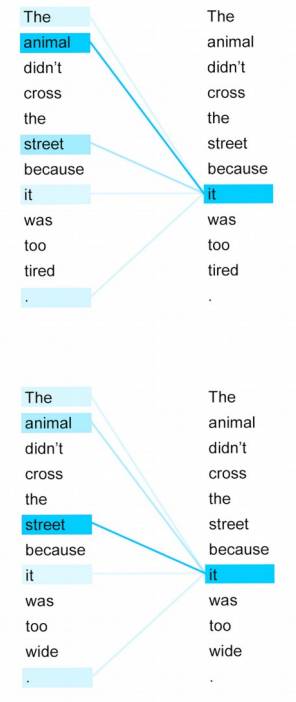
BERT

Is a transformer encoder

Trained on:

- Masked Language Model
  - The ice cream was [MASK] .
- Sentence Adjacency
  - Is B successor of A





UMEÅ UNIVERSITY