Sequence Prediction Models

Loss smoothing & 2D convolutional coding

Maha Elbayad

Supervisors: Laurent Besacier & Jakob Verbeek

June 26, 2018

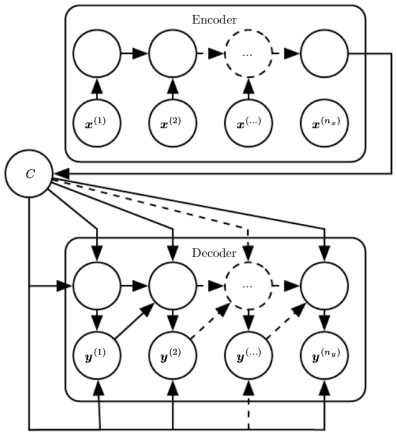
RNN-based encoder-decoder models

For a pair (x, y), we model the conditional distribution:

$$p_\theta(y|x) = \prod_t p_\theta(y^t|x, y^{<t})$$

- (h^t_{enc}) a context to condition the decoder.
- (h^t_{dec}) a predictor for the emission prob.

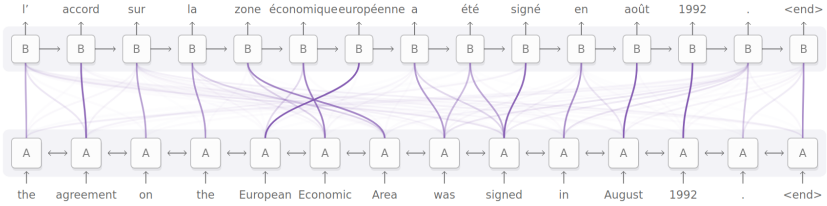
$$h^t = f_\theta(h^{t-1}, y^t) \qquad \text{(recursion)}$$
$$p_\theta(y^{t+1}|h^t) = \sigma(\mathbf{W}h^t) \qquad \text{(prediction)}$$



source: I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press.

Applications

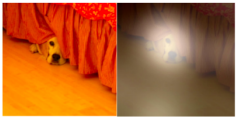
Machine translation



source: distill.pub

Image captioning


A woman is throwing a frisbee in a park.

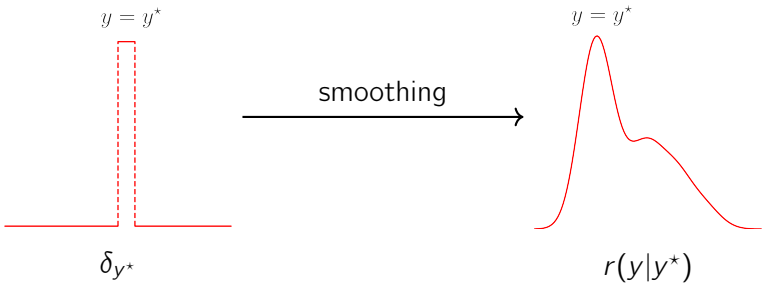

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

OCR, Speech recognition, Times series...

Training objectives: Smoothing

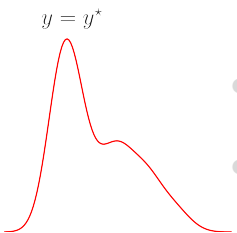
$$\ell_{\mathrm{ML}}(y^\star, x) = -\ln p_\theta(y^\star|x) = D_{\mathrm{KL}}(\delta(y|y^\star)\|p_\theta(y|x)) = \sum_{t=1}^{T} D_{\mathrm{KL}}(\delta(y_t|{y^\star}_t)\|p_\theta(y_t|h_t))$$



- Zero-one Loss, all the outputs $y \neq y^\star$ are treated equally.
- Discrepancy at the sentence level between the training and evaluation reward.

Training objectives: Smoothing

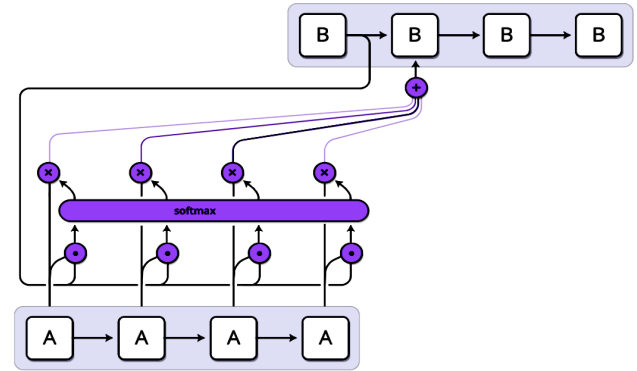
$$\ell_{\text{ML}}(y^\star, x) = -\ln p_\theta(y^\star|x) = D_{\text{KL}}(\delta(y|y^\star)\|p_\theta(y|x)) = \sum_{t=1}^{T} D_{\text{KL}}(\delta(y_t|{y^\star}_t)\|p_\theta(y_t|h_t))$$



- Token-level: smooths w.r.t tokens similarity as assessed by a given word embedding (Word2vec, GloVe,..)
- Sequence-level: smooths w.r.t an evaluation reward via importance sampling.

M. Elbayad, L. Besacier, and J. Verbeek. Token-level and sequence-level loss smoothing for RNN language models. In ACL, 2018

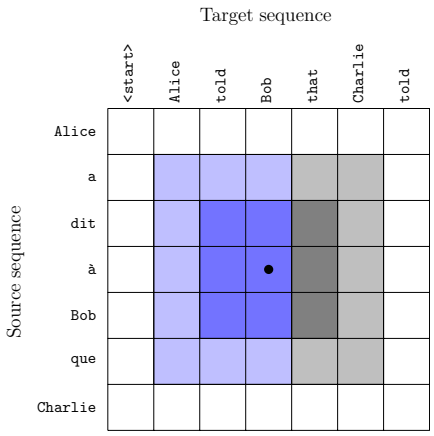
Encoder-decoder with attention



source: distill.pub

- The attending RNN generates a query describing what it wants to focus on and matches it with the source codes.

Our 2D-Convolutional coder



- Every target token has its *own interpretation* of the source tokens.
- Each convolution has the future target tokens masked $\Longrightarrow$ Autoregressive.
- The emission prob. is generated by a **DenseNet**.

Under review.

Thank you for your attention.

Questions?