-SOTA-
Document-level Neural Machine Translation

by Lorenzo Lupo

April 2020

Plan

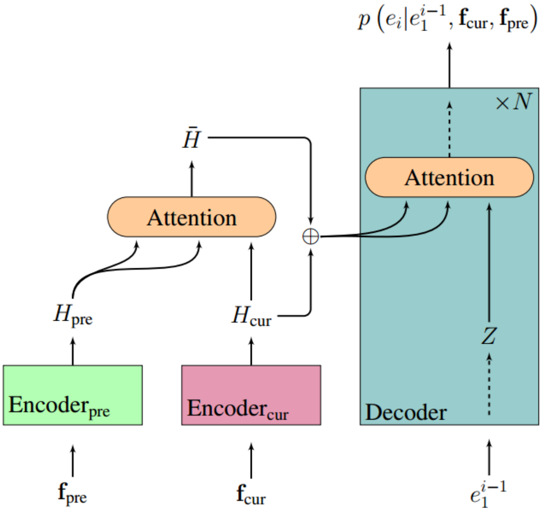
Plan

1. Models
   Remarks and conclusions

Overview

- Single-Encoder Approach
- Multi-Encoder Approach
  - Integration Outside the Decoder
  - Integration Inside the Decoder
    - Sequential Attentions
    - Parallel Attentions

Multi-Encoder - Integration Outside the Decoder



Figure

Multi-Encoder - Integration Outside the Decoder

[Maruf and Haffari, 2018] **Two-pass approach (not training! to explain...). All context (source and target). Fr/De/Et→En. Decoder without attention (integration in the RNN gates).** First pass for sentence representations (filling the memories), second pass for integrating current sentence representation with information stocked into memories (via coarse attention). extention: attention to target memory! FORSE VA MESSO NELLA SEZIONE CACHES

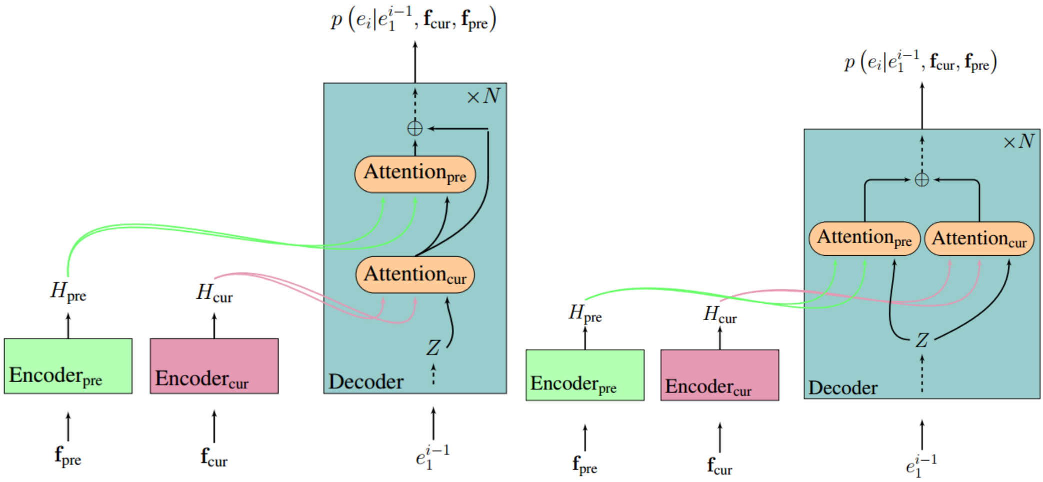
Multi-Encoder - Integration Outside the Decoder

- encoders might encode multiple previous sentences. E.g. [Wang et al., 2017].
- architectures might be RNNs (e.g. [Wang et al., 2017]) or Transformers (e.g. [Zhang et al., 2018])
- integration inside the decoder might happen with a different system than cross-attention. E.g. [Wang et al., 2017] propose to concatenate the context representation to the cell state of the decoder's RNN.
- source-side attention to context can be at both sentence and word level. E.g. [Maruf et al., 2019, Miculicich et al., 2018].
- gating context is a way substitutes residual add.
- despite some have considered target-side context harmful because of the *error propagation* problem [Zhang et al., 2018], now...
- weight sharing is blabla [Voita et al., 2018]. Some use it. In general, it has been proven to be successful by a comparative study [Yamagishi and Komachi, 2019].
- two-step training what is. E.g. [Zhang et al., 2018, Miculicich et al., 2018]. Explain that DL training corpus is small! possible future direction...

Multi-Encoder - Integration Outside the Decoder

Reference	Context	Two-Pass Approach	Outside Integr.	Inside Integr.	Lang. Pair
[Wang et al., 2017]	s:-3		optional	optional	Zh→En
[Voita et al., 2018]	s:-1		yes		En→Ru
[Zhang et al., 2018]	s:-2		yes	sequential	Zh→En
[Miculicich et al., 2018]	s:-3; t:-3		yes		Zh/Es→En
[Maruf et al., 2019]	s:all; t:all	optional	yes		En→De
[Jean et al., 2017]	s:-1			parallel	En→De/Fr
[Bawden et al., 2018]	s:-1; t:-1			parallel	En→Fr
[Fu et al., 2019]	s:all; t:-1	yes		parallel	En/Zh→De/En

Multi-Encoder - Integration Inside the Decoder



Figure

Remarks and conclusions

Possible Future Research Directions

- build a large DL corpus for training systems;

Thank you for your attention!

References I

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018).
Evaluating Discourse Phenomena in Neural Machine Translation.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.
00055.

Fu, H., Liu, C., and Sun, J. (2019).
Reference Network for Neural Machine Translation.
In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3002–3012, Florence, Italy. Association for Computational Linguistics.
00000.

References II

Jean, S., Lauly, S., Firat, O., and Cho, K. (2017).
Does Neural Machine Translation Benefit from Larger Context?
arXiv:1704.05135 [cs, stat].
00038 arXiv: 1704.05135.

Maruf, S. and Haffari, G. (2018).
Document Context Neural Machine Translation with Memory Networks.
In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia.
Association for Computational Linguistics.
00032.

Maruf, S., Martins, A. F. T., and Haffari, G. (2019).
Selective Attention for Context-aware Neural Machine Translation.
arXiv:1903.08788 [cs].
00012.

References III

Miculicich, L., Ram, D., Pappas, N., and Henderson, J. (2018).
Document-Level Neural Machine Translation with Hierarchical Attention Networks.
arXiv:1809.01576 [cs].
00024 arXiv: 1809.01576.

Voita, E., Serdyukov, P., Sennrich, R., and Titov, I. (2018).
Context-Aware Neural Machine Translation Learns Anaphora Resolution.
In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.
00047.

References IV

Wang, L., Tu, Z., Way, A., and Liu, Q. (2017).
Exploiting Cross-Sentence Context for Neural Machine Translation.
In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.
00045.

Yamagishi, H. and Komachi, M. (2019).
Improving Context-aware Neural Machine Translation with Target-side Context.
arXiv:1909.00531 [cs].
00001 arXiv: 1909.00531.

References V

Zhang, J., Luan, H., Sun, M., Zhai, F., Xu, J., Zhang, M., and Liu, Y. (2018).
Improving the Transformer Translation Model with Document-Level Context.
In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.
00028.

Markov Decision Processes

Reinforcement Learning

General class of algorithms that allow an agent to learn how to behave in a stochastic and possibly unknown environment by trial-and-error.

Markov Decision Process (MDP)

stochastic dynamical system specified by $< \mathbb{S}, \mathbb{A}, \mathcal{P}, \mathcal{R}, \gamma >$

1. $(\mathbb{S}, \mathcal{S})$ is a measurable state space
2. $(\mathbb{A}, \mathcal{A})$ is a measurable action space
3. $\mathcal{P} : \mathbb{S} \times \mathbb{A} \times \mathcal{S} \to \mathbb{R}$ is a Markov transition kernel
4. $\mathcal{R} : \mathbb{S} \times \mathbb{A} \to \mathbb{R}$ is a reward function
5. $0 < \gamma < 1$ is the discount factor.

Monte-Carlo Policy Gradient: Pseudocode

Input: Stochastic policy π_θ, Initial parameters θ_0, learning rate $\{\alpha_k\}$
Output: Approximation of the optimal policy $\pi_{\theta*} \approx \pi_*$

1: **repeat**
2: Sample M trajectories $h^{(m)} = \{(s_t^{(m)}, a_t^{(m)}, r_{t+1}^{(m)})\}_{t=0}^{T^{(m)}}$ under policy π_{θ_k}
3: Approximate policy gradient

$$\nabla_\theta J(\theta_k) \approx \frac{1}{M} \sum_{m=0}^{M} \sum_{u=0}^{T^{(m)}-1} \nabla_\theta \log \pi_{\theta_k} \left(s_u^{(m)}, a_u^{(m)}\right) \sum_{v \geqslant u}^{T^{(m)}-1} \gamma^{v-u} r_{v+1}^{(m)}$$

4: Update parameters using gradient ascent $\theta_{k+1} = \theta_k + \alpha_k \nabla_\theta J(\theta_k)$
5: $k \leftarrow k + 1$
6: **until** converged

Episodic PGPE Algorithm: Pseudocode

Input: Controller F_θ, hyper-distribution p_ξ, initial guess ξ_0, learning rate $\{\alpha_k\}$
Output: Approximation of the optimal policy $F_{\xi*} \approx \pi_*$

repeat
 for $m = 1, \dots, M$ **do**
  Sample controller parameters $\theta^{(m)} \sim p_{\xi_k}$
  Sample trajectory $h^{(m)} = \{(s_t^{(m)}, a_t^{(m)}, r_{t+1}^{(m)})\}_{t=0}^{T^{(m)}}$ under policy $F_{\theta^{(m)}}$
 end for
 Approximate policy gradient

$$\nabla_\xi J(\xi_k) \approx \frac{1}{M} \sum_{m=1}^{M} \nabla_\xi \log p_\xi \left(\theta^{(m)}\right) \left[G\left(h^{(m)}\right) - b\right]$$

 Update hyperparameters using gradient ascent $\xi_{k+1} = \xi_k + \alpha_k \nabla_\xi J(\xi_k)$
 $k \leftarrow k + 1$
until converged

Truncated Multiple Importance Sampling Estimator

Importance Sampling

Given a bounded function $f : \mathcal{Z} \to \mathbb{R}$, and a set of i.i.d. outcomes $z_1, \ldots, z_N$ sampled from Q, the importance sampling estimator of $\mu := \underset{z \sim P}{\mathbb{E}} [f(z)]$ is:

$$\widehat{\mu}_{\text{IS}} = \frac{1}{N} \sum_{i=1}^{N} f(z_i) w_{P/Q}(z_i), \tag{1}$$

which is an unbiased estimator, i.e., $\underset{z_i \overset{\text{iid}}{\sim} Q}{\mathbb{E}} [\widehat{\mu}_{IS}] = \mu$.

Truncated Estimator With Balance Heuristic

$$\check{\mu}_{\text{BH}} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N_k} \min \left\{ M, \frac{p(z_{ik})}{\sum_{j=1}^{K} \frac{N_j}{N} q_j(z_{ik})} \right\} f(z_{ik}). \tag{2}$$

OPTIMIST2

Theorem

regretdiscretized Let $\mathcal{X}$ be a d-dimensional compact arm set with $\mathcal{X} \subseteq [-D, D]^d$. For any $\kappa \geqslant 2$, under Assumptions 1 and 2, OPTIMIST2 with confidence schedule $\delta_t = \frac{6\delta}{\pi^2 t^2 \left(1 + \lceil t^{1/\kappa} \rceil^d\right)}$ and discretization schedule $\tau_t = \lceil t^{\frac{1}{\kappa}} \rceil$ guarantees, with probability at least $1 - \delta$:

$$\begin{aligned} Regret(T) \leqslant \Delta_0 + C_1 T^{\left(1-\frac{1}{\kappa}\right)d} + C_2 T^{\frac{1}{1+\epsilon}} \\ \cdot \left[v_\epsilon \left((2 + {}^{d}/_{\kappa}) \log T + d \log 2 + \log \frac{\pi^2}{3\delta} \right) \right]^{\frac{\epsilon}{1+\epsilon}}, \end{aligned}$$

where $C_1 = \frac{\kappa}{\kappa - 1} LD$, $C_2 = (1 + \epsilon) \left(2\sqrt{2} + \frac{5}{3} \right) \|f\|_\infty$, and Δ_0 is the instantaneous regret of the initial arm $\boldsymbol{x}_0$.