Fuzzy Alignments in Directed Acyclic Graph for Non-Autoregressive Machine Translation

Zhengrui Ma · Chenze Shao · Shangtong Gui · Min Zhang · Yang Feng

Paper Code Group

Natural Language Processing Group
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Science

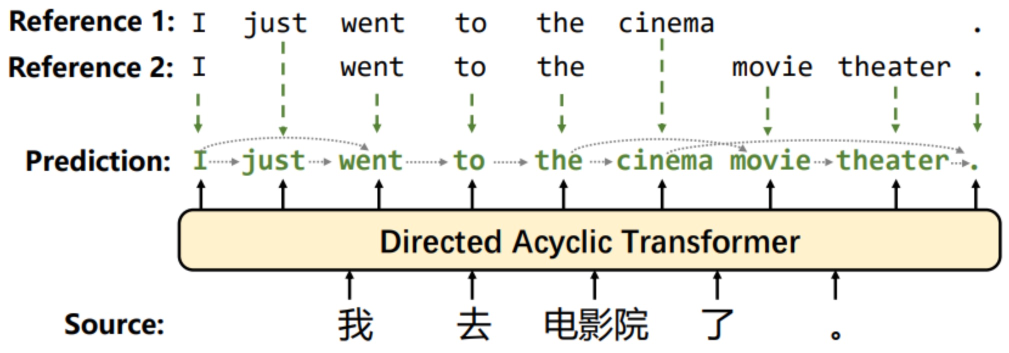
ICLR

Non-autoregressive Translation

- Autoregressive generation
  - $P_\theta(y|x) = \prod_{i=1}^{M} P_\theta(y_i|x, y_{<i})$
- Non-autoregressive generation
  - Conditional independence assumption $P_\theta(y|x) = \prod_{i=1}^{M} P_\theta(y_i|x)$
  - Parallel generation: $\approx 13\times$ speedup
- Performance degradation
  - Inaccurate length prediction
  - Token repetition and omission

Directed Acyclic Transformer

- Directed acyclic decoder
  - Vertices: tokens
  - Transitions: token dependency
  - SOTA NAT
- Distribution modeling
  - Consider all paths aligned with target (Γ_y)
  - $P_\theta(y|x) = \sum_{a \in \Gamma_y} P_\theta(y|a,x) P_\theta(a|x)$
- Problem in NLL training
  - Paths are treated differently!
  - $\frac{\partial \mathcal{L}_{NLL}}{\partial \theta} = \sum_{a \in \Gamma_y} P_\theta(a|y,x) \frac{\partial \mathcal{L}_a}{\partial \theta}, \mathcal{L}_a = -\log P_\theta(y,a|x)$
  - Translations in other modalities will be poorly calibrated!



(Image credits to Huang et al., 2022)

Huang, Fei, et al. "Directed acyclic transformer for non-autoregressive machine translation." International Conference on Machine Learning. PMLR, 2022.

Treatment: Fuzzy Alignment Training

- From verbatim alignment to fuzzy alignment
- How to measure the alignment quality?
  - Sentence against Sentence: Clipped n-gram precision

$$p_n(y', y)$$

  - Path against Sentence: Expected n-gram precision of a path

$$p_n(\theta, a, y) = \mathbb{E}_{y' \sim P_\theta(y'|a,x)}[p_n(y', y)]$$

  - Graph against Sentence: Average all the path

$$p_n(\theta, y) = \mathbb{E}_{a \sim P_\theta(a|x)}[p_n(\theta, a, y)]$$

Estimate Fuzzy Alignment

- Fuzzy alignment objective

$$p_n(\theta, y) = \mathbb{E}_{a \sim P_\theta(a|x)}[p_n(\theta, a, y)]$$

- Impractical for training
  - Spaces of translations and paths both exponentially large!!!
- Approximation
  - Estimate the ratio of the clipped expected count of n-gram matching to expected number of n-grams

$$p'_n(\theta, \boldsymbol{y}) = \frac{\sum_{\boldsymbol{g} \in G_n(\boldsymbol{y})} \min(\mathbb{E}_{\boldsymbol{y}'}[C_{\boldsymbol{g}}(\boldsymbol{y}')], C_{\boldsymbol{g}}(\boldsymbol{y}))}{\mathbb{E}_{\boldsymbol{y}'}[\sum_{\boldsymbol{g} \in G_n(\boldsymbol{y}')} C_{\boldsymbol{g}}(\boldsymbol{y}')]}$$

Efficient Estimation

- Main challenge in optimizing $p'_n(\theta, y)$
  - Intractable $\mathbb{E}_{y'}[C_g(y')]$ and $\mathbb{E}_{y'}[\sum_{g \in G_n(y')} C_g(y')]$
- We find those terms
  - are connected with transition probabilities of a Markov chain (*Eq. 17 & 18*)
  - can be calculated with $O(n + L)$ parallel operations (*Alg. 1*)
- $p'_n(\theta, y)$ is practical for training!

Experiments

- FA-DAT: Fuzzy-Aligned Directed Acyclic Transformer
  - Substantially improves translation quality
  - Further narrows the gap between AT and NAT
  - Sets new SOTA of NAT without knowledge distillation

Model		Iter.	Speedup	WMT14 EN-DE	WMT14 DE-EN	WMT17 ZH-EN	WMT17 EN-ZH
Transformer (Vaswani et al., 2017)		N	1.0×	27.6	31.4	23.7	34.3
Transformer†		N	1.0×	27.54	31.55	24.23	35.19
DA-Transformer†	+ *Greedy*	1	14.2×	26.07	30.69	22.35	33.58
	+ *Lookahead*	1	14.0×	26.56	30.81	22.65	33.62
	+ *Joint-Viterbi*	1	13.2×	26.89	31.09	23.17	33.25
FA-DAT	+ *Greedy*	1	14.2×	27.49**	31.36**	23.78**	33.97*
	+ *Lookahead*	1	14.0×	**27.53****	31.37**	23.81**	34.02**
	+ *Joint-Viterbi*	1	13.2×	27.47**	**31.44***	**24.22****	**34.49****

Codes: https://github.com/ictnlp/FA-DAT

Thank you!

Paper

Code

Group