Efficient Wait-k Models for Simultaneous Machine Translation

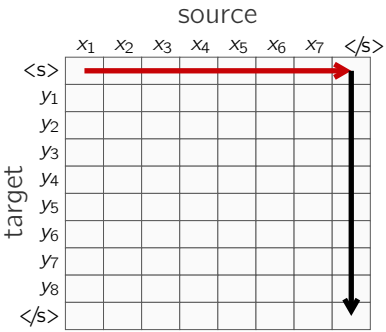
Maha Elbayad[1,2] Laurent Besacier[1] Jakob Verbeek[3]

(1) Université Grenoble Alpes

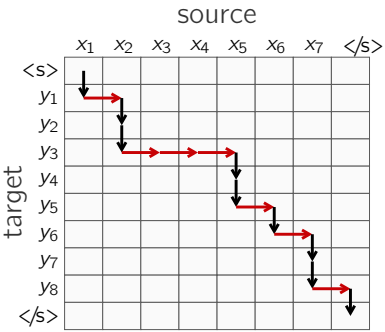
(2) informatiques mathématiques

(3) FACEBOOK AI

Simultaneous Neural Machine Translation

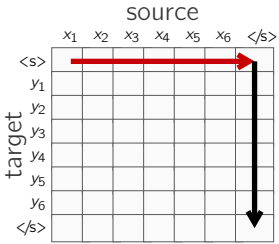


Offline translation

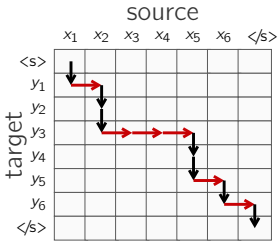


Simultaneous translation

The decoding path z



Offline translation



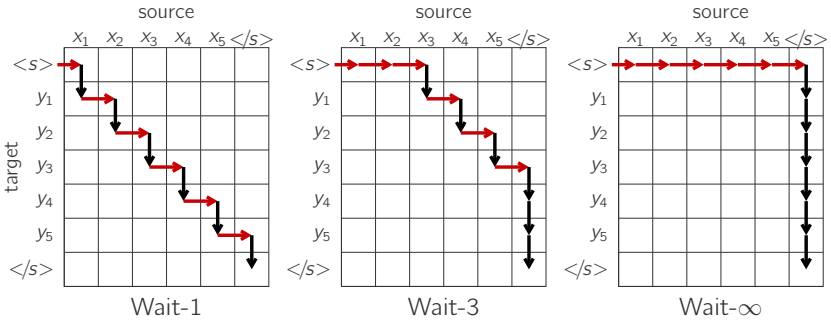
Simultaneous translation

$$\mathrm{p}_\theta(\mathbf{y} \mid \mathbf{x}) = \prod_{t=1}^{|\mathbf{y}|} \mathrm{p}_\theta(y_t \mid \mathbf{y}_{<t}, \mathbf{x}).$$

$$\mathrm{p}_\theta(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) = \prod_{t=1}^{|\mathbf{y}|} \mathrm{p}_\theta(y_t \mid \mathbf{y}_{<t}, \mathbf{x}_{\leq z_t}, \mathbf{z}_{\leq t}).$$

The decoding path z - wait-k

$$\forall t \in [1..|\mathbf{y}|], \quad z_t^{\text{wait-}k} = \min(k + t - 1, |\mathbf{x}|).$$



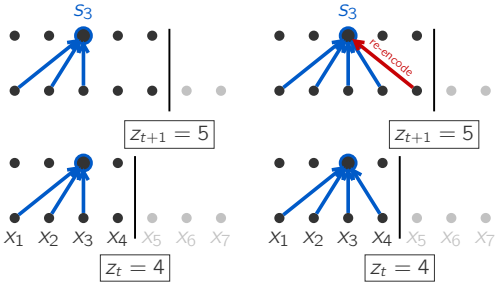
Wait-k or prefix-to-prefix decoding (Dalvi *et al.* 2018; Ma *et al.* 2019).

Adapting Translation Models for Simultaneous Translation

1. Transformer (Vaswani *et al.* 2017)
2. Pervasive Attention (Elbayad *et al.* 2018)

Simultaneous Transformer

- **Encoder self-attention:**
  Unidirectional or bidirectional.



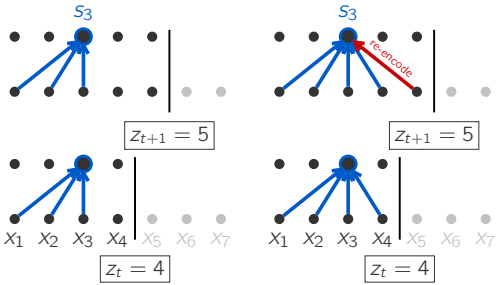
Uni-directional

Bi-directional

(Ma *et al.* 2019)

Simultaneous Transformer

- **Encoder self-attention:**
  Unidirectional or bidirectional.

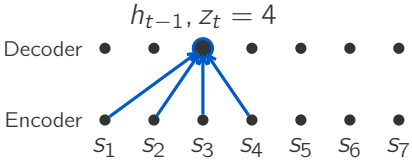


Uni-directional

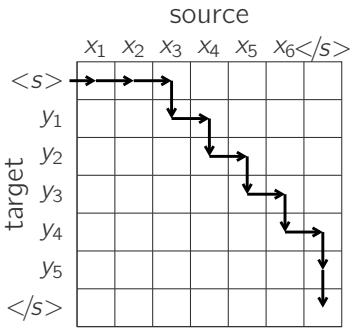
Bi-directional

(Ma *et al.* 2019)

- **Encoder-decoder interaction:**
  Masking the attention energies wrt. $\mathbf{z}_t$.

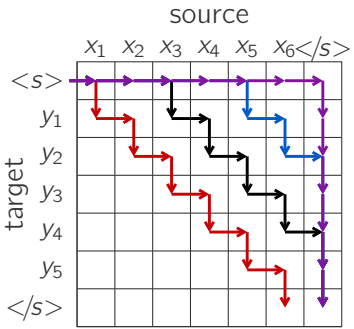


Simultaneous Transformer



- A single wait-k path (Dalvi *et al.* 2018; Ma *et al.* 2019)

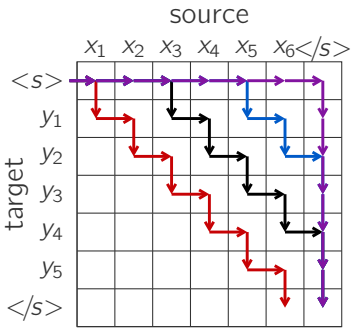
Simultaneous Transformer



- A single wait-k path (Dalvi *et al.* 2018; Ma *et al.* 2019)
- Jointly optimized wait-k paths

$$\mathbb{E}_{\mathrm{K}}\left[\log p(\mathbf{y} \mid \mathbf{x}, \mathbf{z}^{\text{wait-}k})\right] \approx \sum_{k \sim \mathcal{U}(\mathrm{K})} \log p_\theta(\mathbf{y} | \mathbf{x}, \mathbf{z}^{\text{wait-}k}).$$

Simultaneous Transformer

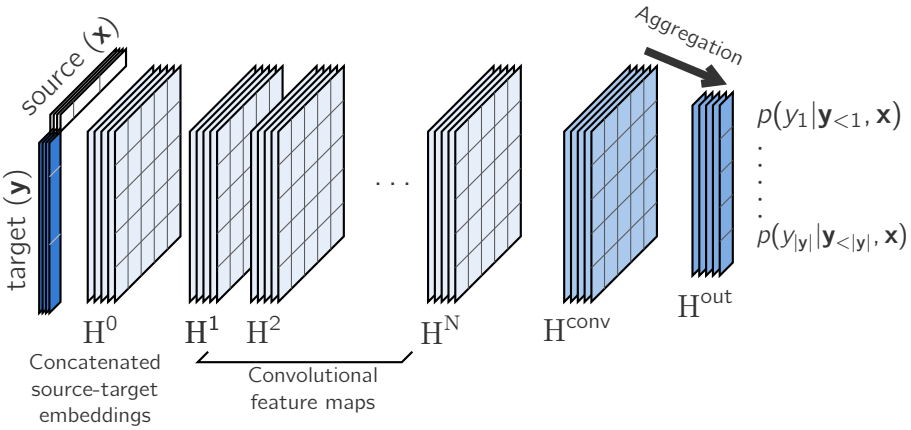


- A single wait-k path (Dalvi *et al.* 2018; Ma *et al.* 2019)
- Jointly optimized wait-k paths

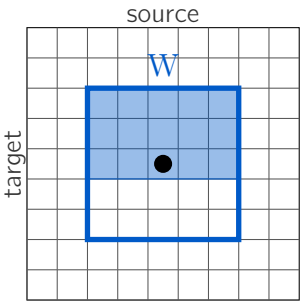
$$\mathbb{E}_{\mathrm{K}}\left[\log p(\mathbf{y} \mid \mathbf{x}, \mathbf{z}^{\text{wait-}k})\right] \approx \sum_{k \sim \mathcal{U}(\mathrm{K})} \log p_\theta(\mathbf{y} | \mathbf{x}, \mathbf{z}^{\text{wait-}k}).$$

- Since we need a decoder forward-pass per path $\mathbf{z}$, we can only optimize a limited set of paths.

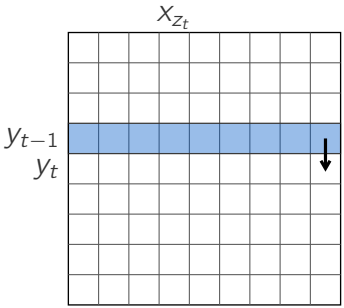
The Pervasive Attention Architecture (Elbayad et al. 2018)



Simultaneous Pervasive Attention

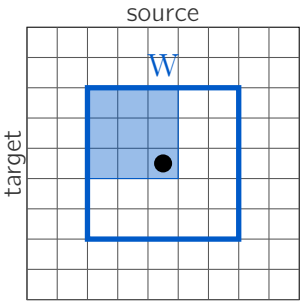


2D causal convolution

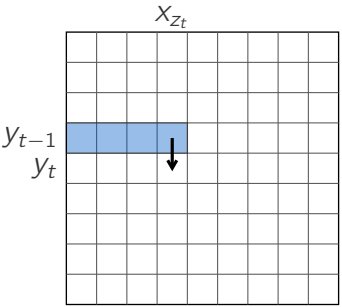


Features aggregation

Simultaneous Pervasive Attention

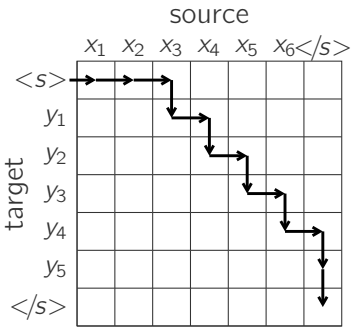


+ Masking the future source for unidirectional encoding.



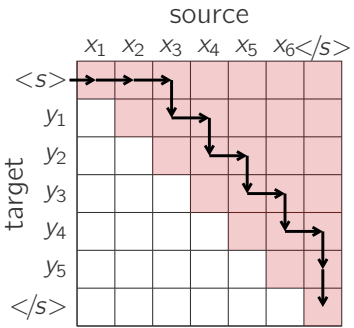
The appropriate context size z_t is controlled during aggregation.

Simultaneous Pervasive Attention



- A single wait-k path.

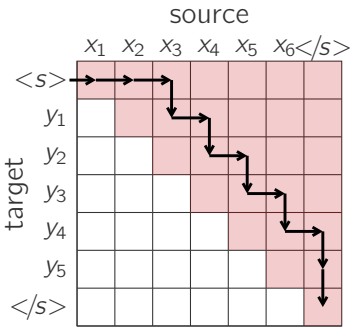
Simultaneous Pervasive Attention



- A single wait-k path.
- Optimizing an area of the grid (*multi-path*).

$$\sum_{t=1}^{|\mathbf{y}|} \sum_{j=1}^{|\mathbf{x}|} \log \mathrm{p}_\theta(y_t | \mathbf{y}_{<t}, \mathbf{x}_{\leq j}) [\![j \geq t]\!].$$

Simultaneous Pervasive Attention



- A single wait-k path.
- Optimizing an area of the grid (*multi-path*).

$$\sum_{t=1}^{|\mathbf{y}|} \sum_{j=1}^{|\mathbf{x}|} \log \mathrm{p}_\theta(y_t | \mathbf{y}_{<t}, \mathbf{x}_{\leq j}) [\![j \geq t]\!].$$

- With a single forward-pass we cover exponentially many paths falling within the optimized area.

Experimental Results

Experimental setup

Dataset	Train	Vocabularies	Baseline models
IWSLT14 En↔De	160K	BPE 8.8K (De) and 6.6K (En)	TF small and PA (N=14)
IWSLT'15 En↔Vi	133K	Words 17k (En) and 7.7K (Vi)	TF base and PA (N=14)
WMT15 De→En	4.5M	BPE 32K (De+En)	TF base and big

TF = Transformer, PA = Pervasive Attention

Quality-Lagging Trade-off

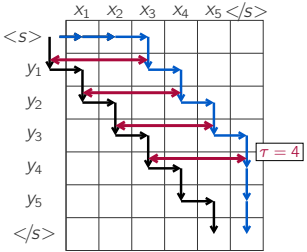
- **Quality:** BLEU of greedily decoded hypotheses.

Quality-Lagging Trade-off

- **Quality:** BLEU of greedily decoded hypotheses.
- **Lagging:** AL (Ma *et al.* 2019) measuring the lagging behind an ideal wait-0.

$$\mathrm{AL} = \frac{1}{\tau(\mathbf{x},\mathbf{z})} \sum_{t=1}^{\tau(\mathbf{x},\mathbf{z})} z_t - \frac{|\mathbf{x}|}{|\mathbf{y}|}(t-1),$$

$$\tau(\mathbf{x},\mathbf{z}) = \min\{t \mid z_t = |\mathbf{x}|\}.$$

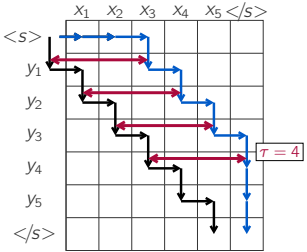


Quality-Lagging Trade-off

- **Quality:** BLEU of greedily decoded hypotheses.
- **Lagging:** AL (Ma *et al.* 2019) measuring the lagging behind an ideal wait-0.

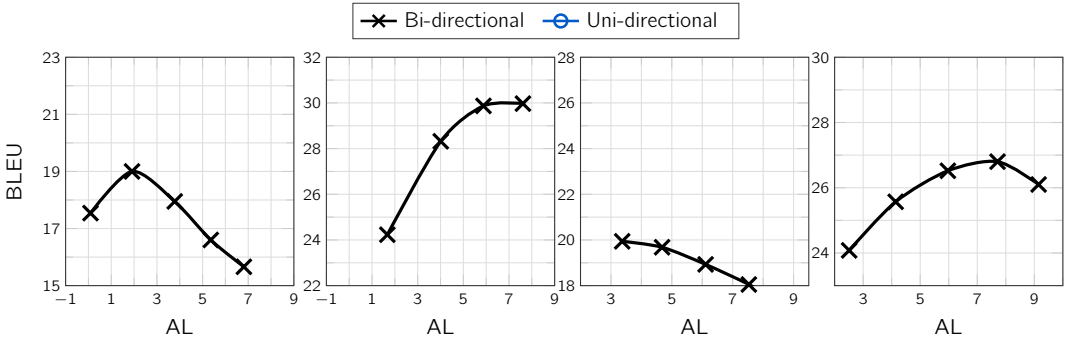
$$\mathrm{AL} = \frac{1}{\tau(\mathbf{x},\mathbf{z})} \sum_{t=1}^{\tau(\mathbf{x},\mathbf{z})} z_t - \frac{|\mathbf{x}|}{|\mathbf{y}|}(t-1),$$

$$\tau(\mathbf{x},\mathbf{z}) = \min\{t \mid z_t = |\mathbf{x}|\}.$$



- We distinguish between k_{train} used for training and k_{eval} used for evaluation.

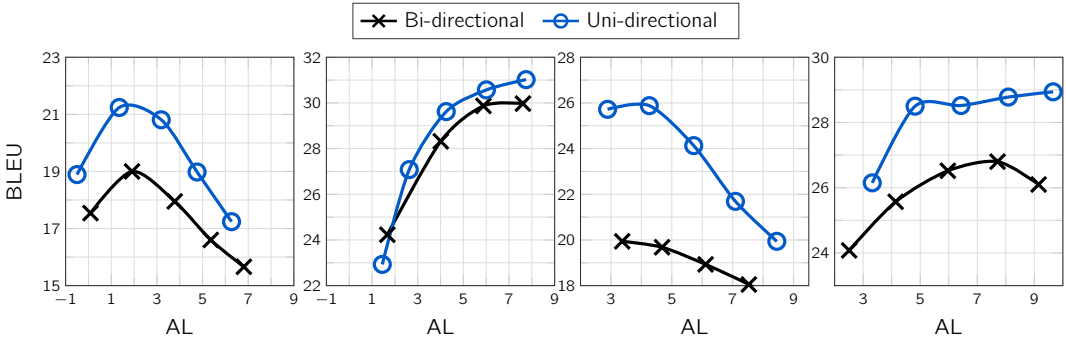
Unidirectional vs bidirectional TF. | IWSLT'14 De→En + IWSLT'15 En→Vi



(a) De→En $k_{\text{train}} = 1$ (b) De→En $k_{\text{train}} = 7$ (c) En→Vi $k_{\text{train}} = 1$ (d) En→Vi $k_{\text{train}} = 7$

Figure: IWSLT'14 De→En and IWSLT'15 En→Vi: Impact of the TF encoder's directionality for wait-k decoding ($k_{\text{train}} \in \{1, 7\}$).

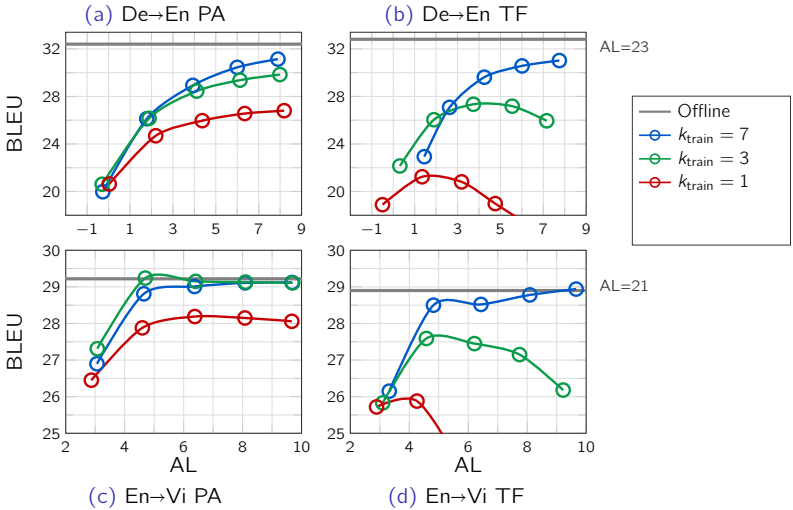
Unidirectional vs bidirectional TF. | IWSLT'14 De→En + IWSLT'15 En→Vi



(a) De→En $k_{\text{train}} = 1$ (b) De→En $k_{\text{train}} = 7$ (c) En→Vi $k_{\text{train}} = 1$ (d) En→Vi $k_{\text{train}} = 7$

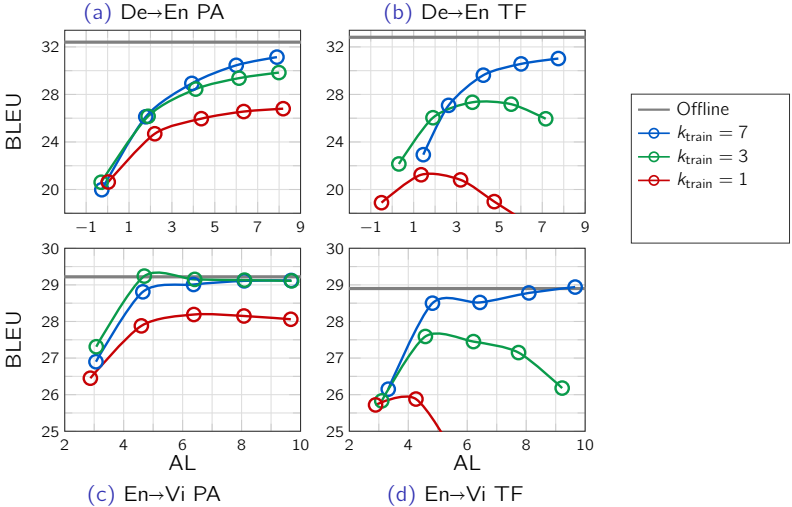
Figure: IWSLT'14 De→En and IWSLT'15 En→Vi: Impact of the TF encoder's directionality for wait-k decoding ($k_{\text{train}} \in \{1, 7\}$).

PA vs. TF | IWSLT'14 De→En & IWSLT'15 En→Vi



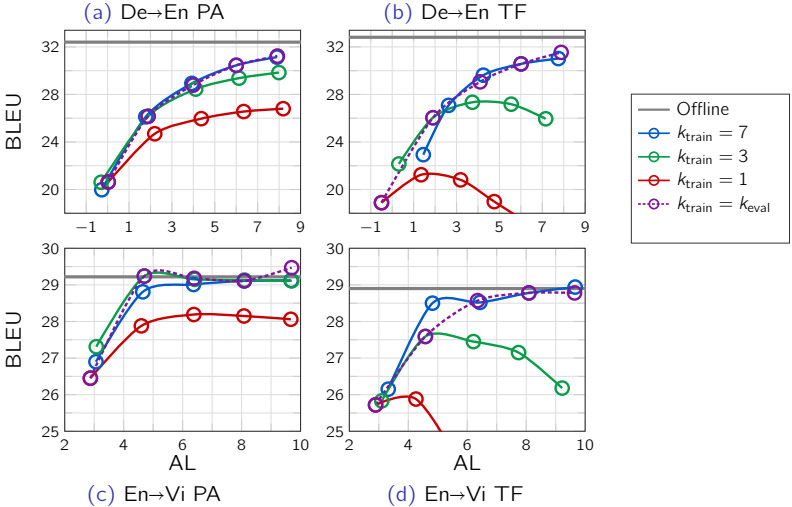
PA vs. TF | IWSLT'14 De→En & IWSLT'15 En→Vi

- Some paths are easier to train with than others.
- PA is more stable outside its training regime.



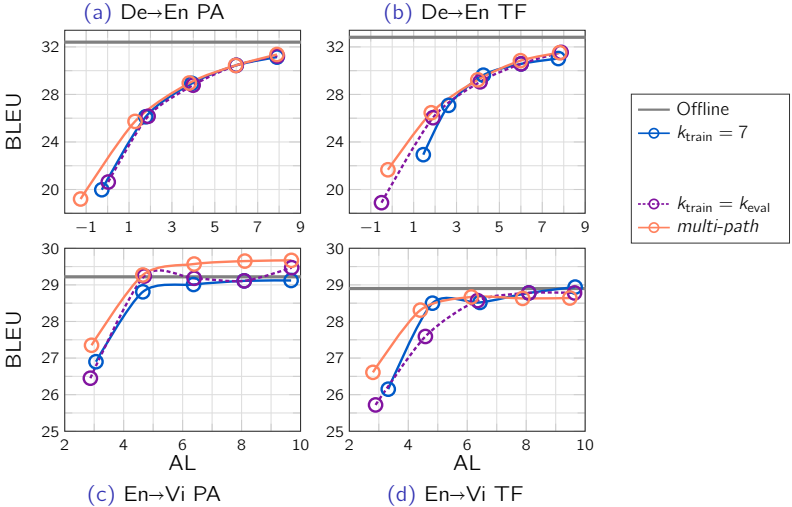
PA vs. TF | IWSLT'14 De→En & IWSLT'15 En→Vi

- Some paths are easier to train with than others.
- PA is more stable outside its training regime.
- We dont't need a model for each k_{eval}.

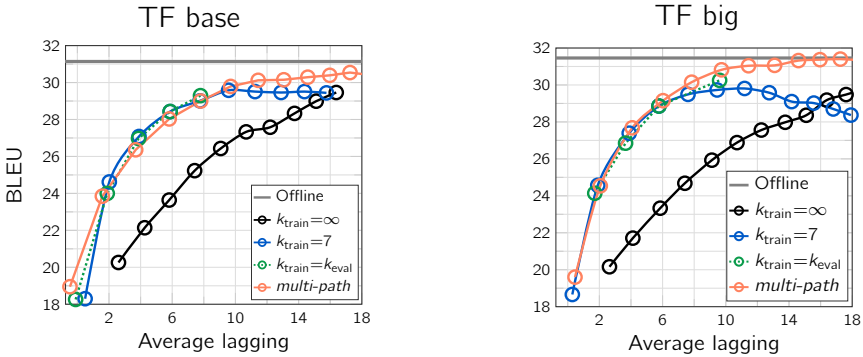


PA vs. TF | IWSLT'14 De→En & IWSLT'15 En→Vi

- Some paths are easier to train with than others.
- PA is more stable outside its training regime.
- We dont't need a model for each k_{eval}.
- We dont't need to tune k_{train}.

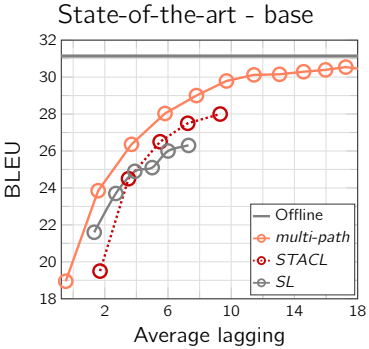


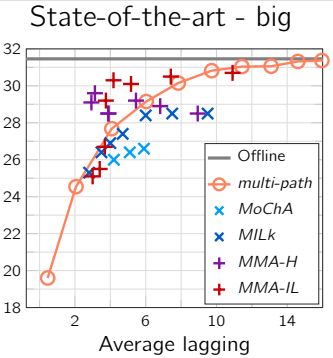
TF | WMT'15 De→En



- Multi-path is better than single path across the AL range.

TF | WMT'15 De→En





- Our models outperform STACL (Ma et al. 2019) and dynamic SL (Zheng *et al.* 2019).
- Our models are competitive with MILk (Arivazhagan *et al.* 2019).
- They form a stronger baseline for MMA (Ma *et al.* 2020).

Conclusion

Experiments on IWSLT'14 De↔En, IWSLT'15 En↔Vi and WMT'15 DE→En show that:

- Unidirectional encoding is more efficient for the task of simultaneous translation.
- Our models are state-of-the-art among deterministic translation strategies.
- Mutli-path models are more versatile than single path models.
- Our simple determinisitc decoding provide a strong baseline for dynamic simultaneous translation.