Adversarial Mask Transformer for Sequential Learning

Hou Lio   Shang-En Li   Jen-Tzung Chien

National Yang Ming Chiao Tung University

jtchien@nycu.edu.tw

ICASSP 2022, Singapore

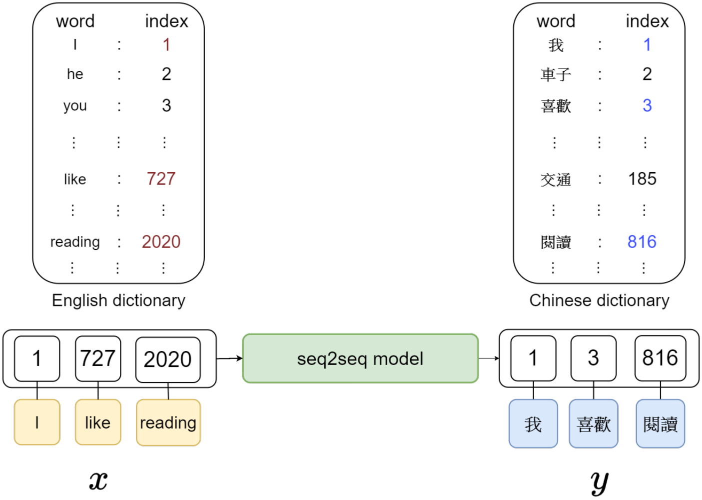
Table of Contents

Outline

Outline

1. Introduction
   - Sequential learning
   - Transformer
2. Adversarial mask transformer
3. Experiments
4. Conclusions and future works

Sequence-to-sequence model



Conditional likelihood maximization

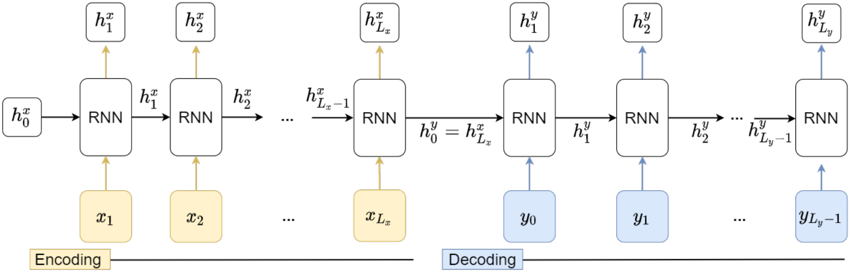
- Given a pair of sequential data $(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ are source and target sequences
- Seq2seq learning aims to learn a mapping function $f_{x \rightarrow y}$
- A standard seq2seq model is an encoder-decoder framework
  - encoder: extract features $\boldsymbol{h}_x$ from $\boldsymbol{x}$
  - decoder: generate target sequence $\boldsymbol{y}$ with condition on $\boldsymbol{h}_x$
- Conditional probability is calculated by

$$p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{y_i \in \boldsymbol{y}} p(y_i | y_1, y_2, \cdots, y_{i-1}, \boldsymbol{x})$$

RNN-based seq2seq model

- Encoder-decoder framework can be implemented by RNN
- Recurrent neural network extracts hidden features $\boldsymbol{h}_t$ via

$$\boldsymbol{h}_t^x = \mathrm{RNN}(\boldsymbol{h}_{t-1}^x, \boldsymbol{x}_t)$$



Outline

1. Introduction
   - Sequential learning
   - Transformer
2. Adversarial mask transformer
3. Experiments
4. Conclusions and future works

Attention mechanism

- Attention was first proposed by Bahdanau et al. (2014)
- Basic attention needs query $\boldsymbol{q}$, keys $\boldsymbol{K}$, values $\boldsymbol{V}$

$$\boldsymbol{c} = \text{attention}(\boldsymbol{q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\boldsymbol{q} \cdot \boldsymbol{K}^T)\boldsymbol{V}$$

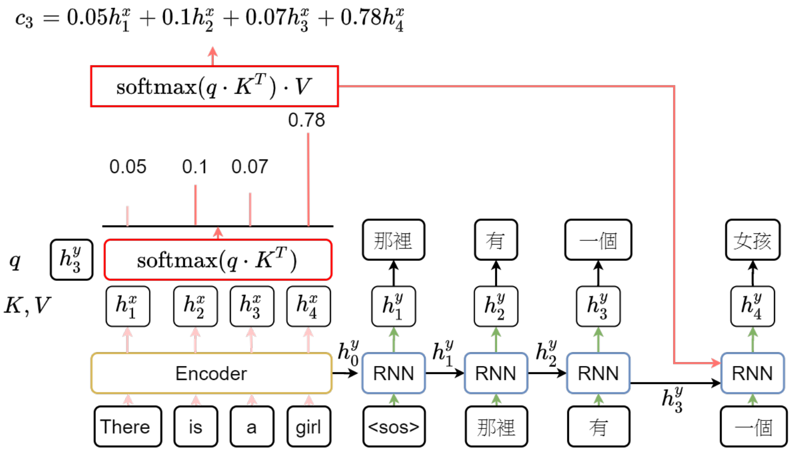
where $\boldsymbol{q} \in R^{1\times d}$, $\boldsymbol{K} \in N_k \times d$, and $\boldsymbol{V} \in N_k \times d$

- $\boldsymbol{q}$ and $\boldsymbol{K}$ determine which rows in values should be focused more

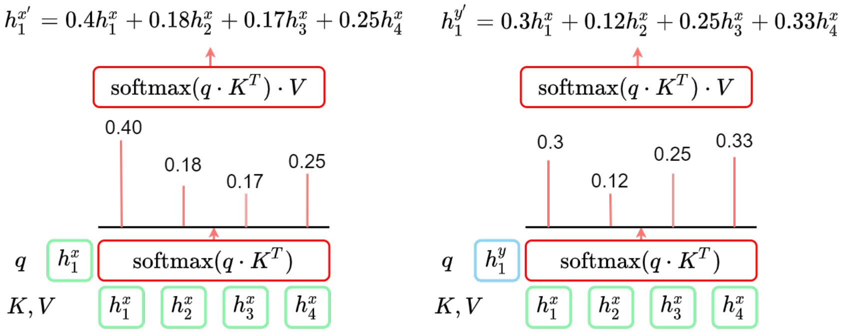
$$\boldsymbol{c} = \sum_{i=1}^{N_k} a_i \boldsymbol{V}[i]$$

where $\boldsymbol{a} = \text{softmax}(\boldsymbol{q} \cdot \boldsymbol{K}^T)$, where $\boldsymbol{a} \in R^{1\times N_k}$

RNN-based seq2seq model with attention

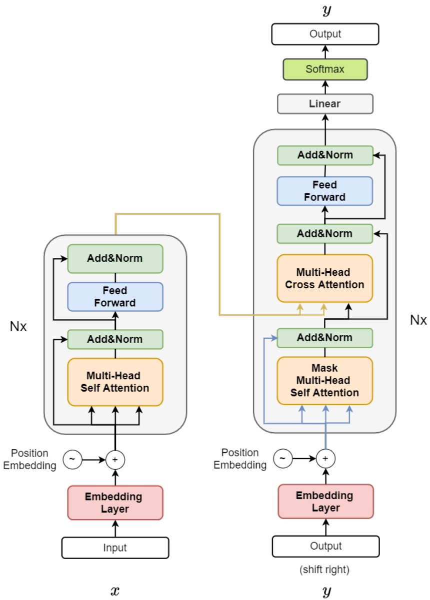


- Self attention
  - queries $\boldsymbol{Q}$, keys $\boldsymbol{K}$, and values $\boldsymbol{V}$ are the same features
  - extracts features from self domain
- Cross attention
  - queries $\boldsymbol{Q}$ are different from Keys $\boldsymbol{K}$ and values $\boldsymbol{V}$
  - obtains features from the other domain

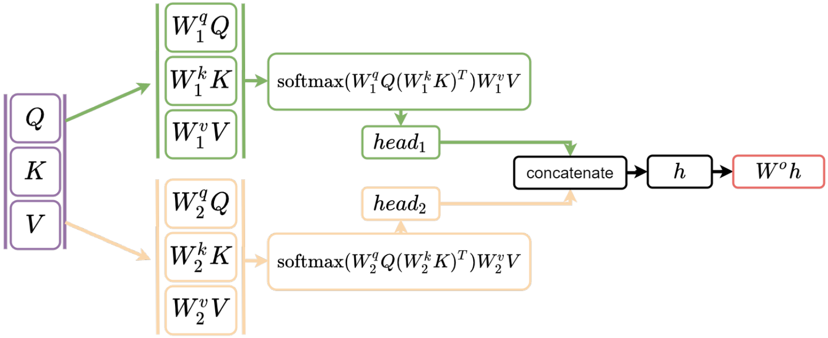


Transformer-based seq2seq model

- Left part: Encoder
- Right part: Decoder
- Main modules (Vaswani et al., 2017)
  - position embedding
  - multi-head attention
  - point-wise feed forward network
  - masked multi-head attention



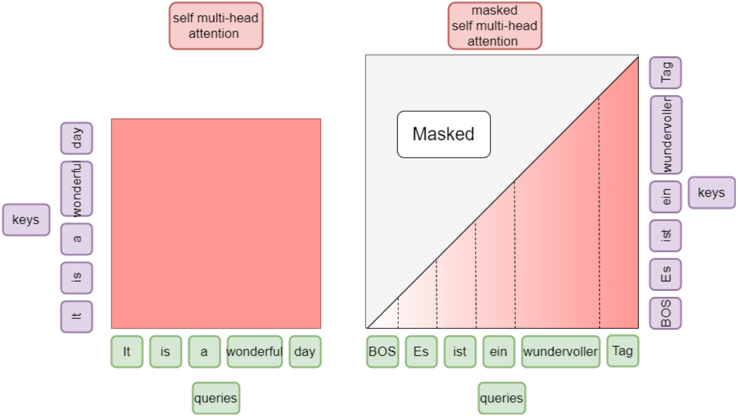
Multi-head self attention



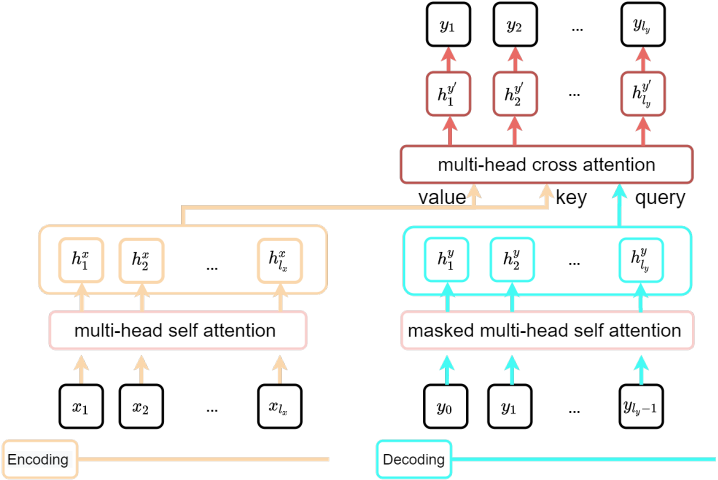
$$\mathrm{MultiHead}(Q, K, V) = W^o \mathrm{Concatenate}(head_1, head_2, \cdots)$$

$$head_i = \mathrm{Attention}(W_i^q Q, W_i^k K, W_i^v V)$$

Masked multi-head self attention



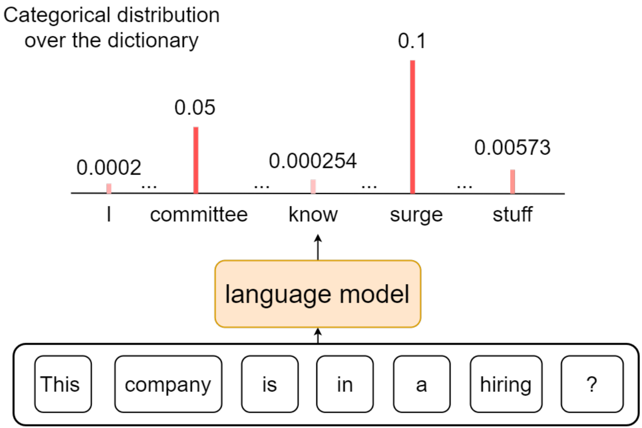
Learning procedure



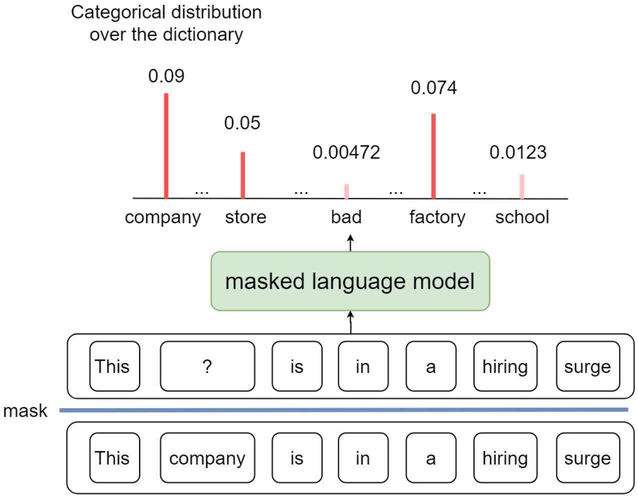
Outline

Outline

Language model

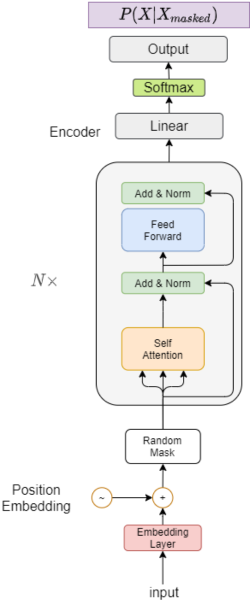


Masked language model



Masked language model

- Objective
  - conditional probability $p(\boldsymbol{x}|\boldsymbol{x}_{masked})$
- Optimization
  - minimize $-\sum_{x_i \in \boldsymbol{x}} \log p(x_i|x_{masked})$



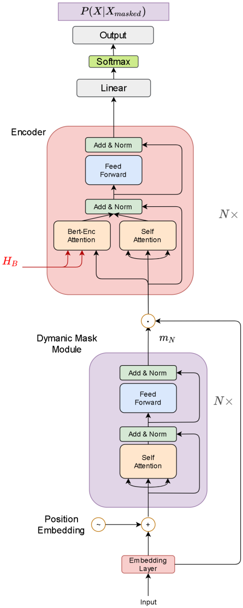
Motivation of this study

- Language understanding plays an important role in NLP
- Masked language model can effectively enhance language understanding. Mask strategy rely on the process of randomization
- Effectively choosing the mask strategy is crucial
- Most cover strategies are mainly based on random mask
  - Randomly select 15% of input tokens. A large dataset is required
    - replace 80% of them with the token [mask]
    - 10% remain unchanged
    - 10% randomly replace other words
- We propose the "adversarial mask transformer"
  - use the adversarial learning to learn different mask strategies
  - adapt the mask strategy to different tasks via their datasets

Outline

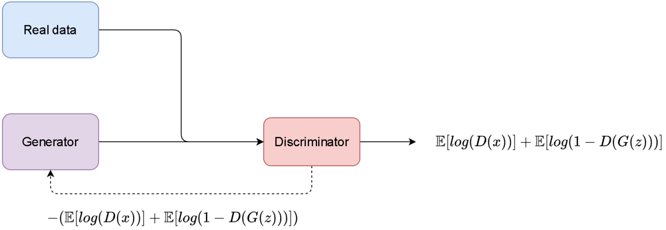
Adversarial masked language model

- Objective
  - conditional policy $\pi(\boldsymbol{m}|\boldsymbol{x})$
  - conditional probability $p(\boldsymbol{x}|\boldsymbol{x}_{masked})$
- Optimization
  - minimize θ
    $$-\sum_{x_i \in \boldsymbol{x}} \log p_\theta(x_i|\boldsymbol{x}_{masked})$$
  - maximize ϕ
    $$E(R(x) \log \pi_\phi(\boldsymbol{m}|\boldsymbol{x}))$$

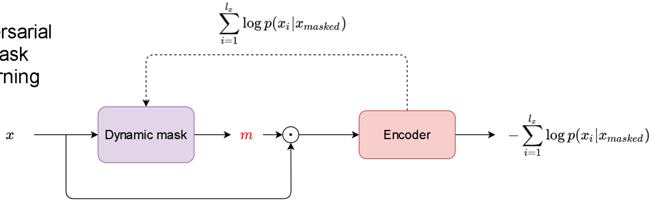


Adversarial learning vs. adversarial mask learning

Adversarial learning



Adversarial mask learning



Policy gradient for adversarial mask learning

- Policy gradient is implemented to choose mask

$$\boldsymbol{\nabla} J(\phi) = \mathbb{E}[\nabla \log \pi_\phi(a|s) R(s)]$$

- Mask is generated by $m \sim \pi(\cdot|x)$
- We define the loss of MLM as an intrinsic reward

$$\mathcal{L}_E(\theta) = -\sum_{i=1}^{l_x} \log p(x_i|x_{masked})$$

- General objective function of adversarial learning considers

$$\min_G \max_D \mathbb{E}[\log(D(x))] + \mathbb{E}[\log(1 - D(G(z)))]$$

- Adversarial mask learning follows the objective

$$J(\theta, \phi) = \min_\theta \max_\phi \mathbb{E}[\mathcal{L}_E(\theta) \log \pi_\phi(m|x))]$$

- Maximization for policy gradient, and minimization for MLM

Algorithm 1: Adversarial mask learning

Masked language model

Initial ϕ, the parameters of approximation function, randomly

for *episode* $e \in \{1, 2, ..., N\}$ **do**

initialize state $s_t = x_t$

$R_t = \sum_{t=1}^{T} r_t$

for $t \in \{1, 2, ...T\}$ **do**

sample a_t from policy distribution $\pi(a_t|x_t)$

given next sentence x_{t+1}

given $r_t = -\mathcal{L}_E(\theta)$

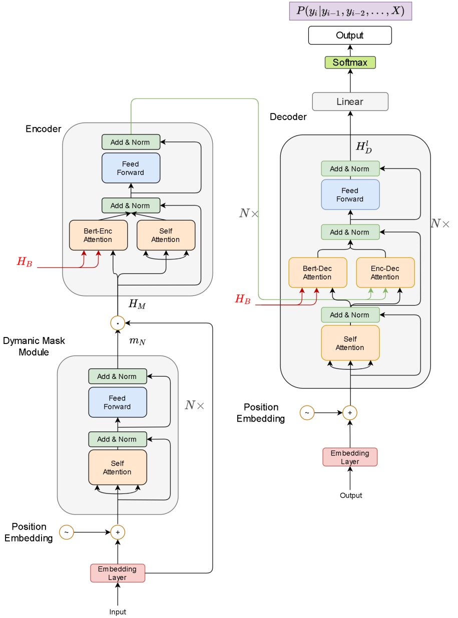
store (x_t, a_t, r_t) into buffer

end

$$\phi \leftarrow \phi + \frac{\alpha}{N} \sum_{e=1}^{N} \sum_{t=1}^{T} R_t \nabla_\phi \log \pi_\phi(a_t|x_t)$$

end

- Adversarial Mask Transformer
- Transformer encoder
  - extracts $\boldsymbol{x}$ features
- Transformer decoder
  - extracts $\boldsymbol{y}$ features
  - grasps $\boldsymbol{x}$ features
- Objective
  - conditional probability $p(y_i|y_1, \cdots y_{i-1}, \boldsymbol{x})$
- Optimization
  - minimize
    $$-\sum_{y_i \in \boldsymbol{y}} \log p(y_i|y_1, \cdots y_{i-1}, \boldsymbol{x})$$



Outline

Outline

Model configuration

- Our model uses the standard setting as transformer
  - number of heads 8
  - number of layers 6
- Our model uses the setting as dynamic mask module
  - number of heads 8
  - number of layers 2
- Word embedding for source and target is 512 units
- Hidden size for source and target is 512 units
- Trained by Adam optimizer
  - batch size 128
  - initial learning rate 0.0005
- BLEU scores are evaluated

IWSLT and WMT machine translation tasks

- IWSLT German(De)-to-English(En) translation task
  - training set 200k pairs of sentences
  - validation set 7k pairs of sentences
  - test set 7k pairs of sentences
  - vocabulary size of 10k words

Language	Sentences
German	oft ist es abwasser , was uns verstopft . was macht man , wenn man solch eine unterbrechung im fluss hat ? stephen palumbi : der spur des quecksilbers folgen sie wären unter meinem niveau .
English	often what jams us up is sewage . what do you do when you have this sort of disrupted flow ? stephen palumbi : following the mercury trail i really thought they were so beneath me .

Outline

- Results on IWSLT translation between German and English

Model	En$\rightarrow$De	De$\rightarrow$En
ConvS2S (Gehring et al., 2017)	26.1	31.9
Transformer (Vaswani et al., 2017)	28.6	34.4
Weighted Transformer (Ahmed et al., 2017)	28.9	35.1
Evolved Transformer (So et al., 2019)	30.4	36.0
BERT-fused model (Zhu et al., 2020)	30.5	36.1
Adversarial Mask Transformer	**30.9**	**36.6**

- Results on WMT English to German

Model	BLEU
ConvS2S (Gehring et al., 2017)	25.2
Transformer (Vaswani et al., 2017)	26.2
Weighted Transformer (Ahmed et al., 2017)	27.2
Evolved Transformer (So et al., 2019)	28.4
BERT-fused model (Zhu et al., 2020)	28.3
Adversarial Mask Transformer	**28.9**

- Results on MLM-fused models on WMT English to German

Model	BLEU
BERT+LM (Devlin et al., 2019)	24.9
Transformer with Mask-Predict (Ghazvininejad et al., 2019)	27.7
MASS (Song et al., 2019)	28.3
Adversarial Mask Transformer	**28.9**

Outline

- Conclusions
  - adversarial mask learning is proposed
  - this method combines adversarial learning and reinforcement learning
  - a small dataset can be used to generalize for a model with large dataset
  - pretrained and then fine-tuned
  - experiments show this model can improve the translation result
- Future works
  - learning of mask strategy can be changed to a specific target task
  - other sequential learning applications
  - text summarization
    - raise the mask strategy from word level to sentence level
    - reward is defined as just guessing some part of the sentence
  - question answering
    - add the masked language model to train decoder

References I

[1] K. Ahmed, N. S. Keskar, and R. Socher, "Weighted transformer network for machine translation," *arXiv preprint arXiv:1711.02132*, 2017.

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of Conference of North American Chapter of Association for Computational Linguistics*, 2019, pp. 4171–4186.

[4] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proc. of International Conference of Machine Learning*, 2017, pp. 1243–1252.

[5] M. Ghazvininejad, O. Levy, Y. Liu, and L. Zettlemoyer, "Mask-Predict: Parallel decoding of conditional masked language models," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 6111–6120.

[6] D. R. So, Q. V. Le, and C. Liang, "The evolved transformer," in *Proc. of International Conference of Machine Learning*, 2019, pp. 5877–5886.

[7] K. Song, X. Tan, T. Qin, J. Lu, and T. Liu, "MASS: masked sequence to sequence pre-training for language generation," in *Proc. of International Conference of Machine Learning*, 2019, pp. 5926–5936.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

References II

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[10] J. Zhu, Y. Xia, L. Wu, D. He, T. Qin, W. Zhou, H. Li, and T.-Y. Liu, “Incorporating BERT into neural machine translation,” in *Proc. of International Conference on Learning Representations*, 2020.