

PARALLEL SENTENCE-LEVEL EXPLANATION GENERATION FOR REAL-WORLD LOW-RESOURCE SCENARIOS

Yan Liu¹, Xiaokang Chen², Qi Dai¹

¹Microsoft Research Asia

²School of Intelligence Science and Technology, Peking University

ABSTRACT

In order to reveal the rationale behind model predictions, many works have exploited providing explanations in various forms. Recently, to further guarantee readability, more and more works turn to generate sentence-level human language explanations. However, current works pursuing sentence-level explanations rely heavily on annotated training data, which limits the development of interpretability to only a few tasks. As far as we know, this paper is the first to explore this problem smoothly from weak-supervised learning to unsupervised learning. Besides, we also notice the high latency of autoregressive sentence-level explanation generation, which leads to asynchronous interpretability after prediction. Therefore, we propose a non-autoregressive interpretable model to facilitate parallel explanation generation and simultaneous prediction. Through extensive experiments on Natural Language Inference task and Spouse Prediction task, we find that users are able to train classifiers with comparable performance 10 – 15× faster with parallel explanation generation using only a few or no annotated training data.

Index Terms— interpretability, parallel explanation generation, low-resource scenarios

1. INTRODUCTION

Recently, deep learning has developed rapidly [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. The interpretability of black-box neural networks has aroused much attention and the importance of interpreting model predictions has been widely acknowledged. Previous interpretation works provide explanations in various forms as the rationale lying behind model decisions, such as attention distribution [14], heatmap [15], input keywords [16], etc. Due to the better human readability, many works exploit generating sentence-level human language explanations to better interpret model predictions and have achieved promising performance [17, 18].

However, sentence-level explanations are hard to achieve in real-world scenarios due to the high latency of autoregressive explanation generation and the severe reliance on human-annotated explanations. For instance, e-INFERSENT[17] autoregressively generates every explanation token, leading to much higher inference latency. In comparison, although some

explanations lack readability[19], such as the attention-based heatmap explanation and post-hoc alignment map explanation, these explanations can be generated almost simultaneously with predictions. Moreover, in spite of readability, previous works that generate sentence-level explanations rely heavily on numerous human-annotated explanations during training. Nevertheless, datasets containing human-annotated explanations are rare due to the high cost.

To alleviate these problems, in this work, we introduce the Classification Non-Autoregressive Transformer (C-NAT) framework for simultaneous classification and parallel sentence-level explanation generation with weakly-supervised and unsupervised learning strategies. To accelerate the explanation generation, we adopt the architecture of the non-autoregressive generation model NAT [20] to generate all tokens in parallel. We also equip the non-autoregressive generation model with a label predictor for simultaneous label prediction. Besides, to better accommodate real-world low-resource scenarios, we propose our weakly-supervised learning and unsupervised learning strategies. Specifically, inspired by [21], we first extract a set of labeling functions and the corresponding explanation templates from a small number of human-annotated samples, and then use these labeling functions and explanation templates to produce pseudo labels and explanations for a large amount of unlabeled data. For the unsupervised learning scenario, we utilize the back-translation mechanism to paraphrase the input sequences as the pseudo explanation targets, and apply a pre-trained language model to refine the predicted explanations during training. We verify the effectiveness of our approach on the Natural Language Inference (NLI) task and the Spouse Prediction (SP) task. Main contributions of this work are three-fold:

- We propose novel weakly supervised learning and unsupervised learning strategies to accommodate interpretable models to real-world low-resource scenarios.
- We introduce our C-NAT to support parallel explanation generation and simultaneous prediction. We also propose to leverage a pre-trained language model as a discriminator to generate more fluent explanations.
- Experimental results show that our C-NAT can generate parallel fluent explanations and improve classification

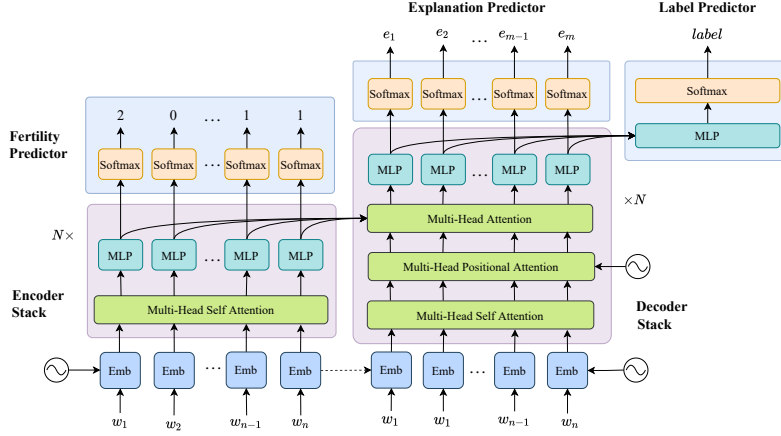


Fig. 1. The overall architecture of our C-NAT.

performance with significant inference speedup, even with few or no human annotations.

2. MODEL ARCHITECTURE

In this section, we introduce the architecture of C-NAT, which modifies the non-autoregressive generation model NAT [20] to support simultaneous label prediction and parallel sentence-level explanation generation. As shown in figure 1, C-NAT consists of the following five modules: an encoder stack, a decoder stack, a fertility predictor, an explanation predictor for parallel explanation tokens generation, and a label predictor for simultaneous label prediction.

2.1. Encoder and Decoder

We adopt the Transformer[22] as the backbone. To enable non-autoregressive interpretation, following [20], the decoder is modified in three aspects: input sequence, self-attention mask, and positional encoding. For input sequence modification, because previously generated tokens are unavailable under the non-autoregressive setting, we use a fertility predictor first to predict the length of the target explanation and produce decoder input with the tokens copied from the encoder input. For the modification of the self-attention mask, because the decoder input is the copied sequence of encoder input, the self-attention module is allowed to attend all positions, rather than only left positions in the conventional Transformer decoder. Therefore, the self-attention mask is replaced with a non-causal mask in our non-autoregressive decoder. For positional encoding modification, different from the self-attention module, the positional attention module uses positional encoding as the query and key, and the hidden representations from the previous layer as the value.

2.2. Fertility predictor

To generate the decoder input sequence for non-autoregressive interpretation, we copy and repeat the tokens from the encoder input. The fertility predictor is used to predict the number of times each token is copied, referred to as the *fertility* of each corresponding token [20]. Specifically, given the input sentence of the encoder $X = \{x_1, x_2, \dots, x_S\}$, the fertility predictor is fed with the encoded feature $H = \{h_1, h_2, \dots, h_S\}$, and generates the fertility sequence $F = \{f_1, f_2, \dots, f_S\}$. Finally, the input sequence of the non-autoregressive decoder is $Y = \{y_1, y_2, \dots, y_T\} = \{\{x_1\}_{i=1}^{f_1}, \{x_2\}_{i=1}^{f_2}, \dots, \{x_S\}_{i=1}^{f_S}\}$ with length $T = f_1 + f_2 + \dots + f_S$, where $\{y_s\}_{i=1}^{f_s}$ denotes the token x_s is repeated for f_s times.

2.3. Explanation Predictor and Label Predictor

The explanation predictor and label predictor are used to generating each token of the explanation sentence and classification label simultaneously. Given the output hidden states of the decoder stack $H^d = \{h_1^d, h_2^d, \dots, h_T^d\}$, each explanation token e_t is generated with the probability $p_E(e_t) = \text{Softmax}(h_t^d)$, and the explanation sentence $E = \{e_1, e_2, \dots, e_T\}$ is generated in parallel with the probability $p_E(E|X; \theta) = \prod_{t=1}^T p_E(e_t|X; \theta)$. Meanwhile, the label predictor projects the hidden states with an MLP layer and the mean pooling operation, resulting in the label prediction L with the probability $p_L(L|X; \theta)$.

3. TRAINING STRATEGY

3.1. Fully-supervised Learning

In the explanation available scenario, the fully-supervised training objective function of our model is the combination of the label prediction loss, the explanation prediction loss, and the fertility prediction loss. Besides, we also apply the pre-trained language model as an extra constraint on the ob-

Datasets	Train	Val	Test	Annotated/Total
e-SNLI	549367	9842	9824	570K/570K
SP-Pseudo	22195	2796	2796	30/22195
SNLI-Pseudo	549367	9842	9824	0/570K

Table 1. Statistics of datasets.

jective function to encourage generating explanations of more fluency and diversity. Then the pre-trained language model with parameters θ_{LM} estimates the log-likelihood of each predicted explanation sentence E' as $\log p_{LM}(E'; \theta_{LM})$. To enable the gradient backpropagation from the pre-trained language model to the C-NAT model, the product of the predicted probability distribution $p_E(e_t|X; \theta)$ and the word embedding vectors is used as the input embedding of the explanation token e'_t in the pre-trained language model. The additional loss term \mathcal{L}_{LM} is adopted to optimize the explanation generation by maximizing the estimated log-likelihood of the pre-trained language model over the training dataset. Finally, the fully-supervised training objective function of our C-NAT model is formulated as:

$$\mathcal{L} = \mathcal{L}_L + \lambda_E \mathcal{L}_E + \lambda_F \mathcal{L}_F + \lambda_{LM} \mathcal{L}_{LM} \quad (1)$$

, where λ_L , λ_E , λ_F and λ_{LM} are hyperparameters for each loss term.

3.2. Weakly-supervised Learning

In the more practical scenario, where only a few human annotated explanations are available, we introduce the weakly-supervised learning strategy to generate the pseudo explanations and pseudo labels for the large-scale unlabeled data. Firstly, we extract the labeling functions along with the explanation templates from a small number of human-annotated samples. Then, we use the labeling functions and the explanation templates to annotate the pseudo labels and explanations for the large-scale unlabeled data. Due to the wide divergence in accuracy and coverage of the labeling functions, the data programming method [23] is applied for label aggregation [21], where a learnable accuracy weight w_m is assigned to each labeling function $f_m(\cdot)$, and the final pseudo label is selected as the label with the largest aggregated accuracy weight. As for the labeling function with the highest contribution to the pseudo label, we select the corresponding explanation template E_m^{temp} and generate the pseudo natural language explanation $\{E^{\text{pseudo}}\}$. Finally, the training data \mathcal{D} is a combination of the small amount of human-annotated data, and a large amount of data with pseudo labels and explanations. We optimize our C-NAT with the fully-supervised training objective function on the combined training dataset.

3.3. Unsupervised Learning

For the real-world scenario where no human annotated explanations are available, we also explore the unsupervised learning strategy for our C-NAT model training. Different from

the autoregressive interpretation approach, golden explanations are only used as the training target but not the decoder input for the non-autoregressive interpretation approach. To mimic the human-annotated training targets, we utilize the back-translation mechanism to generate pseudo explanations as the noisy training targets, and keep refining the explanation generation with a pre-trained language model during training.

4. EXPERIMENTS

4.1. Tasks and Datasets

To verify the effectiveness of our approach, we conduct experiments on the Natural Language Inference (NLI) and Spouse Prediction (SP) tasks. NLI task aims to predict the entailment relationship between two sentences. SP task is to predict whether two people in the given sentence are spouses.

We use three datasets as our testbeds for **fully-supervised**, **weakly-supervised** and **unsupervised** learning respectively: **e-SNLI** [17], **SP** [21], and **SNLI** [24]. SNLI is a standard benchmark for the NLI task, while e-SNLI extends it with human-annotated natural language explanations for each sentence pair. Therefore, we use the e-SNLI dataset to generate explanations with full-supervision, while using the SNLI dataset for unsupervised explanation generation. SP dataset has only 30 samples annotated with human explanations, which we thus adopt for weakly-supervised explanation generation. Besides, as introduced in Section 3.2 and 3.3, we propose two methods to generate pseudo data in low-resource scenarios. For the SP dataset, we extract templates from 30 human explanations, which are then used to generate pseudo explanations and form our **SP-Pseudo dataset**. For the SNLI dataset without human annotated explanations at all, we propose to use a pre-trained NMT model to generate pseudo explanations, which form our **SNLI-Pseudo dataset**. The statistics of all datasets and pseudo data are shown in Table 1.

4.2. Metrics

To evaluate classification performance and explanation quality, we report **NE-Acc**(classification accuracy without generating explanations), **Acc** (classification accuracy), **BLEU** (similarity between generation and ground truth, if any), **PPL** (fluency of generated explanations), **Inter-Rep** (diversity of generated explanations), and **Rationality** (rationality of explanations). Specifically, the Rationality metric is a model-based evaluation metric, which utilizes a pre-trained classifier to evaluate whether the generated explanation is reasonable for corresponding input and prediction.

4.3. Implementation Details

We set the embedding size and the hidden size as 512, and use 8 heads. The layer number of the encoder and decoder are set as both 6. We use the Adam [25] for optimization with

Methods	BLEU [↑]	Rationality [↑]	PPL [↓]	Inter-Rep [↓]	NE-Acc [↑]	Acc [↑]	Latency [↓]	Speedup [↑]
Dataset [†]	22.51	100.00	30.00	0.40	100.00	100.00	-	-
Transformer(AT)	20.33	80.16	27.04	0.51	80.62	79.46	793ms	1.27×
e-INFERSENT(AT)	22.40	84.79	10.58	0.72	84.01	83.96	1006ms	1.00×
C-NAT	21.19	85.10	34.71	0.30	82.41	85.23	47ms	21.40×
w/o LM	20.87	84.51	46.77	0.32	82.41	85.19	47ms	21.40×
w/o NAR	19.33	82.06	61.14	0.47	82.41	80.47	734ms	1.37×
w/o LCE	21.04	84.16	36.21	0.33	36.68	39.31	47ms	21.40×

Table 2. Automatic evaluation results in NLI task with full-supervision. The higher[↑](or smaller[↓]), the better. [†]We evaluate the ground truth with our metrics. Latency is computed as the time to decode a single output sequence without mini batching, averaged over the whole test set. At the bottom, we present the results of the ablation study.

Methods	Rationality [↑]	PPL [↓]	NE-Acc [↑]	Acc [↑]	Latency [↓]	Speedup [↑]
Transformer(AT)	76.51	31.74	81.27	84.11	566ms	1.00×
C-NAT	77.41	30.61	85.01	87.14	34ms	16.65×
w/o LM	75.79	44.31	85.01	86.95	34ms	16.65×
w/o NAR	71.78	47.24	85.01	84.13	518ms	1.09×
w/o LCE	76.09	30.24	42.65	47.15	34ms	16.65×

Table 3. Automatic evaluation results in the Spouse Prediction task with weak-supervision. The higher[↑](or smaller[↓]), the better. At the bottom, we present the results of ablation study.

Methods	Rationality [↑]	PPL [↓]	Inter-Rep [↓]	Acc [↑]
C-NAT	72.69	46.32	0.49	83.12
w/o LM	70.49	57.46	0.52	83.00
w/o SSUp	4.68	37.26	0.89	82.36

Table 4. Automatic evaluation results in NLI task with the unsupervised learning strategy. The higher[↑](or smaller[↓]), the better. At the bottom, we present the results of ablation study.

$\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The learning rate is set to 0.00004, and the dropout rate is set to 0.3.

4.4. Results of Fully-Supervised Learning

Evaluation results on e-SNLI in the full-supervised learning scenario are shown in Table 2. We observe that our C-NAT can achieve the comparable performance of explanation generation and label prediction with more than 20× speedup compared to the baseline autoregressive models. We also conduct the ablation study to evaluate the effectiveness of each component. We find that the BLEU score and PPL score drop significantly with the LM discriminator removed, but the prediction accuracy remains. It indicates that the pre-trained language model can effectively improve the fluency of generated explanations. If we modify C-NAT for autoregressive generation, much higher inference latency would be witnessed, and the performance would also degrade due to the exposure bias problem. Besides, we notice the classification performance drops on NE-Acc and ACC for baseline models, while our C-NAT achieves 2.07 absolute improvement. This demonstrates that our method can improve the inference ability of the classifier with model interpretability

increased, instead of improving interpretability at the cost of classification performance.

4.5. Results of Weakly-Supervised Learning

Table 3 shows the results of our C-NAT model with weakly-supervised learning strategy on the Spouse Prediction dataset that has only 30 human annotated explanations. We augment with pseudo data generated by our template-based approach. Because there are no previous works exploring the weakly-supervised learning method for explanation generation, we choose the modified Transformer model supporting classification as our baseline. Despite the small amount of human-annotated data, with the pseudo labels and explanations, we can still achieve improvement compared to the baseline model on all metrics.

4.6. Results of Unsupervised Learning

We conduct experiments in the Natural Language Inference task under the unsupervised learning scenario where no human-annotated explanations are available. Table 4 shows the experimental results of applying our approach in such a scenario. We observe that the LM clearly affects the persuasion accuracy and the fluency of explanations. Moreover, we also notice that the performance drops a lot without using the unsupervised learning strategy, which confirms the effectiveness of our unsupervised learning approach.

5. CONCLUSION

In this paper, we explore the important problem of generating human-friendly sentence-level explanations in low-resource scenarios. To solve the high inference latency problem of previous interpretable models, we propose our C-NAT to support parallel explanation generation and simultaneous prediction. We conduct extensive experiments in the Natural Language Inference task and Spouse Prediction task in the fully-supervised learning, weakly-supervised learning, and unsupervised learning scenarios. Experimental results reveal that our C-NAT can generate fluent and diverse explanations with classification performance also improved.

6. REFERENCES

- [1] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang, “Context autoencoder for self-supervised representation learning,” *arXiv preprint arXiv:2202.03026*, 2022.
- [2] Xiaokang Chen, Jiahui Chen, Yan Liu, and Gang Zeng, “D³etr: Decoder distillation for detection transformer,” *arXiv preprint arXiv:2211.09768*, 2022.
- [3] Yan Liu and Yazheng Yang, “Enhance long text understanding via distilled gist detector from abstractive summarization,” *arXiv preprint arXiv:2110.04741*, 2021.
- [4] Qiang Chen, Xiaokang Chen, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang, “Group detr: Fast detr training with group-wise one-to-many assignment,” *arXiv preprint arXiv:2207.13085*, vol. 1, no. 2, 2022.
- [5] Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng, “Not all voxels are equal: Semantic scene completion from the point-voxel perspective,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 2352–2360.
- [6] Xiaokang Chen, Fangyun Wei, Gang Zeng, and Jingdong Wang, “Conditional detr v2: Efficient detection transformer with box queries,” *arXiv preprint arXiv:2207.08914*, 2022.
- [7] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang, “Conditional detr for fast training convergence,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3651–3660.
- [8] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang, “Semi-supervised semantic segmentation with cross pseudo supervision,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2613–2622.
- [9] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng, “Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation,” in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*. Springer, 2020, pp. 561–577.
- [10] Xiaokang Chen, Kwan-Yee Lin, Chen Qian, Gang Zeng, and Hongsheng Li, “3d sketch-aware semantic scene completion via semi-supervised structure prior,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4193–4202.
- [11] Xiaokang Chen, Yajie Xing, and Gang Zeng, “Real-time semantic scene completion via feature aggregation and conditioned prediction,” in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 2830–2834.
- [12] Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng, “Compressible-composable nerf via rank-residual decomposition,” *arXiv preprint arXiv:2205.14870*, 2022.
- [13] Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng, “Point scene understanding via disentangled instance mesh reconstruction,” in *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*. Springer, 2022, pp. 684–701.
- [14] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *International conference on machine learning*, 2015.
- [15] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller, “Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models,” *arXiv preprint arXiv:1708.08296*, 2017.
- [16] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal, “Generating token-level explanations for natural language inference,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.
- [17] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom, “e-snli: natural language inference with natural language explanations,” in *Advances in Neural Information Processing Systems*, 2018.
- [18] Yan Liu, Sanyuan Chen, Yazheng Yang, and Qi Dai, “MPII: Multi-level mutual promotion for inference and interpretation,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022, Association for Computational Linguistics.
- [19] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith, “Annotation artifacts in natural language inference data,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 107–112.
- [20] Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher, “Non-autoregressive neural machine translation,” *arXiv preprint arXiv:1711.02281*, 2017.
- [21] Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré, “Training classifiers with natural language explanations,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1884–1895.
- [22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [23] Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré, “Data programming: Creating large training sets, quickly,” *Advances in neural information processing systems*, vol. 29, 2016.
- [24] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning, “A large annotated corpus for learning natural language inference,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015, pp. 632–642.
- [25] Diederik P. Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.