MPII: Multi-Level Mutual Promotion for Inference and Interpretation

Yan Liu[1]*, **Sanyuan Chen**[2,3], **Yazheng Yang**[1], **Qi Dai**[3]
[1]School of Computer Science and Technology, Zhejiang University, China
[2]School of Computer Science and Technology, Harbin Institute of Technology, China
[3]Microsoft Research, Beijing, China
runningmelles@gmail.com, sychen@ir.hit.edu.cn,
yazheng_yang@zju.edu.cn, qid@microsoft.com

Abstract

In order to better understand the rationale behind model behavior, recent works have exploited providing interpretation to support the inference prediction. However, existing methods tend to provide human-unfriendly interpretation, and are prone to sub-optimal performance due to one-side promotion, i.e. either inference promotion with interpretation or vice versa. In this paper, we propose a multi-level **M**utual **P**romotion mechanism for self-evolved **I**nference and sentence-level **I**nterpretation (MPII). Specifically, from the model-level, we propose a Step-wise Integration Mechanism to jointly perform and deeply integrate inference and interpretation in an autoregressive manner. From the optimization-level, we propose an Adversarial Fidelity Regularization to improve the fidelity between inference and interpretation with the Adversarial Mutual Information training strategy. Extensive experiments on NLI and CQA tasks reveal that the proposed MPII approach can significantly outperform baseline models for both the inference performance and the interpretation quality.[1]

1 Introduction

Recently, the interpretability of neural networks has been of increasing concern. In order to break the black-box of neural networks, many works explore the interpretability of neural networks through providing interpretations to support their inference results (Ribeiro et al., 2016; Chen et al., 2018; Liu et al., 2019; Thorne et al., 2019; Kumar and Talukdar, 2020).

Although prior works have made some progress towards interpretable NLP, they tend to provide interpretations that lack human-readability. Existing interpretable models usually extract prominent features or select input key words as explanations, such as attention distribution (Xu et al., 2015), heatmap (Samek et al., 2017), alignment rationale (Jiang et al., 2021), gradients (Li et al., 2016), magnitude of hidden states (Linzen et al., 2016), etc. Considering readability and comprehensibility for humans, some works turn to generate token-level explanations (Liu et al., 2019; Thorne et al., 2019), which are nevertheless prone to cause ambiguity. Figure 1 shows some prevalent forms of interpretations in NLI task. Obviously, human language interpretations seem more acceptable than those chaotic maps, whether it is heatmap or alignment map. As for the token-level interpretation, several discrete tokens without any logical links are vague and ambiguous. Moreover, Thorne et al. (2019) observed that token-level methods tend to predict common tokens (e.g. people, man, dog)

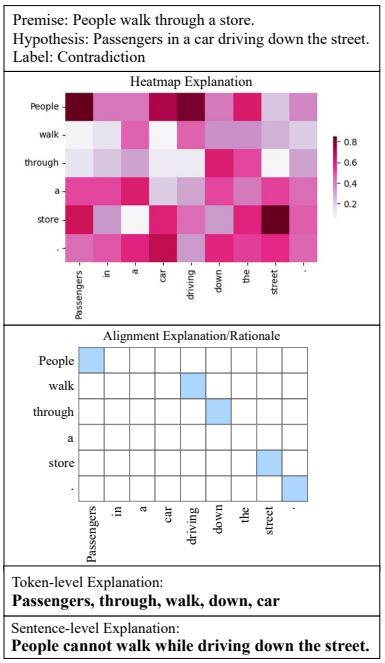

Figure 1: Comparison of different interpretations: heatmap explanation, alignment rationale, token-level NL explanation, and sentence-level NL explanation.

*Work was done during internship at Microsoft Research
[1]Our code is available at https://github.com/theNamek/MPII.git

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7074 - 7084
May 22-27, 2022

rather than keywords. Intuitively, human language sentence-level interpretations containing reasoning logic are the best form for human to understand.

With annotated natural language interpretation datasets available (Camburu et al., 2018; Rajani et al., 2019), methods of generating sentence-level interpretation have been explored recently. Camburu et al. (2018) proposed to first generate interpretation and then predict the label only based on the generated interpretation. Kumar and Talukdar (2020) proposed to first generate sentence-level interpretations with deep pre-trained language models (such as BERT and GPT), then fed those interpretations as extra knowledge to help improve inference performance. We notice that these methods only include one-side promotion: utilizing information contained in interpretation to improve inference, while ignoring the other-side promotion: using inference logic to enhance interpretation. As claimed in Kumar and Talukdar (2020) that their one-side promotion improves predictions' faithfulness to generated interpretations, then the other-side should be able to improve interpretation's faithfulness to inference process. This has aroused our thinking: *Can we deeply fuse these two relevant tasks with ingenious combination skills and achieve mutual promotion for inference and interpretation?*

In this paper, we propose a multi-level Mutual Promotion mechanism for self-evolved Inference and sentence-level Interpretation (MPII). Specifically, from the model-level, we propose a Stepwise Integration Mechanism (SIM) to iteratively update the inference prediction and generate an interpretation token at each decoding step, and deeply integrate hidden representations of the prediction and the token with two fusion modules. In this way, the model learns to refine the inference conclusion as the interpretation proceeds, and the inference procedure can in turn guide the generation of interpretation at each decoding step. From the optimization-level, we propose an Adversarial Fidelity Regularization (AFiRe) to improve the fidelity between inference and interpretation with the Adversarial Mutual Information (AMI) method (Pan et al., 2020), which extends the maximum mutual information optimization objective with the idea of generative adversarial network (Goodfellow et al., 2014). With this training framework, the model is trained against a smart backward network that learns to reward the inference prediction and interpretation of fidelity, which ensures faithfulness and makes the derived interpretation depict the true profile of how the model works (Jiang et al., 2021).

To verify the effectiveness of MPII, we conduct extensive experiments on two inference tasks: Natural Language Inference (NLI) task and Commonsense Question Answering (CQA) task. Experiment results reveal that compared with baseline models, our method can achieve mutual promotion on both model inference performance and sentence-level interpretation quality. Meanwhile, through providing simultaneous inference prediction and human-comprehensible interpretation with deep integration mechanism and adversarial training strategy, our model can perform inference and interpretation of fidelity and generate more robust explanations. Main contributions of this work include:

- Different from the previous works that only include one-side promotion, we mutually promote the inference and sentence-level interpretation from both the model-level and the optimization-level.
- We propose a Stepwise Integration Mechanism to tightly fuse latent prediction and interpretation information at every decoding step, and an Adversarial Fidelity Regularization to further improve the fidelity with the adversarial training strategy.
- Experiment results show that our method achieves significant improvement in both inference accuracy and interpretation quality compared with baseline models.

2 Methodology

In this section, we introduce Stepwise Integration Mechanism (SIM) and Adversarial Fidelity Regularization (AFiRe) in details. Utilizing the autoregressive nature of Transformer decoder, SIM enables deep interaction at every decoding step between inference and interpretation. With the adversarial training strategy, AFiRe enables further integration of latent semantic information between inference and interpretation, and also improves the quality of explanation sentences by bringing them closer to human expressions.

2.1 Task Description

Transformer model (Vaswani et al., 2017) has been firmly established as the dominant approach in text generation tasks, we therefore adopt the Transformer model as backbone. Given a sequence of tokens as input $\mathbf{X} = \{x_0, x_1, ..., x_m\}$ (e.g. for NLI:

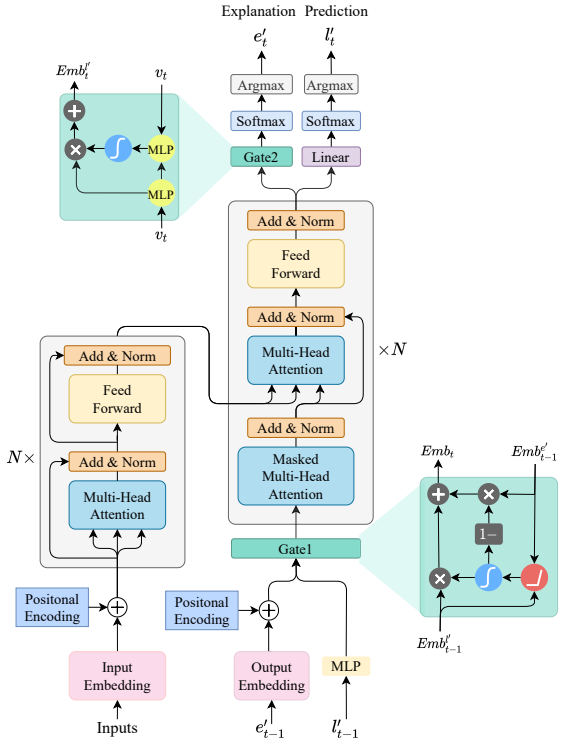

Figure 2: The overall architecture of our model. Both prediction label and explanation token are generated at every decoding step. Two fusion gates are attached to enable deep interaction of their hidden representations.

$\mathbf{X} = \{[\text{CLS}] + \text{Premise} + [\text{SEP}] + \text{Hypothesis}\}$, for CQA: $\mathbf{X} = \{[\text{CLS}] + \text{Question} + [\text{SEP}] + \text{Answers}\}$), Transformer encoder produces a sequence of continuous vectors $\mathbf{H}_{enc}$. Conditioned on $\mathbf{H}_{enc}$, on each decoding step, Transformer decoder takes the embedding of words generated by previous steps as input and predicts the word for current step.

With ground truth prediction $\mathbf{L}$ and explanation $\mathbf{E}$ from human-annotated dataset, the interpretable model is required to generate prediction $\mathbf{L}'$ and explanation sentence $\mathbf{E}' = \{e'_0, e'_1, ..., e'_n\}$ simultaneously.

2.2 Stepwise Integration Mechanism

Prevalent interpretable models share the same encoder and separately adopt a MLP and a decoder to generate predictions and explanations. We analogously adopt the standard Transformer encoder, but apply Stepwise Integration Mechanism to deeply integrate standard MLP and Transformer decoder at every decoding step to simultaneously produce predictions and explanations.

As depicted in Figure 2, at decoding step t, decoder takes the last generated token e'_{t-1} and the predicted label l'_{t-1} at previous step as input. At the first decoding step, we pass the encoder hidden state corresponding to [CLS] token into MLP to get the l'_0. We project the label l'_{t-1} with Multi-Layer Perceptrons (MLP) and obtain $\mathbf{v}^p_{t-1}$, which represents the previous step prediction information. We then fuse the prediction information $\mathbf{v}^p_{t-1}$ and the explanation token e'_{t-1} with gate mechanism. The gate probability at t step is computed by:

$$p'_t = \text{ReLU}(\mathbf{W}_1\,[\mathbf{Emb}^{l'}_{t-1}; \mathbf{Emb}^{e'}_{t-1}] + \mathbf{b}_1) \quad (1)$$

$$p_t = \sigma(\mathbf{W}_2\, p'_t + \mathbf{b}_2) \quad (2)$$

where ";" means concatenation, $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{b}_1$ and $\mathbf{b}_2$ are trainable parameters. $\text{ReLU}(\cdot)$ here denotes the ReLU activation function (Nair and Hinton, 2010), $\sigma(\cdot)$ represents the sigmoid function. We fuse the prediction and interpretation information as below:

$$\mathbf{Emb}_t = p_t \mathbf{Emb}^{l'}_{t-1} + (1 - p_t)\mathbf{Emb}^{e'}_{t-1} \quad (3)$$

where $\mathbf{Emb}_t$ contains the information of prediction and the overall explanation sub-sequence generated in all previous steps.

We utilize the stack of masked self-attention layers f_{sa} used in Transformer decoder to compute the decoder hidden states:

$$\{\mathbf{h}_0, \mathbf{h}_1, ..., \mathbf{h}_t\} = f_{sa}(\{\mathbf{Emb}_0, \mathbf{Emb}_1, ..., \mathbf{Emb}_t\}) \quad (4)$$

The attention vector referring to the source sequence is computed with multi-head attention:

$$\mathbf{v}_t = f_{mha}(\mathbf{H}_{enc}, \mathbf{h}_t) \quad (5)$$

where $\mathbf{H}_{enc}$ represents the encoder hidden states, f_{mha} denotes the multi-head attention module. The $\mathbf{v}_t$ is further passed into a fully connected layer followed with $softmax$ function to obtain the vocabulary distribution of generated explanation token e'_t at t step:

$$e'_t = argmax(softmax(\mathbf{W}\mathbf{v}_t + \mathbf{b})) \quad (6)$$

where $\mathbf{W}$ and $\mathbf{b}$ are both trainable parameters.

The gate mechanism is then used to integrate the explanation information to update the prediction information:

$$p_t = \sigma(\mathrm{MLP}_1([\mathbf{Emb}^{l'}_{t-1}; \mathrm{MLP}_2(\mathbf{v}_t)])) \quad (7)$$

where the two $\mathrm{MLP}(\cdot)$ use different parameters.

$$\mathbf{Emb}^{l'}_t = \mathbf{Emb}^{l'}_{t-1} + p_t \mathrm{MLP}_3(\mathbf{v}_t) \quad (8)$$

We apply the residual connection (He et al., 2016) here, which is easier to optimize in the scenario of many decoding steps. This is similar to the gate mechanism used in Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) that learns to remember important information obtained on each decoding step. At the last decoding step, the model deduces the eventual decision:

$$L' = argmax(softmax(\mathbf{Emb}^{l'}_n)) \quad (9)$$

where n is the length of the generated explanation $\mathbf{E}'$. With this setting, both prediction and explanation are updated at every decoding step. The step-by-step explanation helps the model to do better inference, and the stepwise inference in turn guides the generation of better explanation.

2.3 Adversarial Fidelity Regularization

From the level of optimization objective, we further introduce the Adversarial Fidelity Regularization (AFiRe) to improve the fidelity of inference and interpretation. We leverage the Adversarial Mutual Information (AMI) method (Pan et al., 2020) to extend the maximum mutual information objective among input, inference prediction and the generated explanation with the idea of generative adversarial network (Goodfellow et al., 2014).

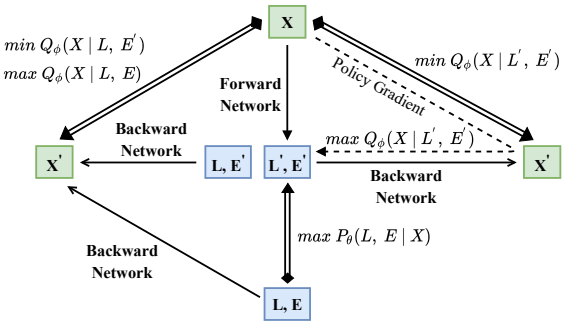

Figure 3: The overview of Adversarial Fidelity Regularization.

Compared to the maximum likelihood estimation (MLE) objective, maximum mutual information (MMI) objective encourages the model to generate the prediction and explanation that are more faithful to the input (Kinney and Atwal, 2014; Stratos, 2019). The mutual information $I(X, L, E)$ among the input X, inference label L and explanation E is formulated as:

$$\begin{aligned} I(X, L, E) &= \mathbb{E}_{P(X,L,E)}\left[\log \frac{P(X, L, E)}{P(X)P(L, E)}\right] \\ &= H(X) - H(X|L, E) \end{aligned}$$

where H denotes the entropy.

Because of the intractability of directly estimating the mutual information in high-dimensional space, we approximate the optimization objective with a Variational Information Maximization lower bound (Chen et al., 2016b; Zhang et al., 2018; Poole et al., 2019):

$$\begin{aligned} \mathrm{I(X, L, E)} &= \mathrm{H(X)} + \mathbb{E}_{P(X,L,E)}\left[\log P(X|L, E)\right] \\ &= \mathrm{H(X)} + \mathbb{E}_{P(X,L,E)}\left[\log Q_\phi(X|L, E)\right] \\ &\quad + \mathbb{E}_{P(L,E)}\left[\mathcal{KL}(P(X|L, E)||Q_\phi(X|L, E))\right] \\ &\geq \mathrm{H(X)} + \mathbb{E}_{P(X)}\mathbb{E}_{P_\theta(L,E|X)}\left[\log Q_\phi(X|L, E)\right] \end{aligned}$$

where $\mathcal{KL}(\cdot||\cdot)$ denotes the Kullback-Leibler (KL) divergence between two distributions. $P_\theta(L, E|X)$ and $Q_\phi(X|L, E)$ denote the forward network (generating L, E conditioned on X) and the backward network (generating X conditioned on L, E) respectively.

Since the entropy term $H(X)$ associates with the training data and does not involve the parameters we optimize, the objective of MMI is equivalent as:

$$\max_{\theta,\phi} \mathbb{E}_{(L',E')\sim P_\theta(L',E'|X)}\left[\log Q_\phi(X|L', E')\right]$$

where θ and ϕ are the parameters of the forward and backward network respectively. L' and E' represent the synthetic prediction label and explanation generated by the forward network.

With the MMI optimization objective, the backward network is trained with only the synthetic label and explanation produced by the forward network, and prone to sub-optimal performance if the synthetic text is uninformative. Since the the backward network provides a reward for optimizing the forward network, a biased backward network may provide unreliable reward scores and mislead the forward network optimization.

To remedy this problem, we leverage the Adversarial Mutual Information (AMI) method (Pan et al., 2020) to extend MMI with the idea of generative adversarial network (Goodfellow et al., 2014).

Specifically, we first bring the min-max adversarial game into training procedure and add an additional objective term $\mathrm{Q}_\phi(X|L,E)$ to maximize the negative likelihood of Q_ϕ when feeding it with the real data:

$$\min_\phi \max_\theta \ \mathbb{E}_{(L',E')\sim P_\theta(L',E'|X)}\left[\log \mathrm{Q}_\phi(X|L',E')\right] - \mathrm{Q}_\phi(X|L,E)$$

With this interactive training strategy and regularizing the backward network with both the synthetic data and real data, the forward network will be trained against a smarter backward network that only rewards prediction and explanation of fidelity.

Besides, we add an objective term $\mathrm{P}_\theta(\mathrm{L},\mathrm{E}|X)$ of maximize the negative likelihood of P_θ to balance the positive samples as teacher-forcing algorithm (Li et al., 2017). The final optimization objective is formulated as:

$$\min_\phi \max_\theta \ \mathrm{P}_\theta(\mathrm{L},\mathrm{E}|X) + \overbrace{\underbrace{\mathbb{E}_{(L',E')\sim P_\theta(L',E'|X)}\left[\log \mathrm{Q}_\phi(X|L',E')\right] - \mathrm{Q}_\phi(X|L,E)}_{\text{Adversarial Training}}}^{\text{Mutual Information}}$$

As depicted in Fig 3, to encourage the forward network to learn a stronger connection between generated explanations and model predictions, we also add $Q_\phi(X|L,E')$ as negative samples for backward network. This explicitly encourages the backward network to be capable of punishing the P_θ when it generates unfaithful explanations.

3 Experiments

We intend to verify the mutual promotion effect of SIM and AFiRe on the inference ability and interpretablity of model. We choose two tasks requiring inference ability: Natural Language Inference (NLI) and Commonsense Question Answering (CQA).

3.1 Datasets

We use six datasets as our testbeds: **SNLI** (Bowman et al., 2015), **e-SNLI** (Camburu et al., 2018), **CQA** (Talmor et al., 2019), **CoS-E** (Rajani et al., 2019), **MultiNLI** (Williams et al., 2018), and **SICK-E** (Marelli et al., 2014).

SNLI is a standard benchmark for NLI task, while e-SNLI extends it with human-annotated natural language explanations for each sentence pair. CoS-E[2] dataset extends CQA dataset with natural language explanations for each QA sample. MultiNLI is another large-scale NLI corpus, which includes a diverse range of genres. SICK-e (Sentences Involving Compositional Knowledge for entailment) provides sentence pairs that are rich in the lexical, syntactic and semantic phenomena. The latter two datasets are used for out-of-domain evaluation.

3.2 Baselines

NLI: We use e-INFERSENT and Transformer as two baseline models for NLI task. The e-INFERSENT model adds a LSTM decoder into INFERSENT (Conneau et al., 2017) for explanations. The classification module and the explanation generation module are separated but share the same encoder. The Transformer model (Vaswani et al., 2017) adds a MLP layer for making predictions. With this baseline, we aim to test whether vanilla transformer without further interaction can achieve good results.

CQA: We use CAGE (Rajani et al., 2019) as the baseline model for CQA task. CAGE adopts the explain-then-predict approach, which firstly finetunes a deep pretrained language model GPT (Radford et al., 2019) to generate explanations, then use a classifier to predict the inference label with the generated explanation and source text as the input.

3.3 Metrics

To evaluate inference performance, we report **Task-specific Accuracy** (NLI Accuracy and CQA Accuracy). To evaluate the quality of generated interpretation, we report **BLEU** (similarity between generation and ground truth), **PPL** (fluency of generated sentences), and **Inter Repetition** (diversity of generated explanations).

[2]https://github.com/salesforce/cos-e

Methods	Inference	Interpretation		
	Task-Accuracy↑	BLEU↑	PPL↓	Inter-Rep↓
NLI Task				
Dataset†	100.00	22.51	30	0.40
e-INFERSENT‡	83.96	22.40	**24**	0.72
Transformer	80.12	23.63	68	0.69
Transformer + MPII (w/o Inference in SIM)	-	28.31	38	0.56
Transformer + MPII (w/o Interpretation in SIM)	85.43	-	-	-
Transformer + MPII (w/o AFiRe)	86.47	27.93	41	0.64
Transformer + MPII	**87.32**	**28.64**	37	**0.52**
BART + MPII (w/o AFiRe)	89.79	31.01	29	0.59
BART + MPII	**91.85**	**31.26**	**27**	**0.51**
Δ	**11.73**↑	**7.63**↑	**41**↓	**0.18**↓
CQA Task				
Dataset†	100.0	100.0	454	0.16
CAGE‡	58.15	4.37	**129**	0.36
BART + MPII (w/o AFiRe)	52.83	3.54	227	**0.13**
BART + MPII	**60.21**	**4.92**	196	**0.15**
Δ	**2.06**↑	**0.55**↑	**67**↑	**0.21**↓

Table 1: Automatic evaluation results on the SNLI and CQA datasets with the annotated explanation from the e-SNLI and CoS-E datasets. The higher↑ (or smaller↓) score indicates the better performance. †We evaluate the ground truth with our metrics. ‡We use the released baseline model and evaluate it with our metrics. Δ indicates the improvement over the Transformer/CAGE baselines.

3.4 Main Results

Table 1 shows automatic evaluation results on the SNLI and CQA datasets with the annotated explanation from the e-SNLI and CoS-E datasets. Compared with the baseline models, our MPII method can achieve significant performance improvement for both the inference and interpretation on two tasks. It indicates that the inference and interpretation process can be mutually promoted with our proposed method. With the ablation study, we notice a performance degradation of the inference and interpretation if we remove either of them, demonstrating the faithfulness between the generated explanation and the model's prediction.

Inference Promotion: We can achieve 11.73 and 2.06 absolute inference accuracy improvements compared to the baselines for the NLI and CQA task, respectively. For the NLI task, with our MPII framework, the Transformer baseline model can improve over 5 absolute accuracy score. The ablation study shows the contribution comes from not only the mutual interaction of inference and interpretation in the Stepwise Integration Mechanism (SIM), but also the adversarial mutual information training objective introduced in the Adversarial Fidelity Regularization (AFiRe). Moreover, with parameters initialized with the pretrained BART model, the accuracy can be further improved by a 4.53 absolute score. For the CQA task, we observe that better performance is still achieved compared with the CAGE baseline model. If we remove the AFiRe, a significant inference degradation would be witnessed. It also indicates the effectiveness of AFiRe for utilizing interpretability to improve the inference ability.

Methods	MultiNLI	SICK-E
Transformer	55.92	53.21
Transformer + MPII (w/o AFiRe)	56.42	53.84
Transformer + MPII	**58.73**	**56.54**

Table 2: Out-of-domain NLI evaluation results on MultiNLI and SICK-E datasets.

Interpretation Promotion: The quality of generated interpretation can also be significantly improved with our mutual promotion method on both NLI and CQA tasks. For the NLI task, combined with our MPII, the Transformer baseline model can provide more accurate, fluent and diverse interpretation with much better results in all metrics. Similar with the inference results, the ablation study shows that both SIM and AFiRe contribute to the performance improvement. With the pretrained BART model, we further improve the BLEU and Inter-Rep performance and get comparable PPL compared with the e-INFERSENT model. For the CQA task, our method performs better in terms of BLEU score and the diversity of generated explanations. We notice that the BLEU scores are pretty low for CQA task, which may stem from the free form of expression for explanations in the dataset, i.e. several different explanations share the same

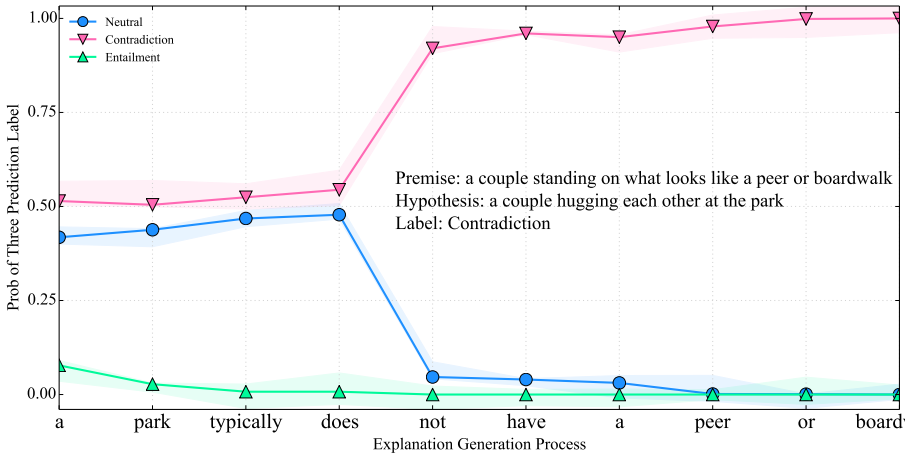

Figure 4: Visualization for mutual promotion evolution of inference and interpretation.

Methods	Critic-Score↑
e-INFERSENT‡	82.41
Transformer	82.27
Transformer + MPII (w/o AFiRe)	92.09
Transformer + MPII (w/o Interpretation in SIM)	93.81
Transformer + MPII (w/o Inference in SIM)	94.50
Transformer + MPII	**95.93**

Table 3: Fidelity evaluation results on SNLI dataset.

commonsense knowledge. We observe that most of the explanations generated by our method are reasonable enough to interpret the predictions even though the BLEU scores are low. Our method also achieves a smaller Inter-Rep score, which shows that our model can provide more diverse explanations to reveal the inference process of making predictions.

3.5 Out-of-Domain Evaluation

As shown in Table 2, we evaluate our method with the Transformer baseline model on two out-of-domain datasets: MultiNLI and SICK-E. The results show that our mutual promotion method enables the Transformer model to be more robust, and achieves about 3 absolute accuracy improvement on both of the out-of-domain datasets without fine-tuning. It is because with our MPII method, the model can generate more reliable and domain-related interpretation, which helps to make more accurate inference prediction. The ablation results demonstrate both the adversarial mutual information training strategy in AFiRe and deep integration in SIM is very effective to improve the model's generalization and robustness.

3.6 Fidelity Evaluation

We propose a model-based evaluation metric *Critic-Score* to evaluate the fidelity between model's inference predictions and interpretations. Inspired by Shen et al. (2017), which applied a trained model to automatically evaluate the text style transfer accuracy in the absence of parallel dataset, we pre-train a well-performed discriminator model to evaluate the fidelity between the predicted label and the generated explanation.

The discriminator is a binary classifier $f : (X, L, E) \mapsto$ Yes/No , which shares similar architecture with the backward network in our Adversarial Fidelity Regularization (Section 2.3). The training dataset is constructed based on the e-SNLI and CoS-E corpus. Given a sample $\langle X_i, L_i, E_i \rangle$ on e-SNLI that serves as a positive sample, we build the negative sample as $\langle X_i, L_i, E_j \rangle$, where explanation $E_j \neq E_i$ is selected from another e-SNLI sample that shares either the same premise or hypothesis. With this dataset, the discriminator model is trained to learn the intrinsic fidelity between the label and its corresponding explanation. The trained discriminator achieves 97% accuracy on its test set and is able to serve as a quantitative way of evaluating fidelity.

As shown in Table 3, with our proposed mutual promotion method, the Transformer model can achieve significant improvement on Critic-Score between prediction and explanation. The ablation results confirm both the deep interaction design in Stepwise Integration Mechanism and the adversarial training strategy in Adversarial Mutual Information can contribute to the improvement of fidelity and faithfulness.

Input X: the premise is : a statue at a museum that no seems to be looking at , while the hypothesis is : the statue is offensive and people are mad that it is on display . **Choices**: (0) entailment, (1) neutral, (2) contradiction.		
Method	Prediction	Explanation
Human:	(1)	not all statues are ignored because they are offensive .
e-INFERSENT:	(0)	people is _end_ .
MPII (w/o AFiRe):	(1)	just because a statue at a museum that no seems to be looking at does n't imply the statue is offensive and people are mad .
MPII:	(1)	a statue that no seems to be looking at it does not imply the statue is offensive or people are mad .
Input X: the premise is : a young family enjoys feeling ocean waves lap at their feet . while the hypothesis is : a family is out at a restaurant . **Choices**: (0) entailment, (1) neutral, (2) contradiction.		
Method	Prediction	Explanation
Human:	(2)	one says the family is inside the restaurant , while the other says they are feeling ocean waves .
e-INFERSENT:	(2)	a family can not be enjoying their feet out at a restaurant and at the same time
MPII (w/o AFiRe):	(2)	the family can not be at the beach and at a restaurant at the same time .
MPII:	(2)	the family can not be feeling ocean waves and be at a restaurant at the same time .

Table 4: Examples randomly selected from the test set of e-SNLI. "Human" here represents the explanations from dataset. Blue and red indicate good and bad explanations respectively, except for those provided by human.

Input X: what do people aim to do at work ? **Choices**: (0) complete job, (1) learn from each other, (2) kill animals, (3) wear hats (4) talk to each other.		
Method	Prediction	Explanation
Human:	(0)	people aim to complete job in work .
CAGE:	(2)	people aim to kill animals .
BART + MPII (w/o AFiRe):	(0)	people want to complete job .
BART + MPII:	(0)	what people aim to do at work is to complete job .
Input X: the lock kept the steering wheel from moving, but the thief still took his chances and began to work on the what ? **Choices**: (0) keep cloesd, (1) train, (2) ignition switch, (3) drawer, (4) firearm.		
Method	Prediction	Explanation
Human:	(2)	it was more safe .
CAGE:	(0)	the lock kept the steering wheel from moving, but the door kept the door from moving .
BART + MPII (w/o AFiRe):	(2)	the lock kept the steering wheel from moving ignition switch .
BART + MPII:	(2)	the ignition switch is the only thing that would work on a car .

Table 5: Randomly selected examples in CQA task. "Human" here represents the explanations from dataset. Blue and red indicate good and bad explanations respectively, except for those provided by human. Predicted label is presented in the parentheses.

3.7 Analysis

Mutual Promotion Visualization: Figure 4 demonstrates the evolution of the inference prediction as the interpretation proceeds. The input of the model is "[CLS] a couple standing on what looks like a peer or boardwalk [SEP] a couple hugging each other at the park", of which the ground truth label is "contradiction". We observe that the model draws an initial conclusion that the entailment relationship between the premise and the hypothesis is not "entailment", and is not able to tell whether it is "neutral" or "contradiction". As the deliberation proceeds, our model comes to judge that it is "contradiction" with the generated interpretation "a park does not have a peer or boardwalk". From the clear split of the red and blue lines when "does" and "not" are generated, we can see that the prediction is very sensitive to explanation, which demonstrates the faithfulness (Kumar and Talukdar, 2020).

Semantic Similarity Evaluation of Interpretation: To better evaluate the quality of generated explanations, we also measure the cosine similarity between generated explanations and human annotated explanations. The results are presented in Fig 5. The cosine similarity of our method concentrates on 0.9 and achieves higher scores than CAGE, which demonstrates the effectiveness of our MPII for generating better interpretation that are closer to human expression.

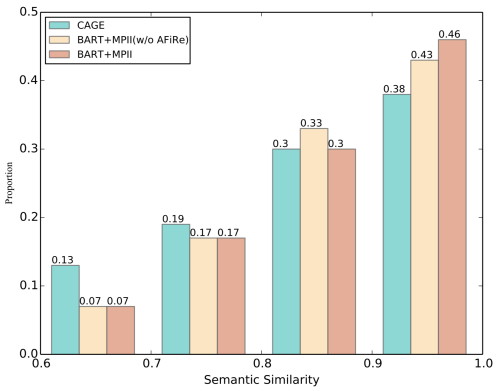

Figure 5: The distribution of cosine similarity with average sentence embedding between human annotation and generated interpretation.

Case Study Table 4 presents examples produced by different models. For the first example, e-INFERSENT fails to make correct prediction and provide reasonable explanation. In contrast, our MPII not only predict the entailment relation correctly, but also produce faithful explanations to

Methods	Fidelity-C$^{\uparrow}$	Fidelity-W$^{\uparrow}$	LAcc$^{\uparrow}$	Fluency$^{\uparrow}$
NLI Task				
e-INFERSENT	3.16	2.74	3.34	4.23
Transformer	3.30	3.21	3.65	3.68
Transformer + MPII (w/o AFiRe)	4.01	4.12	4.33	4.36
Transformer + MPII	**4.17**	**4.38**	**4.57**	**4.51**
CQA Task				
CAGE(GPT, ETP)	3.71	3.18	3.52	4.25
BART + MPII (w/o AFiRe)	4.26	4.13	4.05	4.21
BART + MPII	**4.37**	**4.39**	**4.22**	**4.30**

Table 6: Human evaluation results on Fidelity-C(fidelity between correct prediction and corresponding interpretation), Fidelity-W(fidelity between wrong prediction and corresponding interpretation), LAcc(accuracy of selecting correct lables when only given the generated interpretations), Fluency(fluency of interpretation).

interpret predictions. For the second example, our MPII and MPII with AFiRe removed still capture the entailment relation well, and explain that "at the beach" and "at restaurant" can not be done at the same time. As we can see, these explanations generated by our method are also fluent.

Table 5 shows the randomly selected examples generated by different models in the CQA task. For the first example, CAGE makes wrong prediction, and generates explanation that obviously conflicts with common knowledge. In contrast, our method can make correct predictions and generate more reasonable explanations. Similarly for the second example, CAGE seems to directly copy words from the question that do not actually contain meaningful information. Our MPII still explains well, but fails to explain properly with AFiRe removed, even if the explanation contains the correct answer, which reveals the importance of AFiRe for promotion of interpretation.

Human evaluation: We conduct human evaluation to further evaluate the effectiveness of MPII. We randomly selected 300 examples from the test set of e-SNLI, and asked 4 well-educated annotators to rate every sample with 4 metrics on a 1-5 Likert scale in a strictly blind fashion (Stent et al., 2005). As shown in Table 6, analogous to automatic evaluation results (Section 3.4), our MPII can generate interpretations with best quality and fidelity to corresponding inference predictions, whether correct or wrong.

4 Related Work

With the great success of natual language inference, many recent works explore the interpretability of neural networks through providing interpretation to support their inference results (Ribeiro et al., 2016; Chen et al., 2018; Liu et al., 2019; Thorne et al., 2019; Kumar and Talukdar, 2020). Three forms of interpretation are provided by these works: (1) feature-based interpretation (Chen et al., 2016a, 2018; Ribeiro et al., 2016, 2018; Li et al., 2016; Nguyen, 2018; Feng et al., 2018; Gururangan et al., 2018) such as attention distribution (Xu et al., 2015), heatmap (Samek et al., 2017), alignment rationale (Jiang et al., 2021), gradients (Li et al., 2016), magnitude of hidden states (Linzen et al., 2016), etc.; (2) token-level interpretation that relatively easy to comprehend but prone to ambiguity (Ribeiro et al., 2016; Liu et al., 2019; Thorne et al., 2019), and (3) sentence-level interpretation which has the best human-readability Camburu et al. (2018); Talmor et al. (2019); Kumar and Talukdar (2020). Different from the previous work which only include one-side promotion, we proposed the mutual promotion mechanism that can improve the performance of both inference and sentence-level interpretation.

5 Conclusions

In this work, we propose to mutually promote model inference ability and interpretability from multi-levels. From the model-level, we propose Stepwise Integration Mechanism to enable the model to refine the prediction conclusion as the explaining proceeds and also to guide the generation of better explanation with the inference procedure of reaching prediction conclusion. From the optimization-level, we propose an Adversarial Fidelity Regularization, which leverages the Adversarial Mutual Information method to improve the fidelity between the inference and interpretation, which further guarantees faithfulness. Experiment results show the effectiveness of our proposed method on both NLI and CQA tasks. Future work will involve extending our approaches into other tasks of NLP. We hope that our work can encourage further research in this direction.

Acknowledgements

We would like to acknowledge Chuhan Wu for the helpful discussion. We also want to thank Jiale Xu for his kindness and help.

References

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), pages 632–642.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, pages 9539–9549.

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016a. A thorough examination of the cnn/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367.

Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. 2018. Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 882–891.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016b. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. 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)*, pages 107–112.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zhongtao Jiang, Yuanzhe Zhang, Zhao Yang, Jun Zhao, and Kang Liu. 2021. Alignment rationale for natural language inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5372–5387, Online. Association for Computational Linguistics.

Justin B Kinney and Gurinder S Atwal. 2014. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4(1):521–535.

Hui Liu, Qingyu Yin, and William Yang Wang. 2019. Towards explainable nlp: A generative explanation framework for text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5570–5581.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 1–8.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Icml*.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. 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 1069–1078.

Boyuan Pan, Yazheng Yang, Kaizhao Liang, Bhavya Kailkhura, Zhongming Jin, Xian-Sheng Hua, Deng Cai, and Bo Li. 2020. Adversarial mutual information for text generation. In *International Conference on Machine Learning*, pages 7476–7486. PMLR.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. 2019. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 2019 Conference of the Association for Computational Linguistics (ACL2019)*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, pages 6830–6841.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *international conference on intelligent text processing and computational linguistics*, pages 341–351. Springer.

Karl Stratos. 2019. Mutual information maximization for simple and accurate part-of-speech induction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1095–1104.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. 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, Volume 1 (Long and Short Papers)*, pages 963–969.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems*, pages 1815–1825.