Primary open-angle glaucoma diagnosis from fundus photographs using siamese network

Mingquan Lin[1] MIL4012@MED.CORNELL.EDU
Lei Liu[2] LEI.LIU@WUSTL.EDU
Mae Gordon[3] MAE@WUSTL.EDU
Michael Kass[3] KASS@WUSTL.EDU
Fei Wang[1] FEW2001@MED.CORNELL.EDU
Sarah H. Van Tassel[4] SJH2006@MED.CORNELL.EDU
Yifan Peng[1†] YIP4002@MED.CORNELL.EDU

[1] *Department of Population Health Sciences, Weill Cornell Medicine*
[2] *Institute for Public Health, Washington University School of Medicine*
[3] *Department of Ophthalmology and Visual Sciences, Washington University School of Medicine*
[4] *Department of Ophthalmology, Weill Cornell Medicine*

Editors: Under Review for MIDL 2022

Abstract

Primary open-angle glaucoma (POAG) is one of the leading causes of irreversible blindness in the United States and worldwide. Although deep learning methods have been proposed to diagnose POAG, these methods all used a single image as input. Differently, the glaucoma specialists compare the follow-up image with the baseline image to determine a glaucomatous eye. To simulate this process, we proposed a siamese network model, POAGNet, to identify POAG from fundus photographs. The POAGNet consists of two side-outputs for deep supervision. The POAGNet network was trained and evaluated on two datasets: (1) 37,339 fundus photographs from 1,636 Ocular Hypertension Treatment Study (OHTS) participants, and (2) 3,684 fundus photographs from the sequential fundus images for glaucoma (SIG) dataset. Extensive experiments show that POAGNet performed better on POAG diagnosis in the OHTS test set with an accuracy of 0.91, F-score of 0.5069, and an AUC of 0.9081 than state-of-the-art (accuracy 0.8320; F-score 0.3864; AUC 0.8750). It also outperformed the baseline in the SIG dataset (Accuracy 0.9176 vs 0.8690; F-score 0.1613 vs 0.1010; AUC 0.7518 vs 0.6434). These results highlight the potential of deep learning to assist and enhance clinical POAG diagnosis. The proposed network will be publicly available on https://github.com/bionlplab/poagnet.

Keywords: Deep learning, Primary open-angle glaucoma (POAG), Fundus photographs, Siamese network.

1. Introduction

Primary open-angle glaucoma (POAG) is one of the leading causes of blindness worldwide (Bourne et al., 2013). In the United States, POAG is the most common form of glaucoma and is the leading cause of blindness among African-Americans (Sommer et al., 1991) and Hispanics (Jiang et al., 2018). POAG is usually asymptomatic except it progresses to a late

† corresponding author

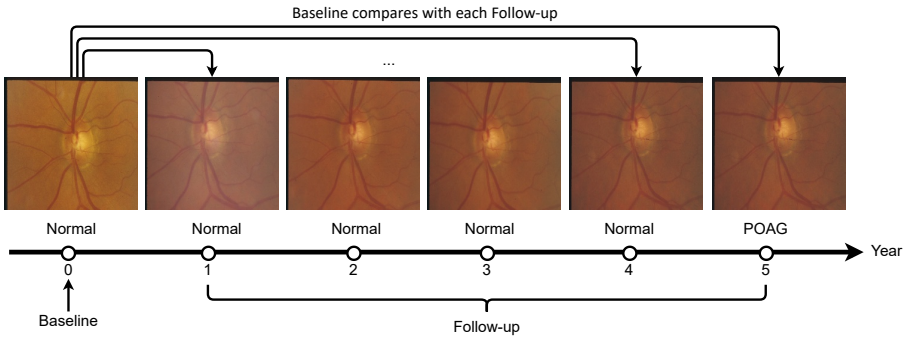

Figure 1: Longitudinal fundus images of a patient.

stage where visual field (VF) loss can happen which may cause a prompt blindness. However, fortunately, it is possible to avoid the most blindness caused by POAG via early diagnosis and treatment (Tatham et al., 2015). Therefore, accurately identifying individuals with glaucoma is crucial to clinical decision-making, which can provide early notice for patient monitoring and medical and surgical treatments (Doshi et al., 2008; Quigley et al., 1992).

Fundus photography is generally considered most helpful for diagnosing glaucoma in eyes showing a classic glaucomatous appearance to expert graders. While it is convenient and inexpensive, its low prevalence and screening limitations are confronted by inexperienced clinicians, making it challenging to conduct daily screenings (Kolomeyer et al., 2021). Therefore, it is important to develop an automatic model to assist clinicians in predicting POAG with high accuracy from the fundus photographs and improve their photograph interpretation skills.

Recently, deep learning methods have demonstrated promising results in biology and medicine (Ching et al., 2018). In the ophthalmology domain, several method have been proposed to detect POAG at its earliest stage (Chen et al., 2015; Li et al., 2019, 2018; Thakur et al., 2020; Christopher et al., 2018). However, these approaches used a single image as input, which might affect the performance of the model. In clinical practice, the glaucoma specialists compare the follow-up image with the baseline image (the image taken at the beginning of a study) to trace out the relevant features. In this paper, we proposed a siamese network model with side output, POAGNet, to simulate this process by comparing the differences between two input images. Different from previous siamese work, POAGNet used a convolution operation, instead of the absolute Euclidean distance, to study the feature difference between two outputs of the network instead of the absolute Euclidean distance between the two outputs. In addition, the POAGNet also consists of side output (Lin et al., 2021) to ease the vanishing gradient problems in training deep models and to drive the hidden layers for favoring discriminative features. To the best of our knowledge, it is the first time in the ophthalmology domain that two fundus images have been utilized and compared for automated glaucoma detection via Siamese networks.

In addition, prior studies may not perform well over real-world problems due to small datasets from a single institution. It makes the methods less generalizable to different

populations and settings. In this paper, we assessed POAGNet on two independent datasets: the Ocular Hypertension Treatment Study cohort (OHTS)(Kass et al., 2002) and Sequential fundus Images for Glaucoma dataset (SIG)(Li et al., 2020).

Interpretability is also a broad topic in the medical image analysis domain. Taking inspiration from the ERASER benchmark (DeYoung et al., 2020), we also curated a new dataset to evaluate why models make POAG predictions. In this dataset, we focus on rationales, i.e., optic disc, that support POAG diagnosis; therefore, mask optic disc over fundus photography. Evaluation on this dataset showed that POAGNet provides rationales aligned with human annotations.

Our work has the following contributions: (1) We leveraged the siamese network to find the differences between two fundus photographs. We then propose a novel POAGNet to jointly fuse two image representations for glaucoma analysis. (2) Our approach achieves superior POAG diagnosis results (90.81% and 75.18% in AUC) against several competitive baselines on two large-scale, multi-institutional benchmarks. Notably, POAGNet actually relies on particular rationales to make predictions. (3) We make codes, models, and pre-processed data publicly available.

The rest of the paper is organized as follows. We describe the POAGNet in Section 2, followed by ourexperimental setup, results, and discussion in Section 3. We conclude with future work in the last section.

2. Methods

2.1. POAGNet architecture

POAGNet comprises two convolutional blocks that share the weight and are followed by seven layers (Figure 2). In the beginning, two fundus image x_1 and x_2 are passed through the convolutional neural network, DenseNet-201 (Huang et al., 2017), respectively. We used the output of last (F_{d1} and F_{d2}) and second last Dense Blocks (F_{d1n} and F_{d2n}). For each output, we concatenated two outputs, followed by use 1×1 convolution, a batch normalization (BN), and rectified linear units (ReLU). In the end, a global average pooling and a fully connected layer with softmax activation is attached.

2.2. Loss function

In this study, we use the binary cross-entropy as the loss function in the POAGNet. While state-of-the-art Siamese networks tend to use contrasting loss or triplet loss when training, our preliminary study found that they were not suited in this task. In addtion, to overcome the severe class imbalance for the POAG classification, we apply the weighted cross-entropy, a commonly used loss function in classification. The adopted weighted cross-entropy was as follow:

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}[\beta y_n \log(\hat{y}_n(x_n, \theta_s)) + (1-\beta)(1-y_n)\log(1-\hat{y}_n(x_n, \theta_s))] \tag{1}$$

N is the number of training examples. β is the balancing factor between positive and negative samples. Here, we used inversely proportional to POAG frequency in the training

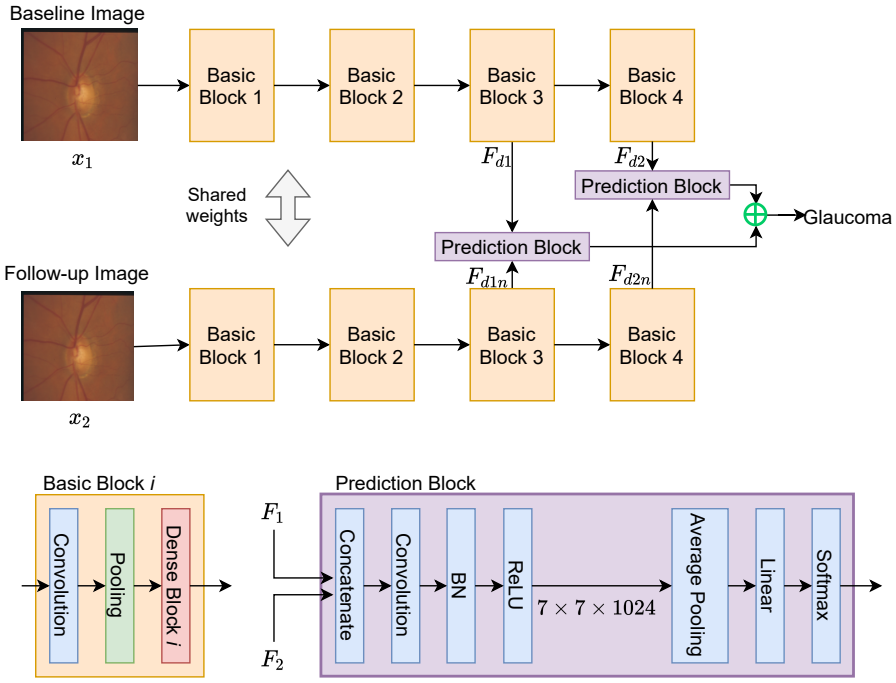

Figure 2: The architecture of the proposed POAGNet.

data. y_n is the ground truth, and $\hat{y}$ is the likelihood predicted by the classifier, and θ_s represents the parameters of the neural network.

The overall loss function is the average of the losses associated with the prediction from the last two blocks:

$$\mathcal{L}_s = \alpha \mathcal{L}^{(1)} + (1 - \alpha)\mathcal{L}^{(2)}, \tag{2}$$

2.3. Evaluation of the model's rationales

In this work, we constructed a new dataset to assess the plausibility of rationales by measuring rationale faithfulness – rationales ought to have meaningfully influenced its prediction (Yu et al., 2019). Specifically, we constructed a contrasting example for a fundus photograph x_i with the bounding box of optic disc r_i masked (Figure 3). Let $m(x_i)$ be the original prediction provided by a model m. We consider the predicted class from the model once the supporting bounding box is stripped. Intuitively, the model should generate a less "correct" class. We then measure this as $\frac{1}{N}\sum_n (m(x_i) - m(x_i/r_i))$. In particular, if we selected glaucomatous images where their labels were predicted by the model correctly ($m(x_i) = 1$), a score of 1.0 indicates that the rationales are indeed influential in the prediction, while 0.0 indicates that the rationales are not the reasons for the prediction.

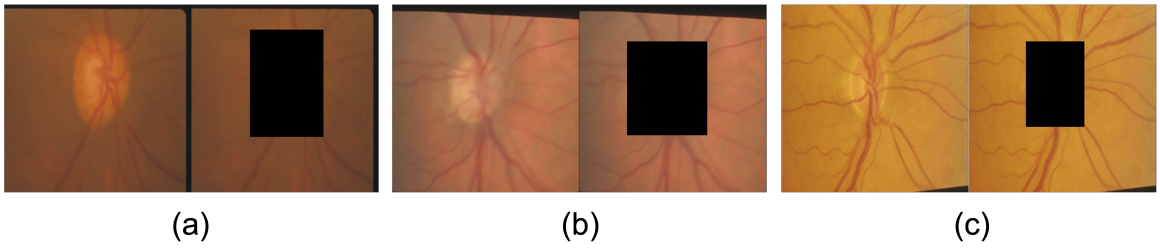

Figure 3: Examples of fundus photographs and their corresponding masked images: (a) POAG due to VF only and (b) POAG due to GON.

3. Results

3.1. Datasets

In this study, we include two independent datasets (Table 1). These two databases are large-scale, cross-sectional, longitudinal, and population-based studies.

The first dataset is obtained from the Ocular Hypertension Treatment Study (OHTS). OHTS is one of the largest longitudinal clinical trials in POAG (1,636 participants and 37,399 images) from 22 centers in the United States. The study protocol was approved by an independent Institutional Review Board at each clinical center (Kass et al., 2002). The participants in this dataset were selected according to both eligibility and exclusion criteria. The gold standard POAG labels were graded at the Optic Disc Reading Center. In brief, two masked certified readers were arranged to detect the optic disc deterioration independently. If there is a disagreement between two readers, a senior reader reviewed it in a masked fashion. The POAG diagnosis in a quality control sample of 86 eyes (50 normal eyes and 36 with progression) showed test-retest agreement at $\kappa = 0.70$ (95% confidence interval [CI], 0.55-0.85). More details of the reading center workflow has been described in Kass et al. (2002). For the OHTS dataset, we split the entire dataset randomly at the patient level. We take one group (20% of total subjects) as the hold-out test set and the remaining as the training set.

The second dataset is obtained from the Sequential fundus Images for Glaucoma (SIG) dataset [1]. SIG contains 3,684 fundus images, of which 153 (4.15%) have POAG. In the SIG dataset, all fundus images are annotated with binary labels of glaucoma, i.e., positive or negative glaucoma. The samples are labeled to positive glaucoma when they satisfy any of the three criteria, i.e., retinal nerve fibre layer defect, rim loss, and optic disc hemorrhage(Li et al., 2020). We used the official training, validation, and testing split in this study.

We compare the baseline image with each follow-up image and they compromise pairs separately. In each pair, the follow-up image and the baseline image come from the same eye.

1. https://github.com/XiaofeiWang2018/DeepGF

Dataset	OHTS			SIG		
	Train	Val	Test	Train	Val	Test
Participants	2503	115	654	300	35	70
Images						
POAG	1774	89	492	110	15	28
Normal	26871	1215	6927	2646	337	701
Pairs						
POAG	1723	86	478	110	15	28
Normal	23620	1068	6289	22236	287	561

Table 1: Characteristics of the OHTS and SIG datasets.

3.2. Evaluation metrics

To evaluate the performance of POAG diagnosis, we compute accuracy, precision, sensitivity (also called recall), specificity, F1-score, and AUC (Area Under the ROC curve).

3.3. Experimental settings

We first trained DenseNet-201 on POAG detection using a single image as input. Then we initialized the subnets in the POAGNet using the DenseNet-201 and fine-tune the entire network in an end-to-end manner.

All images are resized to $224 \times 224 \times 3$ as input of the proposed model. The models were implemented by Keras with a backend of Tensorflow. The proposed network was optimized using the Adam optimizer method (Kingma and Ba, 2014). The learning rate is 5×10^{-5}. α is 0.8. The experiments were performed on Intel Core i9-9960 X 16 cores processor and NVIDIA Quadro RTX 6000 GPU.

3.4. Results and Discussion

3.4.1. POAG diagnosis on the OHTS dataset

We compare our method with three models on POAG diagnosis on the OHTS dataset, including the DenseNet-201 with a single image as input, the traditional Siamese network with Euclidean distance, and POAGNet using the last DenseNet Block (POAGNet w/o side output).

Table 2 shows the performance comparison. Our model achieved the best results, with an accuracy of 0.9100, a precision of 0.4135, a recall of 0.6548, a specificity of 0.9294, an F1-score of 0.5069, and an AUC of 0.9081. Compared to the model with a single image as input (DenseNet-201), POAGNet has higher accuracy (7.80%), precision (15.07%), specificity(9.110%), F1-score (12.23%), and AUC (3.31%).

When compared POAGNet with and without side output, results demonstrated that the side output mechanism could boost the performance of POAGNet.

Method	Accuracy	Precision	Recall	Specificity	F1-score	AUC
DenseNet-201	0.8320	0.2628	0.7291	0.8383	0.3864	0.8750
Siamese network (euclidean distance)	0.8394	0.2581	0.6799	0.8515	0.3740	0.8536
POAGNet w/o side output	0.9372	0.5702	0.4498	0.9742	0.5029	0.8997
POAGNet	0.9100	0.4135	0.6548	0.9294	0.5069	0.9081

Table 2: Comparisons on the test set of OHTS dataset.

3.4.2. POAG diagnosis on the SIG dataset

Table 3 compares the results of POAGNet with DenseNet-201 on the SIG dataset. Our model obtained the better results, with an accuracy of 0.9176, a precision of 0.1471, a recall of 0.1786, an F1-score of 0.1613, and an AUC of 0.7518. Compared to the baseline (DenseNet-201), POAGNet has higher accuracy (5.86%), precision (7.67%), specificity(6.14%), F1-score (6.03%), and AUC (10.84%).

Method	Accuracy	Precision	Recall	Specificity	F1-score	AUC
DenseNet-201	0.8590	0.0704	0.1786	0.8905	0.1010	0.6434
POAGNet	0.9176	0.1471	0.1786	0.9519	0.1613	0.7518

Table 3: Comparisons on the test set of the SIG dataset.

3.4.3. Evaluation on the model's rationales

To evaluate the model's rationales, we randomly selected 100 "correctly predicted" fundus photographs from the OHTS dataset. We then manually masked the disc of these photographs and applied POAGNet to the masked images. POAGNet can achieve a score of 97%, suggesting that optic discs are needed to make the POAG diagnosis.

4. Conclusions

In conclusion, this study proposed a new end-to-end deep learning network that simulates the process for automatic POAG detection from fundus photographs. Two datasets were used to evaluate the proposed model. The results demonstrated that the proposed network has a good performance on POAG diagnosis. Although deep learning models are often considered "black-box" entities, we aimed to improve the transparency of our algorithm by constructing a new dataset by masking the optic disc on the fundus photography. These "contrasting" examples help us understand if the rationales (optic disc) are indeed influential in the POAG diagnosis. These efforts to demystify deep learning models may help improve levels of acceptability to patients and adoption by ophthalmologists.

References

Rupert RA Bourne, Gretchen A Stevens, Richard A White, Jennifer L Smith, Seth R Flaxman, Holly Price, Jost B Jonas, Jill Keeffe, Janet Leasher, Kovin Naidoo, et al.

Causes of vision loss worldwide, 1990–2010: a systematic analysis. *The lancet global health*, 1(6):e339–e349, 2013.

Xiangyu Chen, Yanwu Xu, Damon Wing Kee Wong, Tien Yin Wong, and Jiang Liu. Glaucoma detection based on deep convolutional neural network. In *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 715–718. IEEE, 2015.

Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, Wei Xie, Gail L Rosen, Benjamin J Lengerich, Johnny Israeli, Jack Lanchantin, Stephen Woloszynek, Anne E Carpenter, Avanti Shrikumar, Jinbo Xu, Evan M Cofer, Christopher A Lavender, Srinivas C Turaga, Amr M Alexandari, Zhiyong Lu, David J Harris, Dave DeCaprio, Yanjun Qi, Anshul Kundaje, Yifan Peng, Laura K Wiley, Marwin H S Segler, Simina M Boca, S Joshua Swamidass, Austin Huang, Anthony Gitter, and Casey S Greene. Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society, Interface*, 15(141), April 2018. doi: 10.1098/rsif.2017.0387.

Mark Christopher, Akram Belghith, Christopher Bowd, James A Proudfoot, Michael H Goldbaum, Robert N Weinreb, Christopher A Girkin, Jeffrey M Liebmann, and Linda M Zangwill. Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. *Scientific reports*, 8(1):1–13, 2018.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408.

Vatsal Doshi, Mei Ying-Lai, Stanley P Azen, Rohit Varma, Los Angeles Latino Eye Study Group, et al. Sociodemographic, family history, and lifestyle risk factors for open-angle glaucoma and ocular hypertension: the los angeles latino eye study. *Ophthalmology*, 115 (4):639–647, 2008.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Xuejuan Jiang, Mina Torres, Rohit Varma, Los Angeles Latino Eye Study Group, et al. Variation in intraocular pressure and the risk of developing open-angle glaucoma: the los angeles latino eye study. *American journal of ophthalmology*, 188:51–59, 2018.

Michael A Kass, Dale K Heuer, Eve J Higginbotham, Chris A Johnson, John L Keltner, J Philip Miller, Richard K Parrish, M Roy Wilson, Mae O Gordon, Ocular Hypertension Treatment Study Group, et al. The ocular hypertension treatment study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. *Archives of ophthalmology*, 120(6):701–713, 2002.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Natasha N Kolomeyer, Leslie J Katz, Lisa A Hark, Madison Wahl, Prateek Gajwani, Kanza Aziz, Jonathan S Myers, and David S Friedman. Lessons learned from 2 large community-based glaucoma screening studies. *Journal of Glaucoma*, 30(10):875–877, 2021.

Liu Li, Mai Xu, Hanruo Liu, Yang Li, Xiaofei Wang, Lai Jiang, Zulin Wang, Xiang Fan, and Ningli Wang. A large-scale database and a cnn model for attention-based glaucoma detection. *IEEE transactions on medical imaging*, 39(2):413–424, 2019.

Liu Li, Xiaofei Wang, Mai Xu, Hanruo Liu, and Ximeng Chen. Deepgf: Glaucoma forecast using the sequential fundus images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 626–635. Springer, 2020.

Zhixi Li, Yifan He, Stuart Keel, Wei Meng, Robert T Chang, and Mingguang He. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology*, 125(8):1199–1206, 2018.

Mingquan Lin, Shadab Momin, Yang Lei, Hesheng Wang, Walter J Curran, Tian Liu, and Xiaofeng Yang. Fully automated segmentation of brain tumor from multiparametric mri using 3d context deep supervised u-net. *Medical Physics*, 2021.

Harry A Quigley, Joanne Katz, Robert J Derick, Donna Gilbert, and Alfred Sommer. An evaluation of optic disc and nerve fiber layer examinations in monitoring progression of early glaucoma damage. *Ophthalmology*, 99(1):19–28, 1992.

Alfred Sommer, James M Tielsch, Joanne Katz, Harry A Quigley, John D Gottsch, Jonathan C Javitt, James F Martone, Richard M Royall, Kathe A Witt, and Sandi Ezrine. Racial differences in the cause-specific prevalence of blindness in east baltimore. *New England journal of medicine*, 325(20):1412–1417, 1991.

Andrew J Tatham, Felipe A Medeiros, Linda M Zangwill, and Robert N Weinreb. Strategies to improve early diagnosis in glaucoma. *Progress in brain research*, 221:103–133, 2015.

Anshul Thakur, Michael Goldbaum, and Siamak Yousefi. Predicting glaucoma before onset using deep learning. *Ophthalmology Glaucoma*, 3(4):262–268, 2020.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi S Jaakkola. Rethinking cooperative rationalization: Introspective extraction and complement control. *arXiv preprint arXiv:1910.13294*, 2019.