Proceedings of the Eighth AAAI Conference on Human Computation and Crowdsourcing (HCOMP-20)

How Useful Are the Machine-Generated Interpretations to General Users? A Human Evaluation on Guessing the Incorrectly Predicted Labels

Hua Shen, Ting-Hao (Kenneth) Huang
College of Information Sciences and Technology
The Pennsylvania State University
201 Old Main, University Park, PA 16802, USA
{huashen218, txh710}@psu.edu

Abstract

Explaining to users why automated systems make certain mistakes is important and challenging. Researchers have proposed ways to automatically produce interpretations for deep neural network models. However, it is unclear how *useful* these interpretations are in helping users figure out why they are getting an error. If an interpretation effectively explains to users how the underlying deep neural network model works, people who were presented with the interpretation should be better at predicting the model's outputs than those who were not. This paper presents an investigation on whether or not showing machine-generated visual interpretations helps users understand the **incorrectly predicted labels** produced by image classifiers. We showed the images and the correct labels to 150 online crowd workers and asked them to select the incorrectly predicted labels with or without showing them the machine-generated visual interpretations. The results demonstrated that displaying the visual interpretations did not increase, but rather *decreased*, the average guessing accuracy by roughly 10%.

Introduction

Explaining to users why automated systems make certain mistakes is important. As deep neural network technologies achieve higher performance, they have been applied to important domains, influencing important decisions in healthcare, transportation, and education. However, due to the nonlinear, complicated structures of neural models, the high performance of deep neural networks is achieved at the cost of interpretability. In response, researchers have proposed ways to explain the inner workings of deep neural networks by automatically producing interpretations (Melis and Jaakkola 2018; Selvaraju et al. 2017; Ribeiro, Singh, and Guestrin 2016). Such machine-generated interpretations help various stakeholders (Strobelt et al. 2017): researchers, who develop new deep-learning architectures; machine-learning engineers, who train and optimize existing networks; product engineers, who apply general-purpose pre-trained networks to various tasks; and the general users, who want to understand system outputs (Chu, Roy, and Andreas 2020; Smith-Renner et al. 2020; Selvaraju et al. 2017). This paper focuses on the **end users** – who may not understand the mechanism of the underlying deep neural networks, but are most influenced by their outputs – to investigate whether machine-generated interpretations can help users make sense of errors made by algorithms.

We use the image-classification task as our test bed. Neural image classifiers generate interpretations through two approaches: designing proxies, which are inherently interpretable (*e.g.*, decision tree), to substitute the black-box deep neural networks (Melis and Jaakkola 2018); or generating post-hoc interpretations outside the deep neural network workflow (Selvaraju et al. 2017), which is where our work will focus. Most post-hoc interpretations are in the form of instance-wise interpretation – for example, saliency maps of input images. A saliency map highlights the most informative region of the image with respect to its classification label, unveiling post-hoc evidence of the neural network prediction. This line of work was in part motivated by the need of "end users" (Du et al. 2018; Nourani et al. 2019), "non-expert users" (Ribeiro, Singh, and Guestrin 2016), or "untrained users" (Selvaraju et al. 2017), and the generated interpretations were often evaluated by how much they could boost users' trust of deep neural networks. However, it is still unclear how **useful** these interpretations are in helping users make sense of automated system errors.

The need for interpretability arises due to *Incompleteness* in the problem formalization, making it difficult to make further judgements or optimizations (Doshi-Velez and Kim 2017). When a user observed a few cases where the automated system incorrectly labeled his/her images, it was difficult for the user to decide what to do. Did the errors occur because the system's accuracy level is low? If so, should the user switch to another system? Are the images too complicated for computers, in which case users should not expect reliable image labels? Did the underlying algorithms have biases that worsened with certain types of images? We argue that errors *expose* existing incompleteness in the problem formalization, requiring users to seek interpretations. Namely, an important use case of interpretations is to help users figure out what is going on when they get certain errors. Researchers have proposed evaluations to assess how

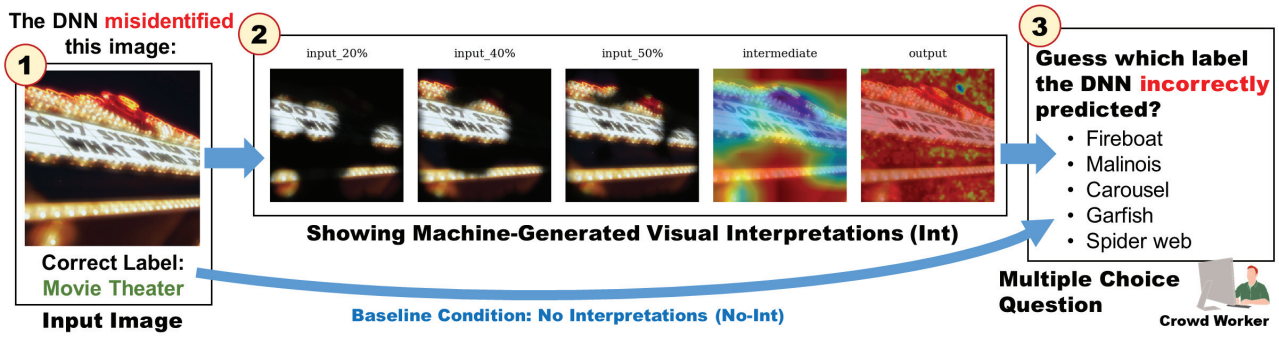

Figure 1: The workflow of "Guessing the Incorrectly Predicted Label" task. Each worker is presented with an image and told that the deep neural network incorrectly predicted its label (Step 1). The worker may also be presented with visual interpretations (Step 2). The worker is then asked to guess the incorrectly predicted label ("Carousel" in this example) from five options, four of them being distractors (Step 3). If an interpretation effectively explains how the underlying deep neural network model works to users, the people who were presented with the interpretation should be better at predicting the model's outputs.

much an interpretation reflects the model's behavior (also known as "fidelity") (Melis and Jaakkola 2018) or boosts users' trust in automated systems (Selvaraju et al. 2017; Ribeiro, Singh, and Guestrin 2016). However, it is unclear how *useful* these interpretations are in helping users figure out why they are getting an error.

This paper introduces a method that uses crowd workers from Amazon Mechanical Turk (MTurk) to directly evaluate the usefulness of interpretations in helping users to reason about the errors of deep neural networks[1]. Figure 1 overviews the workflow. In this task, each worker is presented with an image and told that the deep neural network incorrectly predicted its label. The worker may also be presented with a set of interpretations (*e.g.*, saliency maps) that explain how the deep neural network "perceives" this image and makes the final prediction. The worker is then asked to **guess the incorrectly predicted label** from five options, four of them being distractors. If an interpretation effectively explains how the underlying deep neural network model works to users, the people who were presented with the interpretation should be better at predicting the model's outputs than those who were not.

This paper tried to answer two research questions: First **(RQ1)**, do machine-generated visual interpretations help human users better identify predicted labels? Second **(RQ2)**, when do (and when do not) the visual interpretations help?

Related Work

Interpretation Methods Our work focuses on post-hoc interpretations. These methods generate saliency maps to indicate where the neural networks "look" in the images for their predictions' evidence. Existing methods can be categorized into four lines: *Backprop-Based*: computes the gradient (or variants) of the neural network output to score the importance of each input pixel, such as SmoothGrad (Smilkov et al. 2017); *Representation-Based*: uses the feature maps at intermediate layer of neural networks to generate saliency maps, like GradCAM (Selvaraju et al. 2017); *Meta-Model-Based*: trains a meta-model to predict the saliency map for any given input in a single feed-forward pass, such as RTS (Dabkowski and Gal 2017); *Perturbation-Based*: finds the saliency map by perturbing the input with minimum intervention and observing the change in model prediction, like ExtremalPerturb (Fong, Patrick, and Vedaldi 2019).

Evaluating Interpretations Evaluating the effectiveness of interpretations is critical in practice. Existing evaluations answer two questions: whether the interpretations genuinely reflect neural network behavior (Adebayo et al. 2018), and whether the interpretations are useful for users. To answer the latter question, a set of metrics are proposed to involve human evaluation. For instance, trust assessment and user satisfaction is verified in Smith-Renner et al. (2020) by surveying general users. Mental model evaluations designed by Bucinca et al. (2020) and Chu, Roy, and Andreas (2020) measure whether general users can understand and predict model outputs. Feng and Boyd-Graber (2019) creates a human-computer cooperative task to measure how much interpretation improves human performance. However, more study is needed to investigate how general users perceive and predict neural networks' failure cases, which is of vital importance in building trust and correcting model behavior.

Human-AI Collaboration Although human computation has traditionally played a data annotation role in deep learning systems, there is increasing interest in incorporating it into diverse stages of human-AI hybrid systems (Nourani et al. 2019). Due to its goal of building human understanding and trust in black-box neural networks, interpretation is inherently a human-centric problem. Related efforts involve human perception of different types of interpretation representations in visual interfaces (Roy et al. 2019), etc.

Method

We used a deep neural network to label images and employed several interpreters to generate visual interpretations

[1]The code and interface are available via GitHub: https://github.com/huashen218/GuessWrongLabel

for the images. We showed each image the deep neural network had labeled incorrectly to a group of online crowd workers and asked them to guess which images the deep neural network had mistakenly labelled. Only the workers in the control group were presented with the visual interpretations. We detail the procedure of the study in this section.

Step 1: Labeling Images We trained an image classifier on ImageNet dataset, with its TOP-1 accuracy reaching 78.67% (Xie et al. 2019). We randomly selected images whose labels were incorrectly identified by the classifier.

Step 2: Generating Instance-Wise Interpretations For each image in the misclassified subset, we used three existing interpreters – *i.e.,* input perturbation (Fong, Patrick, and Vedaldi 2019), intermediate feature extraction (Selvaraju et al. 2017), and output backpropagation (Smilkov et al. 2017) – to explain three aspects of this image. **Input perturbation interpretation (column 2-4 in Figure 2)** observes how the output value changes as input is "deleted" in different sub-regions. We used *ExtremalPerturb*, which aims to find a small pixel subset that, when preserved, are sufficient to keep model output stable. Moreover, ExtremalPerturb allows researchers to explicitly constrain the percentage of preserved pixels. We provided three levels of percentage: $a = \{20\%, 40\%, \text{and } 50\%\}$. **Inter-Feature extraction interpretation (column 5 in Figure 2)** looks at intermediate layers of the neural network to indicate the discriminative image regions used by the model for prediction. We used *GradCAM*, which extracts the gradient information flowing into the last convolutional layers, to explain the importance of each pixel. **Output backpropagation interpretation (column 6 in Figure 2)** leverages backpropagation to track information from the model's output back to its input to generate the saliency map. We used *SmoothGrad*, which samples similar images by adding noise to the original image and using the average of the resulting heatmaps to obtain the final interpretation. We eventually generated *(i)* three saliency maps from input perturbation view with 20%, 40% and 50% percentages respectively, *(ii)* one saliency map from intermediate feature extraction view, and *(iii)* one saliency map from the output backpropagation view.

Step 3: Having Crowd Workers Guess the Incorrectly Predicted Label Next, we recruited crowd workers on MTurk to complete tasks[2]. The workers were shown the image and its correct label, and were informed that "a computer algorithm misidentified this image as something else." Only the workers in the control group, as shown in Figure 1, were presented with the visual interpretations. On the interface, we explained that the visual interpretations are "visualizations that try to show how the algorithm *sees* this image,"

[2]Each Human Intelligence Task (HIT) contained one image, and multiple workers were recruited to answer the question. The price of a HIT is $0.05. Four built-in MTurk qualifications are used: Locale (US Only), HIT Approval Rate (≥98%), Number of Approved HITs (≥3000), and the Adult Content Qualification.

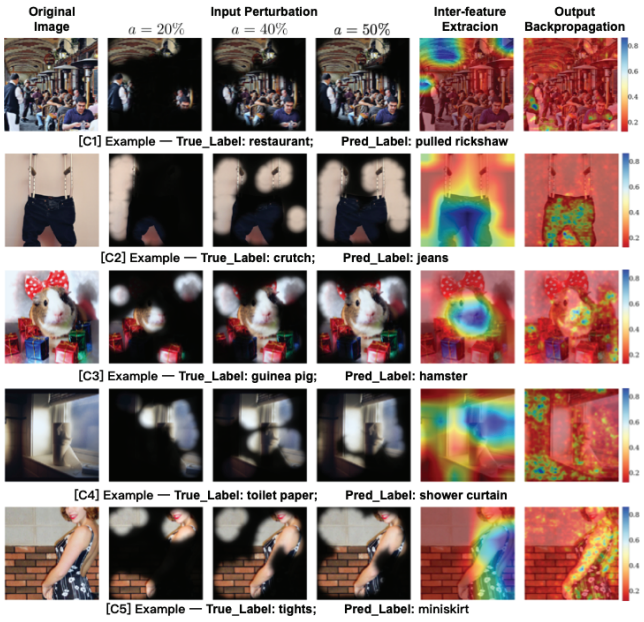

Figure 2: Examples of five types of errors in image classification. The visual interpretations are generated by three existing interpreters (see "Step 2" in the Method section.)

and provided comprehensive descriptions for each interpretation. For example, we explained "input perturbation interpretation" with a 20% mask (column 2 in Figure 2) as "We only allow the algorithm to see 20% of the image and ask the algorithm to choose which 20% is the most important region. The black mask blocks the regions the algorithm pays less attention to." The workers are then asked to **guess the incorrectly predicted label** from five options. One of the options was the incorrect label predicted by the deep neural network model, and the remaining four were randomly selected from the whole label set of ImageNet (*i.e.,* 1,000 labels), excluding the correct gold-standard label.

The assumption is that if the visual interpretations effectively explain how the deep neural network works, the workers who were presented with the interpretations should distinguish the predicted label better than those who were not. Humans alone are sufficient to guess the *correct* label, but it requires workers to take the mechanism of deep neural networks into account to guess the *incorrect* label predicted by deep neural networks. MTurk workers are appropriate participants because they represent general users who do not necessarily understand deep neural network models nor are trained for reasoning about these models' errors.

Categorizing Error Cases Manually To inspect usefulness of interpretation in fine-grained model failure scenarios (RQ2), the authors inspected 1,000 misclassified images and categorized them into five types of errors (Figure 2), in part based on the literature (Arjovsky et al. 2019).

1. **Local Character Inference (C1):** The model arrives at wrong prediction by looking at only part of the object.

For instance, in Figure 2(C1), the error might be due to the model partially capturing the restaurant dome, which looks similar to the canopy of a pulled rickshaw.

2. **Multiple Objects Selection (C2):** For images with multiple objects, the model makes a prediction by choosing another object rather than the labeled one, as in Figure 2(C2).
3. **Similar Appearance Inference (C3):** The model misclassifies the object in the image into another class with a similar appearance, as shown in Figure 2(C3).
4. **Correlation Learning (C4):** The model exploits correlational relationships in training data to apply an incorrect label to the image. For example, in Figure 2(C4), the model predicts a "shower curtain" by identifying the bathroom context, even if no curtain is in the image.
5. **Incorrect Gold-Standard Labels (C5):** The true label of the images might be incorrect in the ImageNet. Figure 2(C5) shows an example.

Experimental Results

Experiment 1: Testing Two Conditions in the Same Batch of HITs Experiment 1 had two conditions: [Interpretation] (*i.e.*, [Int]) and [No-Interpretation] (*i.e.*, [No-Int]). The only difference is that HITs in the [No-Int] group do not show the interpretations to workers in interfaces. We evenly divided 200 randomly selected image samples into two groups. We posted these 200 images in a same batch of HITs at the same time on MTurk, where each HIT recruits nine different workers. A total of 1,800 submissions (900 submissions in each condition) were contributed by 41 workers in [Int] and 40 workers in [No-Int] conditions respectively. We did not control the workers' participation, so a worker could participate in both groups. Thirty-six out of 45 workers participated in both conditions.

Surprisingly, in Experiment 1, **showing the workers machine-generated visual interpretations *reduced* their average accuracy in guessing the incorrectly predicted labels.** We calculated the accuracy as the percentage of correctly inferring the classifier's prediction among all 900 submissions in each condition. The accuracy collected in [Int] was 0.73, while the accuracy in [No-Int] was 0.81. The difference was statistically significant (unpaired t-test, $p<0.05$, N=100). Based on the results, the machine-generated interpretation did not help, but instead hurt, the workers' ability to guess the incorrectly predicted labels. The by-category analysis (Table 1) shows that displaying interpretations significantly lowers human accuracy in cases where the errors were probably caused by similar appearances between items (C3) or by mistakenly learning from the background or scenes of the image (C4).

Experiment 2: Testing with Two None Overlapping Sets of Workers Experiment 2 was controlled more strictly. We randomly selected another 200 images (different from those used in Experiment 1), and used the same photo in both [Int] and [No-Int] conditions. We used custom MTurk qualifications to control the participants: workers who participated in one condition could not accept HITs in the other condition. We recruited 10 different workers for each image, in which five workers were in the [Int] group and the other five were in the [No-Int] group. A total of 2,000 submissions (with 1,000 submissions in each condition) were collected, contributed by 42 workers in the [Int] condition and 63 workers in the [No-Int] condition respectively.

	C1	C2	C3	C4	C5	Overall
Int	0.77	0.83	0.71	0.54	0.71	0.73
#images	29	23	28	15	5	100
No-Int	0.76	0.77	****0.87**	****0.75**	0.78	***0.81**
#images	25	10	47	12	6	100

Table 1: Results of Experiment 1. Showing the workers machine-generated visual interpretations *reduced* their average accuracy in guessing the incorrectly predicted labels. (Unpaired t-test. *: $p<0.05$, **: $p<0.01$.)

	C1	C2	C3	C4	C5	Overall
Int	0.57	0.74	0.66	0.41	0.67	0.63
No-Int	0.52	0.71	****0.84**	***0.59**	0.77	****0.73**
#images	44	20	112	18	6	200

Table 2: Results of Experiment 2. The machine-generated visual interpretation again *reduced* the average human accuracy in inferring model misclassification. (Paired t-test. *: $p<0.05$, **: $p<0.01$.)

In Experiment 2, the machine-generated visual interpretation again **reduced the average human accuracy in inferring model misclassification** (Table 2.) The accuracy of [Int] was 0.63, whereas accuracy in [No-Int] condition was 0.73. The difference was again statistically significant (paired t-test, $p<0.01$, N=200). On average, humans do not benefit from interpretations when inferring incorrect predictions in image classification tasks. Similarly to Experiment 1, the by-category analysis showed that displaying interpretations significantly lowers human accuracy in C3 and C4 (Table 2) errors. We also noticed that the accuracy for C1 and C2 images increased in both experiments when showing visual interpretations, although the differences were not statistically significant.

Discussion

Our experiments showed that, in the case of image classification, machine-generated visual interpretations are not necessarily useful in helping users understand deep neural network failures. It could even be harmful, as in the cases where the errors were probably caused by similar appearances between items (C3) or by mistakenly learning from the background or scenes of the images (C4). System designers should use caution when displaying machine-generated interpretations to users.

Why It Did Not Help More research is required to discover why showing interpretations was ineffective. Here, we submit several of our hypotheses with the goal of helping future explorations. First, the interpreters are not good enough to help humans. The representational power – including the correctness, sensitivity, etc., of the interpretation model – might not be sufficient to augment human reasoning about errors. Although machine-generated interpretations captured some of the deep neural network's behaviors, it may not be good enough to help humans. Second, the format is insufficient. The saliency maps may not be the most efficient format to convey information to humans. For example, when a saliency maps model changes an inner parameter, this change might not be obvious enough to be noticeable by humans, but could still affect the final predictions. Third, the interpreters may work poorly in cases where the image classifier failed.

Limitations We are aware that this work has several limitations. First, the sample size was relatively small. Given that classifiers incorrectly labelled more than 10,000 images in the ImageNet validation set alone, 200 images are relatively small portion of the data. Second, we only tested three particular types of interpretations, and also presented the interpretations together on the same page. This experimental setup introduces the possibility of missing out on the "best" interpretations, or different interpretations might affect each other and reduce their effectiveness. Third, we recruited MTurk workers with certain qualifications to simulate general users. It is difficult to eliminate data noise stemmed from workers' misunderstanding or incognizance of images or options. Finally, we only tested visual interpretations for image classifiers. It requires more research to study if similar effects could be generalized to other tasks.

Conclusion

The goal of this study was to evaluate the usefulness of machine-generated visual interpretations for general users' reasoning about model errors. To this end, we utilized the "guess incorrectly predicted labels" task to examine the usefulness of visual interpretations. Our two sets of control experiments, with 3,800 submissions contributed by 150 online crowd workers, suggest that showing the interpretations does not increase, but rather *decreases*, the average accuracy of human guesses by roughly 10%.

Acknowledgements

We thank Ting Wang for his support. We also thank the workers on MTurk who participated in our studies.

References

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity checks for saliency maps. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 9505–9515.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Bucinca, Z.; Lin, P.; Gajos, K. Z.; and Glassman, E. L. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI)*, 454–464.

Chu, E.; Roy, D.; and Andreas, J. 2020. Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248*.

Dabkowski, P., and Gal, Y. 2017. Real Time Image Saliency for Black Box Classifiers. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Doshi-Velez, F., and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Du, M.; Liu, N.; Song, Q.; and Hu, X. 2018. Towards explanation of dnn-based prediction with guided feature inversion. In *Proceedings of ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 1358–1367.

Feng, S., and Boyd-Graber, J. 2019. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI)*, 229–239.

Fong, R.; Patrick, M.; and Vedaldi, A. 2019. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2950–2958.

Melis, D. A., and Jaakkola, T. 2018. Towards robust interpretability with self-explaining neural networks. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Nourani, M.; Kabir, S.; Mohseni, S.; and Ragan, E. D. 2019. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, 97–105.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of ACM International Conference on Knowledge Discovery and Data Mining (KDD)*.

Roy, C.; Shanbhag, M.; Nourani, M.; Rahman, T.; Kabir, S.; Gogate, V.; Ruozzi, N.; and Ragan, E. D. 2019. Explainable activity recognition in videos. In *IUI Workshops*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 618–626.

Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; and Wattenberg, M. 2017. SmoothGrad: Removing Noise by Adding Noise. In *International Conference on Machine Learning Workshop on Visualization for Deep Learning*.

Smith-Renner, A.; Fan, R.; Birchfield, M.; Wu, T.; Boyd-Graber, J.; Weld, D. S.; and Findlater, L. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.

Strobelt, H.; Gehrmann, S.; Pfister, H.; and Rush, A. M. 2017. Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE transactions on visualization and computer graphics* 24(1):667–676.

Xie, Q.; Dai, Z.; Hovy, E.; Luong, M.-T.; and Le, Q. V. 2019. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.