Counterexample Contrastive Learning for Eliminating Spurious Correlations

Author Name
Affiliation
pcchair@ijcai-22.org

Abstract

Bias datasets cause models to learn spurious correlations. Compared with constructing new counterfactual samples, we consider that making full use of the samples in the dataset can also eliminate spurious correlation. Through analytical experiments, we find that the counterexamples in the dataset can play an important role in avoiding model utilizing spurious correlation. Inspired by the above conclusion, we propose counterexample contrastive (CeCon) loss which treats counterexamples as negatives in contrastive loss. This method utilizes contrastive learning to pull the samples with different bias feature in the same class and push the samples with the same bias feature in different class, so as to eliminate the spurious correlation caused by bias. Experimental results show that our proposed method can achieve state-of-the-art results when the bias features are known.

1 Introduction

Machine learning models have achieved a significant performance in the fields of computer vision and natural language processing [He *et al.*, 2016; Devlin *et al.*, 2018]. However, many recent works have shown that biases existing in the dataset will lead models to learn bias features highly correlated to labels [Agrawal *et al.*, 2018]. This spurious correlation will make models fail to make correct prediction. Even models make correct prediction, they do not depend on the correct basis.

As shown in Fig.1, in the hair color classification task, if women in the image dataset are mostly blond and men are mostly non-blond, the model will take gender as a key basis for predicting hair color. In this situation, the model learns the spurious correlation between gender and hair color, and gender is referred as bias feature. This phenomenon is more obvious when the distribution of test dataset differed from that of training dataset, because spurious correlation no longer help model make correct prediction.

To address this problem, a common and effective way is using data augmentation to generate counterfactual samples [Chen *et al.*, 2020a; Liang *et al.*, 2020]. The counterfactual samples in the dataset can balance the correlation between bias features and labels, that is, to avoid a certain bias feature highly correlated with only one or a few labels. Therefore, the model will rely less on spurious correlation and more on invariant features. However, in order to meet the requirements of researchers for counterfactual sample style, generating counterfactual samples through data augmentation also needs other mature technologies (object detection [Redmon and Farhadi, 2018], semantic segmentation [Chen *et al.*, 2017], etc.) or a large number of manual annotation [Plumb *et al.*, 2021], and the authenticity of the augmented image is poor.

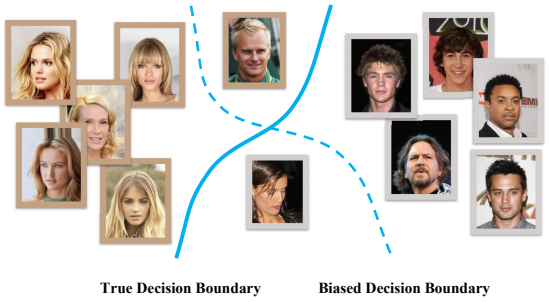


Figure 1: A subset of a bias dataset. The images in the brown box belong to the blond class and the images in the grey box belong to the non-blond class.

Inspired by the spurious correlation detection work [Dancette *et al.*, 2021; Wang and Culotta, 2020], we observe that compared with a large number of counterfactual samples generated based on human priors, there already exist samples similar to counterfactual samples in the original dataset named **counterexamples**. For example, in Fig.1, the images of "non-blond female" and other "blond female" have the same feature "female" but different classes, and images of "non-blond female" are referred to as counterexamples to other images of "blond female". The relationship between a sample and its counterexample and the relationship between a sample and its counterfactual sample are very similar, that is two samples with similar features but their classes are different. Therefore, is it possible to avoid models utilizing spurious correlations by making full use of the existing counterex-

amples in the dataset without generating new counterfactual samples?

To verify the above assumption and inspire our solutions, We conduct analytical experiments and draw two conclusions: (1) counterexamples play a key role in avoiding model utilizing spurious correlation; (2) however, based on the current regular training method, counterexamples play a limited role in eliminating spurious correlation. The reason is that the representation of counterexamples in the feature space will be affected by bias features and be closer to the class center of the wrong class. Therefore, we propose to use Counterexample Contrastive (CeCon) loss to solve the above problem. Contrastive learning can pull samples with different bias features in the same class and push samples(i.e. counterexamples) with the same bias feature in different classes simultaneously.

Our main contributions are summarized as follows:

- Through experiments, we analyze the role of counterexamples in avoiding model utilizing spurious correlation, and discuss the reasons why counterexamples can not play a full role even if there already exist counterexamples in the dataset.
- We propose Counterexample Contrastive (CeCon) loss, which does not generate new counterfactual samples, but uses counterexamples existing in the dataset to avoid model utilizing spurious correlations. Moreover, experiments show that our method can achieve state-of-the-art results when the bias features are known.

To verify the role of counterexamples in avoiding model learning bias and inspire our solutions, we answer two questions in this section: (1) does the existence of counterexamples play a role in avoiding model utilizing spurious correlation? and (2) why does the model still utilize spurious correlation when there already exist counterexamples in the dataset?

Definition of Counterexamples. We define counterexample referring to the description of that in VQA [Dancette *et al.*, 2021]. Considering the input image and label (x, y) in dataset D as random variable sampling from $P(X, Y)$ and decoupling the input x as feature (x^t, x^b) where x^t is the invariant feature correlated with y , and x^b is the bias feature without causal correlation with y. We define the counterexample as follows: for any two sample x_i and x_j , x_i is the "counterexample" of x_j when $y_i \neq y_j$ and $x_i^b \approx x_j^b$, vice versa.

For convenience, in the later section we use x_{ce} denoting the "counterexample" of any example x_s following the above definition. We find that x_s can also be called as a counterexample of x_{ce}. In addition, we find the features of samples belong to same feature (e.g. background, color) are closer than others, so we can view the samples belonging to the same labeled features but different labels as counterexamples, especially examples belonging to same bias class but being distributed unbalancedly across classes. In real world, the unbalanced distribution of some features easily and often results in bias, and the category of attribute approximate category of bias, so we can't differentiate the both in the following sections. Based on unbalanced feature (bias) we can find the counterexamples in dataset. As shown in Fig.1, woman pictures of class non-blond is the counterexample relative to woman pictures of class blond. Further, the counterexample in the following specifically refers to the samples that account for less in the dataset.

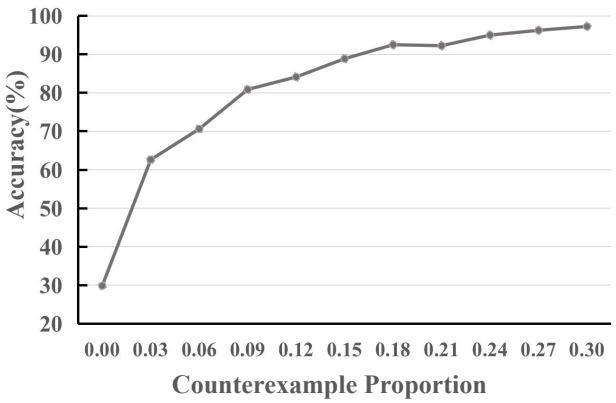


Figure 2: The changing trend of Accuracy with different counterexample proportion.

2 Analysis

Important Effect of Counterexamples. To explore the role of counterexamples in avoiding model utilizing spurious correlation, we first discuss how counterexamples play a role in human judgment. As shown in Fig.3, for a group of binary classification data, if people only know the class labels of the two types of data, but not the specific meaning of the labels, some people will take hair color as the classification basis, while others will take gender as the classification basis. However, if the two types of data contain one counterexample image respectively, humans are easy to determine the classification target and focus on the invariant features that are truly correlated to the class. Furthermore, no matter whether the distribution of the test set is consistent with that of the training set, human judgment will no longer rely on spurious correlation.

Further, through similar experimental settings, we analyze the role of counterexamples in avoiding model utilizing spurious correlation. Taking the hair color classification as the classification task, in the training set, we only use the image of "blond, female" as class 1 and the image of "non blond, male" as class 2, while in the test set, the image of "blond, male" as class 1 and the image of "non blond, female" as class 2. Further, we ensure that the total number of training samples remains unchanged, and gradually increase the proportion $\rho \in [0, 0.30]$ of counterexample images in the training set, increasing the "blond, male" image of class 1 and the "non blond, female" image of class 2, reducing other images. Finally, as shown in Fig.2, as the number of counterexamples in the training set varies from zero to few, the accuracy of the test set has been significantly improved. This proves that the existence of counterexamples plays a decisive role in avoiding model utilizing spurious correlation. However, with the

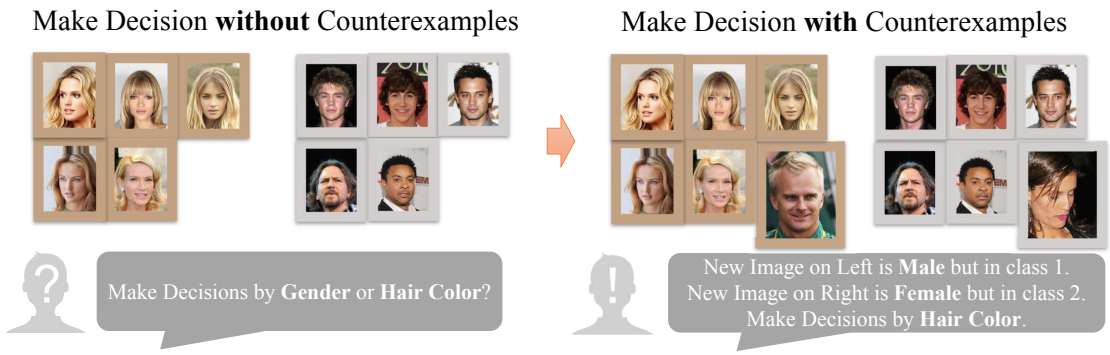
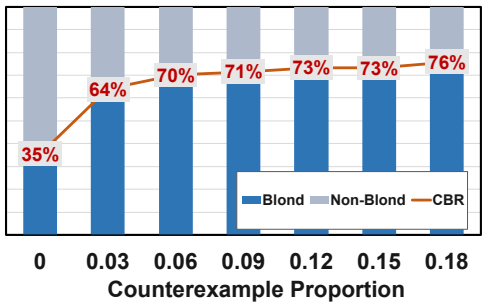
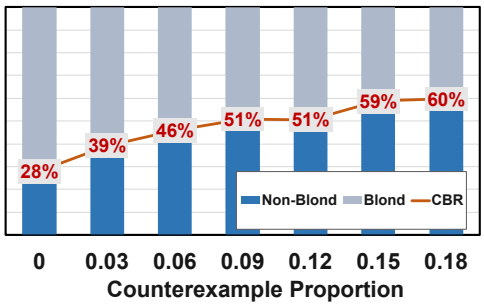


Figure 3: Illustration that humans adjust their decisions with small amount of counterexamples



(a) The changing trend of CBR about counterexample "non blond female"



(b) The changing trend of CBR about counterexample "blond male"

Figure 4: The increase of CBR value with the increase of the proportion of counterexamples in the training set.

increase of the number of counterexamples, although the accuracy of the test set increases gradually, the growth rate will slow down. This phenomenon reflects the positive correlation between the role of counterexamples and the number of counterexamples, but the model can not avoid spurious correlation by using only a small number of counterexamples like human. In conclusion, counterexamples play a key role in avoiding model utilizing spurious correlation. However, based on the current regular training method, there are limits to the role of counterexamples in eliminating spurious correlation.

Representation of Counterexamples in Feature Space. Through the above analysis, it can be found that whether for human or model, the existence of counterexamples plays an important role in avoiding utilizing spurious correlation. However, why does the model still utilize spurious correlation when counterexamples already exist in the dataset? To answer this question, we analyze the representation of samples in feature space. We calculate the distance between the counterexample (such as blond male) and the class center of its class (blond) and the distance between the counterexample and the class center of its non-class (non-blond), and define CBR(Counterexample Belongness Ratio):

$$CBR = \frac{Avg_{ce} dis(e_{x_{ce}}, c_{y_s})}{Avg_{ce} dis(e_{x_{ce}}, c_{y_s}) + Avg_{ce} dis(e_{x_{ce}}, c_{y_{ce}})} \quad (1)$$

In the formulation, $dis(\cdot, \cdot)$ is the calculation function of Euclidean distances. $e_{x_{ce}}$ is representation of one counterexample, $c_{y_{ce}}$ and c_{y_s} represent respectively centers of class, which counterexamples belong to and not belong to , and $Avg_{ce}(\cdot)$ is calculated average value to all counterexamples. The index reflects the distance relationship between the counterexample

and the centers of the two classes. The larger the value of CBR, the closer the counterexample is to the class center of the correct class and the farther to the wrong class, vice versa. As shown in Fig.4(a), the horizontal axis is the proportion ρ of the counterexample in the training set, and the vertical axis is CBR. The length of the light blue column represents the distance between the counterexample and the center of the correct class, and the length of the deep blue column represents the distance between the counterexample and the center of the wrong class. Taking the counterexample of blond female as an example, we can see that when the counterexample proportion in the training concentration is relatively low, the counterexample (non blond female) is closer to the class center that does not belong to the class (blond) than to the class center of the class (non blond). This reflects the defect of the representation of Counterexamples in feature space.

However, it can also be seen from Fig.4(a) that with the increase of the proportion of counterexamples in the training set, the counterexamples are farther and farther away from the class center of the wrong class, and the value of CBR is larger. This shows that with the increase of the proportion of counterexamples in the training set, the representation of counterexamples in the feature space is more reasonable. Due to the reasonable representation, the accuracy is also improved. With the increase of the proportion of counterexamples and the accuracy, the index also increases.

Therefore, this phenomenon inspires us that if the number of counterexamples in a bias training data set cannot be changed, we can try to adjust and optimize the embedded representation of counterexamples in the feature space directly to make it closer to the class center of the correct class and farther from the class center of the wrong class. In this way, without changing the number of counterexamples in the training set, we make the model achieve the effect similar to increasing the proportion of counterexamples in the training set in the above experiment, that is, to avoid the model using spurious correlation

The index reflects the distance relationship between the counterexample and the two centers of classes. The larger the value of the index, the closer the counterexample is to the class center of the correct class and the farther it is to the class center of the wrong class, indicating that the counterexample is less affected by the bias features; vice versa. As shown in Fig.4(b), with the increase of the proportion of counterexamples and the accuracy, the index also increases. Therefore, this phenomenon inspires us that if the number of counterexamples in a dataset cannot be changed, we can directly optimize the embedded representation of counterexamples to make them closer to the class center of the correct class and farther from the class center of the wrong class.

3 Method

Problem Definition. We divide the features of samples into invariant features and irrelevant features. Invariant features are those that can truly be used as the classification basis of classification tasks, while irrelevant features are meaningless to classification tasks and should not be used as the classification basis. The conditional probability of irrelevant features and classes is balanced in unbiased datasets but imbalanced in bias datasets. We take such irrelevant features as bias features. In such a bias dataset, when deploying the conventional training based on cross-entropy loss, and the model will utilize the bias features to predict the class. This is exactly the problem that the model will take shortcuts and rely on spurious correlation to make predictions. To verify the assumption of avoiding model learning bias features and spurious correlation by making full use of counterexamples in the dataset, we suppose that (1) the bias features in the dataset are known and (2) only use the existing samples in the dataset instead of the new samples that can be augmented by counterfactual data.

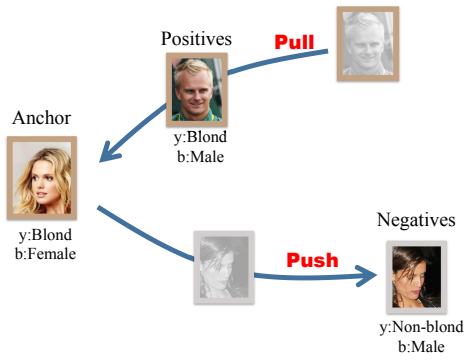


Figure 5: Illustration of Counterexample Contrastive Loss. In feature space, CeCon loss pull the samples with different bias feature in the same class and push the samples with the same bias feature in different class.

Therefore, the goal of this method is that the model can avoid using the bias features as the judgment basis of classification task in a biased dataset, under the assumption that the bias feature is known.

3.1 Self-Supervised Contrastive Loss

In traditional method of contrastive learning, for a batch of N randomly sampled pairs $I = \{x_k, y_k\}_{k=1}^{N}$,we can get the *multi-viewed* batch $\{\tilde{x_k}, \tilde{y_k}\}_{k=1}^{2N}$ by two random data augmentations, let $\tilde{I} = \{1, \cdots, 2N\}$ is index of the *multi-viewed* batch. In self-supervised contrastive learning [Chen *et al.*, 2020b], defined contrastive loss as follows:

$$\mathcal{L}_{self} = -\frac{1}{2N}\sum_{i\in\tilde{I}}\frac{exp(z_i \cdot z_{j(i)}/\tau)}{\sum_{\alpha\in\tilde{I}\setminus\{i\}} exp(z_i \cdot z_\alpha/\tau)} \quad (2)$$

where z_i is the L_2 normalized feature of input x_i from feature-extractor, $z_i \cdot z_j$ denotes inner dot, τ is the temperature parameter and $j(i)$ is the index of the other augmented sample originating from the same source sample. Generally, the sample x_i is named as anchor, the sample $x_{j(i)}$ is named as positive and all others are named as negatives. The above Self-supervised Contrastive Loss $\mathcal{L}_{self}$ can align the align anchor and positive, and push away the anchor and negatives.

3.2 Counterexample Contrastive Loss

The experiments of section 3 show that since the influence of bias feature, the features of counterexample are much

Table 1: The unbiased/bias-conflict accuracy and standard error of the model trained on the CelebA [Liu *et al.*, 2015] dataset.

Task	Acc. Type	Vanilla	LNL	DI	EnD	BiasCon	CeCon
Blond	Unbiased	79.0±0.1	80.1±0.8	90.9±0.3	86.9±1.0	90.0±0.2	**90.2±0.6**
	Bias-Conflict	59.0±0.1	61.2±1.5	86.3±0.4	76.4±1.9	85.1±0.4	**85.5±1.0**
Makeup	Unbiased	76.0±0.8	76.4±2.3	74.3±1.1	74.8±1.8	**77.5±0.7**	74.4±3.9
	Bias-Conflict	55.2±1.9	57.2±4.6	53.8±1.6	53.3±3.6	**61.3±1.6**	55.0±8.0

Table 2: The unbiased/bias-conflict accuracy of the model trained on the UTKFace [Zhang *et al.*, 2017] dataset.

Bias	Acc. Type	Vanilla	LNL	DI	EnD	BiasCon	CeCon
Race	Unbiased	87.4±0.3	87.3±0.3	88.9±1.2	88.4±0.3	90.3±0.2	**91.6±0.3**
	Bias-confflict	79.1±0.3	78.8±0.6	89.1±1.6	81.6±0.3	88.8±0.5	**91.0±0.6**
Age	Unbiased	72.3±0.3	72.9±0.1	75.6±0.8	73.2±0.3	75.7±0.2	**78.4±0.9**
	Bias-confflict	46.5±0.2	47.0±0.1	60.0±0.2	47.9±0.6	61.7±0.5	**81.3±2.5**

closer to the center of the wrong class than to the center of own class. There is a consensus that the model don't learned a well representation resulting in biased prediction. To solve this problem and inspired from self-supervised contrastive learning, we modified the $\mathcal{L}_{self}$ as Counterexample Contrastive Loss $\mathcal{L}_{CeCon}$ to help push together samples of same class and different bias features, and push away samples of different classes and same bias features, see Fig.5. The formulaic definition is as follows: first, we define b_i as the bias feature of the i sample, such as gender in the hair color classification task above; and further, we define $J(i) = \{j \in \tilde{I} | y_j = y_i, b_j \neq b_i\}$ is the index set of samples which have different bias feature in same class with sample x_i and $K(i) = \{k \in \tilde{I} | y_k \neq y_i, b_k = b_i\}$ is the index of samples which have same bias feature in different class with sample x_i, further define the counterexample contrastive (CeCon) loss applied in *multi-viewed* batch training:

$$\mathcal{L}_{CeCon} = -\frac{1}{2N} \sum_{i \in \tilde{I}} \frac{1}{|J(i)|} \sum_{j \in J(i)} log \frac{exp(z_i \cdot z_j / \tau)}{\sum_{\alpha \in J(i) \cup K(i)} exp(z_i \cdot z_\alpha / \tau)} \tag{3}$$

We combine our $\mathcal{L}_{CeCon}$ with standard cross-entropy loss $\mathcal{L}_{CE} = -\frac{1}{N} \sum_{k \in I} \log p(y_k | x_k)$:

$$\mathcal{L} = \alpha \mathcal{L}_{CE} + \mathcal{L}_{CeCon} \tag{4}$$

where α is hyper-parameter as weight the cross-entropy loss , and especially we don't use the samples to calculate cross-entropy after data augmentation. In addition, since there are very few counterexamples in training data, random sample a batch I may included few counterexamples result in there are few positives and negatives, even get a bad performance. So we use an over-sample method [Hong and Yang, 2021] suggested by to make sure adequate counterexamples in a batch.

4 Experiments

4.1 Dataset

In order to verify the effectiveness of our method, we conducted experiments on two bias datasets: CelebA [Liu *et al.*, 2015] and UTKFace [Zhang *et al.*, 2017], under the setting of [Hong and Yang, 2021]. For CelebA, we set up two binary classification tasks about whether the picture is HeavyMakeup and whether it is BlondHair. Both tasks have the bias feature of gender (male or female). In the dataset, compared with men, women have a larger number of pictures of HeavyMakeup and BlondHair. For UTKFace, gender will be the classification target of the binary classification task, while race and age will be the bias features. Compared with men, there exist more pictures of older women and non-white women in the dataset.

4.2 Metric and Baselines

We used two important evaluation indicators: unbiased accuracy [Bahng *et al.*, 2020; Nam *et al.*, 2020] and bias conflict accuracy [Nam *et al.*, 2020]. Unbiased accuracy avoids that the test data is bias data by calculating the accuracy mean of different bias features and label combinations. Bias conflicting accuracy only counts the bias conflicting samples in the test set (for example, blond male and non blond female). Therefore, bias conflict accuracy can better reflect whether the model avoids using bias features as classification basis.

We compare our method with "Vanilla" and other baselines that also utilize the bias feature when training. "Vanilla" is the method without using any debiasing technique. Other baselines include LNL [Kim *et al.*, 2019],DI [Wang *et al.*, 2020],EnD [Tartaglione *et al.*, 2021] and BiasCon [Hong and Yang, 2021]. Among them, BiasCon also uses contrastive learning to debias. BiasCon only pull the samples with different bias features in the same class. Besides that, our method also pushes the samples with same bias feature in the different classes.

4.3 Result

Results on CelebA and UTKFace are shown in Table 1 and 2. Experimental results show that our method CeCon outperforms previous methods in most cases, proving the effectiveness of our method. The results of our method on the Makeup task are not ideal. Through observation, it is found that the samples "Male with Makeup" mainly affect the accuracy. Fig.6 shows the samples "Male with Makeup" in the test

Table 3: CBR of different counterexample with different trained models.b/m stand for blond male, non b/f stand for non blond female, f/w stand for white female, m/non w stand for non white male.

	CelebA		UTKFace	
	b/m	non b/f	f/w	m/non w
vanilla	0.494	0.790	0.582	**0.729**
BiasCon	0.574	**0.836**	0.695	0.663
CE-CL	**0.591**	0.814	**0.706**	0.678

Table 4: Results of different forms of utilizing counterexamples.

Bias	Acc. Type	BiasCon	W-BiasCon	CeCon
Race	Unbiased	90.3±0.2	90.8±0.3	**91.6**±0.3
	Bias Conflict	88.8±0.5	89.2±0.7	**91.0**±0.6
Age	Unbiased	75.7±0.2	77.8±4.6	**78.4**±0.9
	Bias Conflict	61.7±0.5	71.1±5.1	**81.3**±2.5

set. From a human perspective, most of these images are classified as the class of "without Makeup". Moreover, the number of these samples is very small (only 9), so the results are random. This may be the reason for our poor performance.

4.4 Feature Space Analysis

To see whether our method has the effect of pulling samples of the same class with different bias features closer and pushing away samples of different classes with the same bias feature, we calculated the CBR metric on three tasks where our method performed well (see Section 2).

As shown in Table 4, it can be found that our method does pull close the distance between the counterexamples and the class centers of the correct class in most cases. This shows that our method makes the representation of counterexamples in the feature space more reasonable and less susceptible to bias features.

4.5 Different Forms of Utilizing Counterexamples

To demonstrate that counterexamples can play an important role in avoiding models utilizing spurious correlations, we also consider other simpler forms to take advantage of counterexamples. For example, based on the method [Hong and Yang, 2021], we directly set a weight(greater than 1) for the counterexample-related term in the BiasCon loss. We call this method W-BiasCon.

The experimental results are shown in Tabel 3. W-BiasCon is equally effective, showing the power of counterexamples to avoid models utilizing spurious correlations. However, the results of W-BiasCon are not as good as those of CeCon. This shows that CeCon is more effective in the form of only using counterexamples as negatives, and this idea is similar to the idea of selecting difficult negatives in contrastive learning [Robinson *et al.*, 2020].

5 Conclusion

In this work, we propose counterexample contrastive (CeCon) loss, which can avoid the model learning and utilizing spurious correlation by pulling close the samples with different bias features of the same class and pushing far the samples with the same bias features of different classes. Experiments on CelebA and UTKFace show that our method can achieve state-of-the-art results when the bias features are known.

Figure 6: The testset of task "Makeup".

In addition, our method also has some limitations. That is, the bias features need to be known and labeled. In the future, we will try to obtain bias features that possibly be utilized by models through clustering, spurious correlation detection and other methods. Further we determine the counterexamples in the dataset without relying on the previous features labels, and realize the end-to-end counterexample contrastive learning method.

Besides, the methods of counterfactual data augmentation do not conflict with the method of using counterexamples in dataset proposed in this paper. In the future, we will try to fully consider the advantages of both the methods and put forward a novel method of removing spurious correlation by effectively combining the two methods.

References

[Agrawal *et al.*, 2018] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980, 2018.

[Bahng *et al.*, 2020] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning debiased representations with biased representations. In *International Conference on Machine Learning*, pages 528–539. PMLR, 2020.

[Chen *et al.*, 2017] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[Chen *et al.*, 2020a] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809, 2020.

[Chen *et al.*, 2020b] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework

for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[Dancette *et al.*, 2021] Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. *arXiv preprint arXiv:2104.03149*, 2021.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

[Hong and Yang, 2021] Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[Kim *et al.*, 2019] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019.

[Liang *et al.*, 2020] Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3285–3292, 2020.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[Nam *et al.*, 2020] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. *arXiv preprint arXiv:2007.02561*, 2020.

[Plumb *et al.*, 2021] Gregory Plumb, Marco Tulio Ribeiro, and Ameet Talwalkar. Finding and fixing spurious patterns with explanations. *arXiv preprint arXiv:2106.02112*, 2021.

[Redmon and Farhadi, 2018] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[Robinson *et al.*, 2020] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

[Tartaglione *et al.*, 2021] Enzo Tartaglione, Carlo Alberto Barbano, and Marco Grangetto. End: Entangling and disentangling deep representations for bias correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13508–13517, 2021.

[Wang and Culotta, 2020] Zhao Wang and Aron Culotta. Identifying spurious correlations for robust text classification. *arXiv preprint arXiv:2010.02458*, 2020.

[Wang *et al.*, 2020] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020.

[Zhang *et al.*, 2017] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818, 2017.