Aprimorando a Técnica Small Loss Approach para Lidar com Amostras Ruidosas em Modelos Deep Learning

1st Vitor Bento
Departamento de Engenharia Elétrica
Pontifícia Universidade Católica do Rio de Janeiro
Rio de Janeiro, Brasil
vitorbsousa@ica.ele.puc-rio.br

2nd Manoela Kohler
Departamento de Engenharia Elétrica
Pontifícia Universidade Católica do Rio de Janeiro
Rio de Janeiro, Brasil
manoela@ele.puc-rio.br

3rd Marco Pacheco
Departamento de Engenharia Elétrica
Pontifícia Universidade Católica do Rio de Janeiro
Rio de Janeiro, Brasil
marco@ele.puc-rio.br

***Abstract*—A área de estudo *noisy samples*, do qual se refere ao treinamento de modelos de *Deep Learning* com amostras com rótulos equivocados tem grande importância em aplicações reais de *Deep Learning*. Atualmente a técnica *Small Loss Approach* vem sendo amplamente utilizada nos modelos do estado da arte da área. Essa técnica consiste em selecionar as amostras com rótulos corretos do conjunto de dados de treino, excluindo as amostras com rótulos errados. Nesse trabalho demostramos como aumentar o desempenho dessa técnica utilizando o *K-Nearest Neighbors Classifier* para aumentar o número de amostras com rótulos corretos selecionados. Foi utilizado o *dataset* CIFAR-10 com ruído inserido artificialmente para validar o processo, foi obtido ganhos de performance sobre a acurácia de teste e sobre o *Label Precision* de 2% ao comparar o modelo *Co-teaching* com a *Small Loss Approach* com os aprimoramentos apresentados nesse trabalho, e o modelo *Co-teaching* com a *Small Loss Approach* convencional. Os próximos passos do estudo irá consistir na ampliação dos datasets benchmarks utilizados para validar o processo e na aplicação dos modelos do estado da arte com a *Small Loss Approach* aprimorada numa demanda real de importância ambiental de classificação de algas calcárias.**

***Index Terms*—Noisy Samples, Amostras Ruidosas , Small Loss Approach, Deep Learning**

I. Introdução

Modelos de Deep Learning para classificação de imagens [1], [4], [6]–[8], [11], [14], [15], [18] alcançaram o estado da arte em um vasto campo de aplicações. Atualmente, um dos desafios da área, oriundos de aplicações reais, é o aprendizado destes modelos com amostras ruidosas [5]. O termo "amostras ruidosas" se refere a amostras com rótulos errados, *labels* equivocados, presentes no dataset, modelos de Deep Learning treinados nesse cenário possuem baixo desempenho, o que é altamente indesejado. Algumas fontes comuns de ruído em datasets são *web queries* [10], *crowdsourcing* [13], anotação feita por não especialistas ou até mesmo especialistas em tarefas de anotações muito desafiadoras.

Os métodos atuais encontrados na literatura para lidar com esse problema se concentram na abordagem de seleção de amostras limpas, ou seja, amostras com os rótulos corretos presentes no dataset, excluindo do treino as amostras ruidosas, e.g. [16] e [5]. Os modelos do estado arte (SOTA, do inglês State Of The Art), Co-teaching [5], Co-teaching+ [17] e Jocor [12] são desta última categoria, e para selecionar as amostras limpas, a técnica *Small loss Approach* (SLA) é utilizada [5]. A técnica SLA exclui do treinamento as amostras com os maiores resultados na função de custo. Todos os modelos do SOTA citados utilizam diferentes estratégias em cima do conjunto de amostras limpas selecionadas para implementar o desempenho do modelo.

O desempenho na seleção de amostras limpas da SLA é medido através da métrica *Label Precision* [5] e varia de acordo com o ruido inserido e do dataset benchmark analisado, variando entre 70% a 80% sobre a *Label Precision* [5]. Espera-se que um modelo de Deep Learning treinado com menos amostras ruidosas tenha um desempenho melhor do que um modelo treinado com mais amostras ruidosas, pois as amostras ruidosas vão guiar a rede a aprender uma função densidade de probabilidade distinta da desejada.

Nesse trabalho, apresentamos uma forma para aumentar a performance da técnica *Small loss Approach*, aumentando a quantidade de amostras selecionadas como limpas, além de um aumento no desempenho da *Label Precision*. O aumento da quantidade de amostras selecionas como limpas é crucial para auxiliar no aprendizado da rede, pois modelos de Deep Learning exigem uma vasta quantidade de amostras para um aprendizado robusto.

Para aumentar o desempenho da SLA utilizou-se o *k-Nearest Neighbors Classifier*(KNN) [3]. Inicialmente, com o conjunto de amostras limpas previamente selecionadas pela SLA se realiza uma clusterização com o KNN. A clusterização

é feita a partir das *features* das amostras limpas retirada de uma camada interna da rede. Posteriormente, com o conjunto de amostras previamente selecionadas como ruidosas pela SLA verifica-se a qual cluster as features dessas amostras pertencem. Caso o rótulo de uma dada amostra, anteriormente selecionada como ruidosa, seja equivalente ao rotulo do cluster mais próximo de suas features essa imagem é então considerada com o rótulo correto, não sendo mais excluída do treinamento.

Nesse artigo apresentamos os estudos inicias realizados para validar essa técnica sobre o dataset benchmarks CIFAR-10 [9] com 45% de ruído inserido sinteticamente [16]. Sendo a principal contribuição desse trabalho a demostração de como o desempenho da técnica *Small loss Approach* pode ser aprimorada. Os próximos passos previstos do estudo são: ampliação da validação da técnica sobre outros datasets benchmarks como CIFAR-100 [9] e MNIST [2]. Além disso pretende-se utilizar o modelo JOCOR-SLR e Co-teaching com as melhorias na SLA sobre um problema real de importância ambiental de classificação de algas calcarias.

A organização do trabalho é dada por: Na **Seção 2** é apresentado os trabalhos relacionados com esse; Na **Seção 3** o modelo desenvolvido nesse trabalho; Na **Seção 4** os detalhes experimentais; Na**Seção 5** os resultados e discussões; E por fim ,na **Seção 6** a conclusão e próximos passos.

II. Trabalhos Relacionados

A Small Loss Approach (SLA) apresentada em [5] é uma técnica amplamente utilizada pelos modelos atuais para lidar com amostras ruidosas. Ela consiste em selecionar as amostras limpas, ou seja, que possuem o rótulo correto durante o treinamento e excluir do treino as amostras ruidosas. Ela é baseada no fato de um modelo de Deep Learning treinado com menos amostras ruidosas deve ter uma performance melhor do que um mais ruidoso.

Durante o processo de treinamento, as amostras que apresentam os menores valores de custo são selecionadas como amostras limpas (rótulos corretos). Geralmente, esse processo é realizado a cada iteração das épocas de treino. Uma vez que o conjunto de amostras limpas é selecionado, os diferentes modelos do SOTA exploram esse conjunto com diferentes técnicas para aprimorar o treinamento.

Essa técnica é baseada na observação que modelos de Deep Learning tendem a primeiro a aprender as amostras mais simples do conjunto de dados, e gradualmente passam a aprender as amostras mais complicadas. Portanto, é esperado que as amostras com rótulos equivocados sejam aprendidas em épocas mais avançadas durante o treinamento. Esse processo resulta em valores maiores para a função de custo das amostras ruidosas quando comparados com as amostras limpas. A SLA é amplamente utilizada [5], [12], [16], [17].

No trabalho em que a SLA foi introduzido, Co-teaching [5], o autor propõem o uso de duas redes neurais sendo treinadas de forma simultânea com a SLA para lidar com amostras ruidosas. Cada rede deve ser capaz de identificar amostras ruidosas distintas o que torna o processo de treinamento mais robusto contra ruido.

Durante o treinamento, cada rede recebe como entrada um *mini-batch* contendo amostras limpas e ruidosas. Então, através da SLA, cada rede seleciona quais das amostras presentes são limpas e devem ser utilizadas no processo de treinamento da outra rede. As redes em treinamento compartilham a mesma estrutura, porém não os mesmos parâmetros de pesos iniciais. A razão para utilizar duas redes nesse processo é pelo fato que duas redes podem identificar tipos diferentes de ruídos nos rótulos, uma vez que possuem habilidades diferentes, i.e. pesos iniciais diferentes. Portanto, quando duas redes compartilham as amostras limpas de cada mini-batch, o erro deve ser mutualmente reduzido.

III. Aprimorando a *Small Loss Approach*

A SLA exclui do processo de treinamento uma porcentagem de amostras equivalente ao ruido presente no conjunto de dados [12]. Uma das limitações dessa abordagem é que as amostras excluídas podem conter *features* importantes da classe da qual pertencem. Portanto, excluindo tal amostra do treinamento, o modelo de Deep Learning será privado de aprender essa *feature*, da qual pode melhorar a generalização do modelo.

Dessa forma, propomos nesse trabalho uma forma de identificar as amostras que foram excluídas erroneamente pela SLA do processo de treinamento. Ou seja, identificamos as amostras com rótulos corretos que o processo SLA identificou como equivocado e as retornamos ao treino. É importante que esse processo tenha uma acurácia elevada, pois aumentar o número de amostras ruidosas ao treinamento guiará a rede a associar as novas *features* disponíveis as classes incorretas, entretanto se as amostras selecionadas para retornarem ao treinamento tiverem os rótulos corretos as novas *features* disponíveis para a rede podem ser associadas as classes corretas, sendo benéficas ao treinamento.

A. Método

Na Figura 1 ilustramos o processo de seleção de amostras limpas apresentado no trabalho em [5]. O nosso método consiste em realizar uma clusterização com o k- Nearest Neighbors Classifier(KNN) utilizando as amostras que foram selecionadas como limpas inicialmente pela SLA. Ou seja, na Figura 1, as imagens utilizadas para a clusterização seriam as com bordas pretas que geraram os menores valores de custo na função de perdas.

Para o processo de clusterização utiliza-se como dados de entrada do KNN as *features*, das imagens selecionadas como limpas, geradas pela rede em treinamento de uma camada interna. Após o processo de clusterização, verifica-se a qual cluster as imagens que foram inicialmente excluídas pela SLA pertencem. Caso a imagem seja classificada com a mesma classe outrora atribuída a imagem, essa imagem é considerada com o rótulo correto. Sendo então transferida ao conjunto de imagens limpas e assim utilizada no processo de treinamento do modelo. Na Figura 2 esse processo está ilustrado.

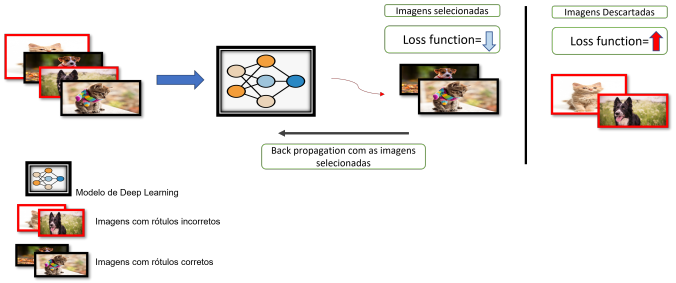

Fig. 1. Representação da seleção de amostras limpas pelo SLA. As imagens com bordas vermelhas representam as amostras com rótulos incorretos, as com bordas pretas rótulos corretos. Após o calculo da função de custo, são selecionadas as amostras com menores valores de custo como corretas, essas imagens são utilizadas para realizar o treinamento do modelo de Deep Learning. As imagens com maiores valores de custo são excluídas do processo de treinamento.

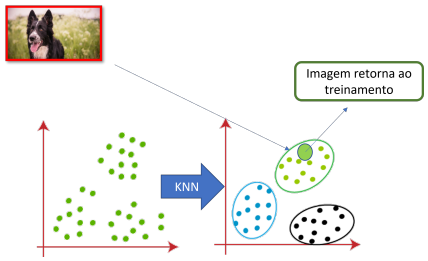

Fig. 2. Representação de uma amostra anteriormente excluída no processo de treinamento do modelo retornando ao treinamento. No exemplo, as *features* da imagem do cachorro pertencem ao cluster formado pelas imagens com rótulos corretos de cachorro. Caso o rótulo da imagem em análise, no exemplo imagem de cachorro, seja cachorro, então esse rótulo é considerado correto e essa imagem retorna ao treinamento.

B. Algoritmo

Nesse trabalho acrescentamos o aprimoramento da SLA no fluxo de treino do modelo Co-teaching, no Algoritmo 1 está apresentado a estrutura completa do treinamento. No Algoritimo 1, o valor fixo τ representa a porcentagem de ruido presente no dataset, $R(t)$ determina quantas amostras são selecionadas por época, *CE* representa a função de custo cross entropy [18] . Observa-se que o procedimento de treinamento do modelo Co-teaching foi conservado, sendo acrescentado apenas ao aprimoramento da SLA. Sendo o conjunto de amostras selecionadas pela SLA dada pela Equação 1:

$$\bar{D} = argmin_{D:|D|\geq R(T)} CE(\theta_1, D) \quad (1)$$

Onde $D = (x_i, y_i)_{i=1}^{z}$ é um conjunto de dados com amostras ruidosas, CE a função de custo para um modelo $w_f(\theta_1)$ sobre o conjunto de dado D.

IV. Experimental

Nesse trabalho comparamos o modelo Co-teaching com o aprimoramento da SLA e o modelo Co-teaching original, utilizamos a métrica acurácia sobre o conjunto de teste e a Label Precion [5]. A comparação foi feita sobre o dataset CIFAR-10 com ruído inserido sinteticamente, pois esse dataset originalmente possui apenas amostras limpas.

Algorithm 1 Aprimoramento da SLA

Require: Rede $w_f(\theta_1)$ e Rede $w_g(\theta_2)$, learning rate η, fixo τ , época T_k, T_{max}, Dataset Ruidoso $D = (x_i, y_i)_{i=1}^{z}$, $R(t)$,iteração N_{max}:

for $T = 1, 2..., T_{max}$ **do**
  Embaralhe o conjunto de dados D
  for $N = 1, 2..., N_{max}$ **do**
  Obtenha o mini-batch D_n de D
  Obtenha as amostras limpas $\bar{D}_f$ utilizando o procedimento aprimorado da SLA sobre D_n com a rede $w_f(\theta_1)$
  Obtenha as amostras limpas $\bar{D}_g$ utilizando o procedimento aprimorado da SLA sobre D_n com a rede $w_g(\theta_2)$
  Atualize: $w_f(\theta_1) = w_f(\theta_1) - \eta \nabla CE(\theta_1, \bar{D}_g)$
  Atualize: $w_g(\theta_2) = w_g(\theta_2) - \eta \nabla CE(\theta_2, \bar{D}_f)$
Atualize: $R(t) = min[\frac{\tau t}{T_k}, \tau]$
  end for
end for
=0

O ruído inserido ao dataset seguiu a estrutura pair-flip [12] como usualmente é utilizado na área de estudo de *noisy samples* [16]. O conjunto de dados foi separado em treino e teste, onde que o ruido foi inserido apenas no conjunto treino seguindo o protocolo dos trabalhos [5], [17]

A rede utilizada para comparação dos modelos consistiu da estrutura apresentada na tabela 1, sendo a função de custo a CE e foi implementada utilizando o Tensorflow 2.4. O batch size foi de 128, o learning rate de 0.001, com otimizador Adam [8].

V. Resultados e Discussões

Na Figura 3 são apresentados os resultados da acurácia de teste para os modelos *Co-teaching* e o modelo *Co-teaching* com a SLA aprimorada do qual nos referenciamos como KNN Co-teaching. Foi observado que o aprimoramento da SLA melhorou o desempenho do modelo em 1.4%.

Na Tabela II apresentamos o desempenho dos modelos para o *Label Precision* para as épocas 10 e 150. Observamos que o pico de desempenho da *Label Precision* ocorre na época 10 coincidindo com o pico de desempenho da acurácia de teste para ambos os modelos, sendo o pico de desempenho superior no KNN Co-teaching em 2% .

Destaca-se que o pico de desempenho ocorre na época 10 na acurácia de teste, do qual também ocorre no *Label Precision*, sendo equivalente ao hiper parâmetro Tk [5]. Nos próximos passos desse trabalho pretende-se explorar se o valor do TK influencia no desempenho das métricas.

TABLE I
BACKBONE UTILIZADO PARA COMPARAÇÃO DOS MODELOS

Camada	Detalhe
Entrada	**28x28x3**
Convolucional	**Filtro 128, kernel(3,3) Ativação:LeakyReLU**
Batch Normalization	**momentum=0.1,epsilon=0.00001**
Convolucional	**Filtro 128, kernel(3,3) Ativação:LeakyReLU**
Batch Normalization	**momentum=0.1,epsilon=0.00001**
Convolucional	**Filtro 256, kernel(3,3) Ativação:LeakyReLU**
Batch Normalization	**momentum=0.1,epsilon=0.00001**
Max Pooling	**strides=2,2**
Dropout	**rate=0.25**
Convolucional	**Filtro 256, kernel(3,3) Ativação:LeakyReLU**
Batch Normalization	**momentum=0.1,epsilon=0.00001**
Convolucional	**Filtro 256, kernel(3,3) Ativação:LeakyReLU**
Batch Normalization	**momentum=0.1,epsilon=0.00001**
Convolucional	**Filtro 128, kernel(3,3) Ativação:LeakyReLU**
Batch Normalization	**momentum=0.1,epsilon=0.00001**
Max Pooling	**strides=2,2**
Dropout	**rate=0.25**
Convolucional	**Filtro 128, kernel(3,3) Ativação:LeakyReLU**
Batch Normalization	**momentum=0.1,epsilon=0.00001**
Convolucional	**Filtro 128, kernel(3,3) Ativação:LeakyReLU**
Batch Normalization	**momentum=0.1,epsilon=0.00001**
Convolucional	**Filtro 128, kernel(3,3) Ativação:LeakyReLU**
Batch Normalization	**momentum=0.1,epsilon=0.00001**
Max Pooling	**strides=2,2**
Dropout	**rate=0.25**
Convolucional	**Filtro 128, kernel(3,3) Ativação:LeakyReLU**
Batch Normalization	**momentum=0.1,epsilon=0.00001**
Convolucional	**Filtro 128, kernel(3,3) Ativação:LeakyReLU**
Batch Normalization	**momentum=0.1,epsilon=0.00001**
Convolucional	**Filtro 128, kernel(3,3) Ativação:LeakyReLU**
Batch Normalization	**momentum=0.1,epsilon=0.00001**
GlobalAveragePooling2D	
Dense	**activation= Softmax**

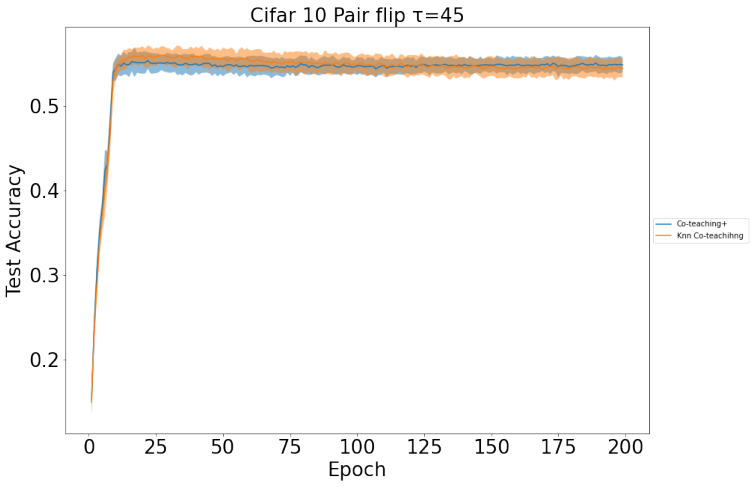

Fig. 3. Comparação do desempenho no sobre a acurácia de teste entre os modelos *Co-teaching* e o modelo *Knn-Coteaching*.

TABLE II
COMPARAÇÃO DO DESEMPENHO NO LABEL PRECISION PARA AS ÉPOCAS 10 E 150

Modelo	Época	Label Precision
Co-teaching	**10**	74%
Knn Co-teaching	**10**	76%
Co-teaching	**150**	72%
Knn Co-teaching	**150**	72%

VI. CONCLUSÃO

Nesse trabalho apresentamos os resultados iniciais utilizados para aprimorar a técnica *Small Loss Approach* amplamente utilizada no estado da arte para o tema de pesquisa noisy label. A técnica consiste em utilizar a *K-Nearest Neighbors Classifier* para recuperar uma parcela maior de amostras com rótulos corretos do processo de separação de amostras limpas e ruidosas da SLA. O primeiro experimento realizado sobre o dataset CIFAR-10 apontou uma melhoria de 1.2% no Test Accuracy e de 2% no Label Precision.

Os próximos passos desse estudo irá consistir na ampliação da validação do processo, explorando o aprimoramento da SLA em outros datasets benchmarks. Além disso, os modelos com a SLA aprimorada serão utilizadas em uma demanda real de importância ambiental para classificação de algas calcarias.

ACKNOWLEDGMENT

O autor gostaria de agradecer o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) e a Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) pelo suporte financeiro para o trabalho.

REFERENCES

[1] Coulibaly, S., Kamsu-Foguem, B., Kamissoko, D., Traore, D.: Deep convolution neural network sharing for the multi-label images classification. Machine Learning with Applications **10**, 100422 (2022)

[2] Deng, L.: The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE signal processing magazine **29**(6), 141–142 (2012)

[3] Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K.: Knn model-based approach in classification. In: On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings. pp. 986–996. Springer (2003)

[4] Guo, Q., Feng, W., Zhou, C., Huang, R., Wan, L., Wang, S.: Learning dynamic siamese network for visual object tracking. In: Proceedings of the IEEE international conference on computer vision. pp. 1763–1771 (2017)

[5] Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. Advances in neural information processing systems **31** (2018)

[6] Hu, M., Han, H., Shan, S., Chen, X.: Multi-label learning from noisy labels with non-linear feature transformation. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14. pp. 404–419. Springer (2019)

[7] Huang, L., Zhang, C., Zhang, H.: Self-adaptive training: bridging the supervised and self-supervised learning. arXiv preprint arXiv:2101.08732 (2021)

[8] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[9] Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
[10] Liu, W., Jiang, Y.G., Luo, J., Chang, S.F.: Noise resistant graph ranking for improved web image search. In: CVPR 2011. pp. 849–856. IEEE (2011)
[11] Sanderson, T., Scott, C.: Class proportion estimation with application to multiclass anomaly rejection. In: Artificial Intelligence and Statistics. pp. 850–858. PMLR (2014)
[12] Wei, H., Feng, L., Chen, X., An, B.: Combating noisy labels by agreement: A joint training method with co-regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13726–13735 (2020)
[13] Welinder, P., Branson, S., Perona, P., Belongie, S.: The multidimensional wisdom of crowds. Advances in neural information processing systems **23** (2010)
[14] Yang, Y., Newsam, S.: Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems. pp. 270–279 (2010)
[15] Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., Lyman, K.: Learning to diagnose from scratch by exploiting dependencies among labels. arXiv preprint arXiv:1710.10501 (2017)
[16] Yao, Y., Sun, Z., Zhang, C., Shen, F., Wu, Q., Zhang, J., Tang, Z.: Jo-src: A contrastive approach for combating noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5192–5201 (2021)
[17] Yu, X., Han, B., Yao, J., Niu, G., Tsang, I., Sugiyama, M.: How does disagreement help generalization against label corruption? In: International Conference on Machine Learning. pp. 7164–7173. PMLR (2019)
[18] Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems **31** (2018)