Appendix for Forming Effective Human-AI Teams: Building Machine Learning Models that Complement the Capabilities of Multiple Experts

Patrick Hemmer[1*], **Sebastian Schellhammer**[1,2*], **Michael Vössing**[1], **Johannes Jakubik**[1] and **Gerhard Satzger**[1]
[1]Karlsruhe Institute of Technology
[2]GESIS - Leibniz Institute for the Social Sciences
{patrick.hemmer, michael.voessing, johannes.jakubik, gerhard.satzger}@kit.edu,
sebastian.schellhammer@gesis.org

A Experimental Details and Results

All experiments were conducted on a Red Hat Enterprise Linux 8.2 system and a NVIDIA A100-PCIE with 40GB RAM. We used PyTorch 1.7.1 for the implementation.

A.1 Hate Speech and Offensive Language Dataset

In the following section, we provide further details on the experimental evaluation conducted on the hate speech and offensive language dataset.

Implementation Details. We use a preprocessed version of the hate speech and offensive language dataset [Davidson *et al.*, 2017] using code from Keswani *et al.* [2021][1]. We use 100-dimensional GloVe embeddings[2] as the features to train our approach and the algorithmic baselines. For our approach and all baselines except *JSF*, we use the validation split loss to perform early stopping. For the *JSF* baseline, we use the accuracy on the validation split because the weight of the system loss increases with each epoch. Hence, the validation split loss is not a reliable indicator of team performance. Our findings remain consistent when performing early stopping based on the validation accuracy with our approach and the other algorithmic baselines.

A.2 CIFAR-100 Dataset

In the following section, we provide further details on the experimental evaluation conducted on the CIFAR-100 dataset.

Implementation Details. We use the torchvision dataset provided by PyTorch[3]. The feature extractor is a ResNet-18 model provided by PyTorch[4] and is pretrained on the ImageNet dataset. We horizontally flip (with a probability of 0.5) and randomly rotate (± 15 degrees) the training and validation images. The test images are not augmented. To make use of the pretrained ImageNet weights, we upsample the 32x32 CIFAR-100 images to 224x224, i.e., the size of the images in the ImageNet dataset. For our approach and all baselines except *JSF*, we use the validation split loss to perform early stopping. For the *JSF* baseline, we use the accuracy on the validation split because the weight of the system loss increases with each epoch. Hence, the validation split loss is not a reliable indicator of team performance. Our findings remain consistent when performing early stopping based on the validation accuracy with our approach and the other algorithmic baselines.

Results. Figure A1 displays the team accuracy development of our approach and the respective baselines over the number of human experts. In general, all methods that rely on human experts except the *Random Expert* baseline benefit from an increasing number of human team members. Whereas our approach (*Classifier & Expert Team*) outperforms all baselines, the *JSF* baseline assigns the instances to the best human expert and therefore does not leverage the existing potential for performance improvements due to disagreement among human experts. Moreover, from the comparison of our approach with the *Expert Team* baseline, we find that the contribution of the classifier in our approach diminishes with increasing team member size. As more human experts enter the team, larger areas of the feature space in which they make accurate predictions are potentially covered.

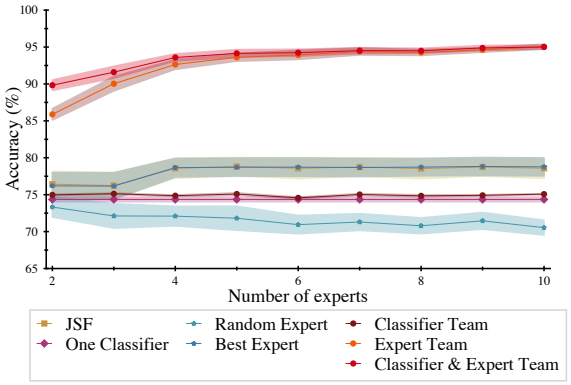

Figure A1: Team accuracy of our approach and the baselines with increasing team size on CIFAR-100. Shaded regions display standard errors.

A.3 NIH Dataset

In the following section, we provide further details on the experimental evaluation conducted on the NIH dataset.

[*]These authors contributed equally.
[1]https://github.com/vijaykeswani/Deferral-To-Multiple-Experts
[2]https://nlp.stanford.edu/projects/glove/
[3]https://pytorch.org/vision/stable/datasets.html
[4]https://pytorch.org/vision/stable/models.html

Implementation Details. We load the NIH Chest X-ray dataset provided by the NIH Clinical Center[5] [Wang *et al.*, 2017]. The feature extractor is a ResNet-18 model[6] pretrained on 224x224 images from the CheXpert dataset [Irvin *et al.*, 2019]. We downsample the NIH images to 224x224 and convert them to RGB. Furthermore, we use individual radiologist labels as well as adjudicated labels[7]. The latter serve as "gold standard" [Majkowska *et al.*, 2020].

We perform a 10-fold cross-validation, repeat the experiment 5 times with different seeds and report the mean performance. To calculate the cross-validation performance for one seed, we first concatenate the predictions for the test fold of each of the 10 runs to obtain the predictions for the whole dataset. Then we compare these predictions against the ground truth labels. For our approach and all baselines except *JSF*, we use the loss on the 2 validation folds to perform early stopping. For the *JSF* baseline, we use the accuracy on the 2 validation folds because the weight of the system loss increases with each epoch. Hence, the loss on the 2 validation folds is not a reliable indicator of team performance. Our findings remain consistent when performing early stopping based on the validation accuracy with our approach and the other algorithmic baselines. Additionally, we use stratification based on the performance of the two radiologists per patient. For each patient, we calculate the performance of both radiologists in terms of accuracy. We then concatenate both accuracies as the target variable to obtain the stratified 10-fold cross-validation splits. This ensures a low variance between a radiologist's performance across the 10 folds.

B Derivations

In this section, we provide an intuition about the loss function's inner workings based on an analysis of the partial derivatives of $\mathcal{L}_{team}$ (Equation (6)) with respect to the allocation system's output $\boldsymbol{a}$ and the classifier's output $\boldsymbol{z}$. The team probability $P_{team}(Y = i|\boldsymbol{x})$ of observing class i given input $\boldsymbol{x}$ will be denoted by P_i from now on.

First, we consider the partial derivatives of $\mathcal{L}_{team}$ (Equation (6)) with respect to the outputs of the allocation system $\boldsymbol{a}$. Using the chain rule we have

$$\frac{\partial \mathcal{L}_{team}}{\partial a_j} = \sum_{i=1}^{k} \sum_{l=1}^{m+1} \frac{\partial \mathcal{L}_{team}}{\partial P_i} \frac{\partial P_i}{\partial w_l} \frac{\partial w_l}{\partial a_j} j = 1, \ldots, m+1\,. \tag{7}$$

Looking at Equation (6), we note that only P_y contributes to the loss. Thus, the partial derivatives of $\mathcal{L}_{team}$ with respect to the conditional probabilities P_i are

$$\frac{\partial \mathcal{L}_{team}}{\partial P_i} = \begin{cases} -\frac{1}{P_y}, & \text{if} \quad y = i \\ 0, & \text{otherwise}\,, \end{cases} \qquad i = 1, \ldots, k\,, \tag{8}$$

where y is the ground truth label. Further, differentiating Equation (5), we have the partial derivatives of P_i with respect to the probabilities $\boldsymbol{w}$

$$\begin{aligned} \frac{\partial P_i}{\partial w_l} &= \begin{bmatrix} \frac{\partial P_1}{\partial w_1} & \cdots & \frac{\partial P_1}{\partial w_{m+1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial P_k}{\partial w_1} & \cdots & \frac{\partial P_k}{\partial w_{m+1}} \end{bmatrix} \\ &= \begin{bmatrix} T_{1,1} & \cdots & T_{1,m+1} \\ \vdots & \ddots & \vdots \\ T_{k,1} & \cdots & T_{k,m+1} \end{bmatrix} \\ &= T_{i,l} \qquad i = 1, \ldots, k \text{ and } l = 1, \ldots, m+1\,, \end{aligned} \tag{9}$$

with T combining the one-hot encoded predictions of the human experts and the conditional probability distribution of the classifier. Next, differentiating the softmax function in Equation (2), the partial derivatives of $\boldsymbol{w}$ with respect to the allocation system outputs $\boldsymbol{a}$ are

$$\begin{aligned} \frac{\partial w_l}{\partial a_j} &= \begin{bmatrix} \frac{\partial w_1}{\partial a_1} & \cdots & \frac{\partial w_1}{\partial a_{m+1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial w_{m+1}}{\partial a_1} & \cdots & \frac{\partial w_{m+1}}{\partial a_{m+1}} \end{bmatrix} \\ &= w_l(\delta_{lj} - w_j) \qquad\qquad j,\, l = 1, \ldots, m+1\,, \end{aligned} \tag{10}$$

where δ_{lj} is the Kronecker delta with

$$\delta_{lj} = \begin{cases} 1 & \text{if} \quad l = j \\ 0 & \text{otherwise} \end{cases} \qquad j,\, l = 1, \ldots, m+1\,. \tag{11}$$

Using Equations (8) to (10) we can rewrite Equation (7) with $y = i$ to

$$\frac{\partial \mathcal{L}_{team}}{\partial a_j} = w_j - w_j \frac{T_{y,j}}{P_y} \qquad j = 1, \ldots, m+1\,. \tag{12}$$

The quotient in the gradient of Equation (12) indicates whether the probability that team member j predicts the correct class y is higher than the probability of the team. As a result, the allocation system increases the weights of those team members with better predictions than the team prediction. Conversely, the weights of team members whose predictions are worse than the team prediction are decreased.

Second, we consider the partial derivatives of $\mathcal{L}_{team}$ with respect to the outputs of the classifier $\boldsymbol{z}$ using the chain rule

$$\frac{\partial \mathcal{L}_{team}}{\partial z_j} = \sum_{i=1}^{k} \sum_{l=1}^{k} \frac{\partial \mathcal{L}_{team}}{\partial P_i} \frac{\partial P_i}{\partial c_l} \frac{\partial c_l}{\partial z_j} \qquad j = 1, \ldots, k\,. \tag{13}$$

In Equation (8) we have shown the partial derivative of $\mathcal{L}_{team}$ with respect to P_i. Furthermore, the influence of the classifiers' conditional probabilities $\boldsymbol{c}$ to the conditional team probabilities P_i is weighted by w_{m+1}. The partial derivative of P_i with respect to the conditional probabilities $\boldsymbol{c}$ then is

$$\frac{\partial P_i}{\partial c_l} = \begin{cases} w_{m+1}, & \text{if} \quad i = l \\ 0, & \text{otherwise} \end{cases} \qquad i, l = 1, \ldots, k\,. \tag{14}$$

[5]https://nihcc.app.box.com/v/ChestXray-NIHCC

[6]https://github.com/habbes/chest-xrays

[7]https://cloud.google.com/healthcare/docs/resources/public-datasets/nih-chest

Differentiating the softmax function in Equation (3), the partial derivatives of $\boldsymbol{c}$ with respect to the output of the classifier $\boldsymbol{z}$ are

$$\frac{\partial c_l}{\partial z_j} = \begin{bmatrix} \frac{\partial c_1}{\partial z_1} & \cdots & \frac{\partial c_1}{\partial z_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial c_k}{\partial z_1} & \cdots & \frac{\partial c_k}{\partial z_k} \end{bmatrix} = c_l(\delta_{lj} - c_j) \qquad l, j = 1, \ldots, k\,. \tag{15}$$

Combining Equations (8), (14) and (15) we can rewrite Equation (13) to

$$\frac{\partial \mathcal{L}_{team}}{\partial z_i} = \begin{cases} \frac{w_{m+1}\, c_y}{P_y}\,(c_y - 1), & \text{if } y = i \\ \frac{w_{m+1}\, c_y}{P_y}\, c_i, & \text{if } y \neq i \end{cases} \quad i = 1, \ldots, k\,. \tag{16}$$

The gradients in Equation (16) bear a close resemblance to the gradients of the standard cross-entropy loss. The direction of the gradients is defined by the ground truth label y. We see that the gradient is non-positive for the output z_y corresponding to the ground truth label y and non-negative for the remaining outputs. Further, the magnitude of the gradients is influenced by the probability the allocation system assigns to the classifier w_{m+1} and by the quality of the classifier's prediction compared to the team prediction, i.e., the division of c_y by P_y. These two factors address the joint learning and give weight to those instances where the classifier is assigned a high probability by the allocation system and/or where the classifier performs better than the team.

References

[Davidson *et al.*, 2017] Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *ICWSM*, 2017.

[Irvin *et al.*, 2019] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, et al. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, 2019.

[Keswani *et al.*, 2021] Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Towards unbiased and accurate deferral to multiple experts. In *AIES*, 2021.

[Majkowska *et al.*, 2020] Anna Majkowska, Sid Mittal, David F Steiner, Joshua J Reicher, Scott Mayer McKinney, Gavin E Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Kalidindi, et al. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 2020.

[Wang *et al.*, 2017] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, 2017.