Data-limited Domain Adaptation and Transfer Learning for Learning Latent Expression Labels of Child Facial Expression Images

Megan A. Witherow[1], Winston J. Shields[1], Manar D. Samad[2], Khan M. Iftekharuddin[1]
[1] Vision Lab, Dept. of Electrical and Computer Engineering, Old Dominion University
[2] Dept. of Computer Science, Tennessee State University

Background & Motivation

Child facial expression analysis (FEA) may be used to leverage the psychophysical information in facial expressions to characterize differences in social communication and identify markers for developmental and behavioral disorders in children. These markers may be used to measure and monitor symptoms to support early diagnosis and intervention. There are **two primary challenges** in child FEA:

1. **Scarcity of labeled facial expression images for children:**
   This scarcity limits the performance of deep learning models due to **overfitting**. Therefore, adult ground truth data are almost invariably used to train, validate, and test FEA models even for classifying child facial expressions.
2. **Child expressions differ from adult expressions:**
   Developmental differences in face proportions and motor ability results in a **distribution shift** between the adult and child expression domains, making adult ground truth data inappropriate for child facial expression classification.

The **deep learning-based approach** to classification learns a function to classify unseen samples based upon available training examples.

- Assuming that samples are independently and identically distributed (i.i.d.), as the number of training samples increases, the trained model converges to the optimal model.
- In practice, the **i.i.d. assumption is often violated** and training examples are limited, resulting in suboptimal models that may not generalize well outside of the domain on which they are trained.

Transfer learning and **domain adaptation** attempt to improve upon the generalizability of models to a target domain that differs from the source domain on which a model is trained.

- **Transfer Learning (TL)** applies information learned from one domain to a task on another domain. It assumes a relationship between learning tasks on the source and target domains, which may have different input and output spaces with different data distributions.
- **Domain adaptation (DA)** seeks to learn a domain-invariant latent representation, and thus requires the input and output spaces to be the same. The only difference between the two domains is a shift in the data distributions.

We hypothesize that combining DA+TL will offer superior results to TL alone on a child facial expression classification task when training with very few labeled samples.

Data

Two data sets of facial expression images: **adult "source" domain** [1] and **child "target" domain** [2,3]. Seven expression categories: **'anger', 'disgust', 'fear', 'happy', 'neutral', 'sad', 'surprise'.**

Source Data	Target Data
Extended Cohn-Kanade (CK+) [1]	Child Affective Facial Expression (CAFE)[2,3]
$\boldsymbol{D_S} = \{(x_i^S, y_i^S) \mid x_i^S \in \boldsymbol{\mathcal{X}}, y_i^S \in \boldsymbol{\mathcal{Y}}\}_{i=1}^{N_S}$	$\boldsymbol{D_T} = \{(x_i^T, y_i^T) \mid x_i^T \in \boldsymbol{\mathcal{X}}, y_i^T \in \boldsymbol{\mathcal{Y}}\}_{i=1}^{N_T}$
$x_i^S \sim p_S$	$x_i^T \sim p_T$
N_S is the total number of samples	N_T is the total number of samples

input space $\mathcal{X} \subset \mathbb{R}^{n \times n}$ is the set of all possible input images
output space $\mathcal{Y} = \{0, \ldots, K\}$ is the set of K class labels

References

[1] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z. and Matthews, I., "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Comput. Soc. Conf. Comput. Vis. pattern recognition-workshops, 94–101, IEEE (2010).
[2] LoBue, V. and Thrasher, C., "The Child Affective Facial Expression (CAFE) set" (2014).
[3] LoBue, V. and Thrasher, C., "The Child Affective Facial Expression (CAFE) set: Validity and reliability from untrained adults," Front. Psychol. 5, 1532 (2015).
[4] Witherow, M. A., Samad, M. D. and Iftekharuddin, K. M., "Transfer learning approach to multiclass classification of child facial expressions," Proc. SPIE - Int. Soc. Opt. Eng. 11139 (2019).
[5] Witherow, M. A., Shields, W. J., Samad, M. D. and Iftekharuddin, K. M., "Learning latent expression labels of child facial expression images through data-limited domain adaptation and transfer learning", Proc. SPIE 11511, Applications of Machine Learning 2020, 115110E (2020)
[6] Motiian, S., Piccirilli, M., Adjeroh, D. A. and Doretto, G., "Unified Deep Supervised Domain Adaptation and Generalization," Proc. IEEE Int. Conf. Comput. Vis. (2017).

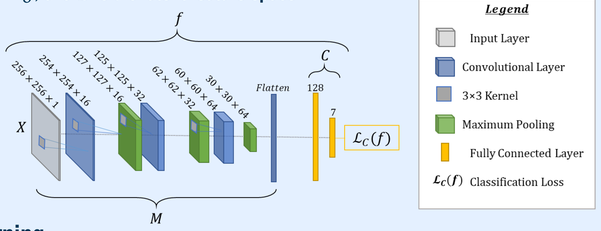
Methods

Model Architecture

We model the **deep learning approach for classification** as:

$$Y = f(X),$$

where the goal is to learn a model $f(\cdot)$ that optimally predicts class labels $Y \in \mathcal{Y}$ given input images $X \in \mathcal{X}$. The model $f(\cdot)$ can be further defined as $f = M \circ C$, where feature extractor $M : \mathcal{X} \to \mathcal{Z}$, classifier $C : \mathcal{Z} \to \mathcal{Y}$, and $\mathcal{Z}$ is the **latent feature space.**



Transfer Learning

The learned solution for source task $\boldsymbol{T_S}$ **(adult facial expression classification)** is leveraged to solve related target task $\boldsymbol{T_T}$ **(child facial expression classification).**

T_S corresponds to learning $f_S: X_S \to Y_S$ where $X_S = \{x_i^S\}_{i=1}^{N_S} \in \mathcal{X}$, $Y_S = \{y_i^S\}_{i=1}^{N_S} \in \mathcal{Y}$,

T_T corresponds to learning $f_T: X_T \to Y_T$ where $X_T = \{x_i^T\}_{i=1}^{N_T} \in \mathcal{X}$, $Y_T = \{y_i^T\}_{i=1}^{N_T} \in \mathcal{Y}$

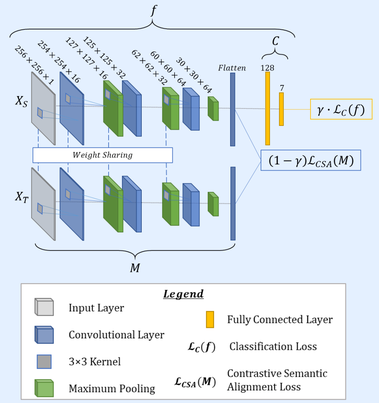
Full implementation and training details available in our publications [4,5].

Domain Adaptation

We assume the distribution shift can be attributed to **covariate shift, i.e.** $\boldsymbol{p_S(x) \neq p_T(x)}$, rather than a shift in the label distributions and that D_T has very few labeled samples.

Based on these assumptions, we adapt the **classification and contrastive semantic alignment (CCSA) approach** [6], for our adult-to-child facial expression adaptation task. This approach uses a CCSA loss composed of **classification loss** $\boldsymbol{\mathcal{L}_C(f)}$ and **contrastive semantic alignment loss** $\boldsymbol{\mathcal{L}_{CSA}(M)}$. The approach employs a Siamese network architecture with input streams for source samples X_S and target samples X_T.

Full implementation and training details in [5].



Proposed DA+TL

We leverage the principles of deep TL to improve the CCSA DA approach on limited data. We define **Source Weights Initialization (SWI)** as initializing $M(\cdot)$ and $C(\cdot)$ in the CCSA model architecture with trained weights from the source model.

We consider two variants of our proposed DA+TL model: (1) **DA+TL with SWI only (DA+SWI),** i.e. weights in all layers are considered trainable, and (2) **DA+SWI with layer freezing (DA+SWI+LF),** i.e. weights in the last convolutional block of $M(\cdot)$ and all weights of $C(\cdot)$ are considered trainable while the remainder of the architecture is frozen. Full implementation and training details in [5].

Experiments

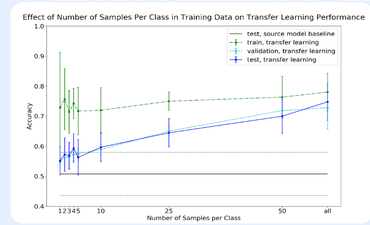
To demonstrate the **effect of the number of target samples per class in the training data on the TL approach**, we train/validate/test TL models varying the number of samples per class.

We consider **1, 2, 3, 4, 5, 10, 25, and 50 target samples per class**, as well as **'all'** target training samples not held out for testing.

To evaluate the **efficacy of proposed models with our data-limited child facial expression classification task**, we train/validate/test **five model configurations**:

(1) source baseline (2) transfer learning (3) CCSA (4) DA+SWI (5) DA+SWI+LF

Results & Discussion



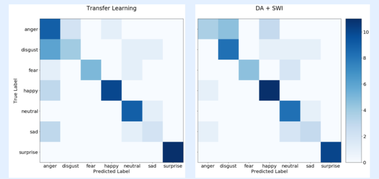
In the figure (left), the performance of transfer learning models is substantially impacted by the number of target samples per class. When the number of target samples per class is in the single digits, there is a 15% gap between training accuracy and validation and test accuracies, indicating overfitting. Mean accuracy increases with the number of target samples per class to reduce the magnitude of overfitting.

Comparison Across Approaches of Mean Test Accuracies from 10-fold Subject Independent Cross Validation

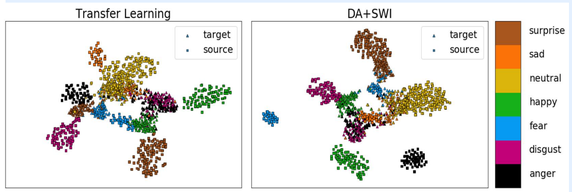
Model	Number of Target Samples per Class in Training Set							
	0	1	2	3	4	5	10	All
Source Baseline	50.76 % ± 7.18 %	-	-	-	-	-	-	-
Transfer Learning	-	55.15% ± 4.69%	57.27 % ± 5.60 %	56.85 % ± 4.43 %	59.38 % ± 4.73 %	56.28 % ± 5.95 %	59.67 % ± 4.71 %	74.79 % ± 6.06 %
CCSA	-	59.67 % ± 6.14 %	62.79 % ± 5.13 %	63.08 % ± 5.53 %	67.04 % ± 3.45 %	65.74 % ± 5.52 %	70.29 % ± 4.53 %	-
DA+SWI	-	61.95 %± 3.31 %	63.63 % ± 4.45 %	63.95 % ± 5.35 %	68.45 % ± 4.44 %	68.58 % ± 7.27 %	**71.84 % ± 6.14 %**	-
DA+SWI+LF	-	58.42 %± 4.00 %	62.79 % ± 4.53 %	64.50 % ± 3.30 %	67.19 % ± 3.69 %	66.18 % ± 5.65 %	70.00 % ± 4.97 %	-

The source baseline model performs the worst out of all approaches. The best performing model is the transfer learning model using all target training samples. When possible, collecting more training data is the best way to improve model performance. However, when collecting additional data is difficult/ not possible, DA approaches may offer substantial increases in model performance over TL. The DA approaches outperform the TL approach for all values of number of target samples per class 1 to 10. Of the three DA approaches, DA+SWI has a higher mean test accuracy than CCSA across all number of samples per class.

Representative confusion matrices obtained from the same test fold for the transfer learning approach using all target training samples and the DA+SWI approach using 10 target samples per class are shown (right). The models follow similar patterns of class confusion. Neither model misclassifies 'surprise' samples because of their distinctive open-mouth appearance. The 'disgust' and 'anger' expressions are the most confusing for both models.



Representative confusion matrices of the same cross validation fold for the transfer learning approach using all samples and the DA+SWI approach using 10 target samples per class.



2D visualization of the latent space using the t-SNE algorithm. The left plot is produced using the transfer learning model trained using all available target training samples. The right plot is produced using the DA+SWI model trained with 10 target samples per class.

The t-SNE algorithm is used to visualize the latent space of both models (left). Here, we see that confusing classes, such as 'anger' and 'disgust' as well as 'sad' and 'neutral' are mapped closely together in the latent space. This may be because these pairs are visually similar.

Conclusion & Future Work

This work:

- Demonstrates the **advantage of the DA approach over traditional TL** for child facial expression classification data with 10 or fewer samples per class
- Shows that **initializing the model architecture with pretrained weights** learned from adult facial expression data improves model performance
- Suggests that even a **small amount of annotated data** may be leveraged to **substantially improve expression classification performance for children**

In future work, we plan to improve upon the class alignment between domains in the latent space and extend to the more challenging child FEA task of facial action coding system (FACS) AU classification.

Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1753793. This research was supported by the Research Computing clusters at Old Dominion University under National Science Foundation Grant No. 1828593.

Contact Email: mwith010@odu.edu (Megan Witherow)