Neural Computing and Applications (2023) 35:10535–10549
https://doi.org/10.1007/s00521-023-08248-y

ORIGINAL ARTICLE

Meta-transfer learning for emotion recognition

Dung Nguyen[1,3] · Duc Thanh Nguyen[1] · Sridha Sridharan[2] · Simon Denman[2] · Thanh Thi Nguyen[1] · David Dean[2] · Clinton Fookes[2]

Received: 3 February 2022 / Accepted: 6 January 2023 / Published online: 24 January 2023

Abstract
Deep learning has been widely adopted in automatic emotion recognition and has lead to significant progress in the field. However, due to insufficient training data, pre-trained models are limited in their generalisation ability, leading to poor performance on novel test sets. To mitigate this challenge, transfer learning performed by fine-tuning pr-etrained models on novel domains has been applied. However, the fine-tuned knowledge may overwrite and/or discard important knowledge learnt in pre-trained models. In this paper, we address this issue by proposing a PathNet-based meta-transfer learning method that is able to (i) transfer emotional knowledge learnt from one visual/audio emotion domain to another domain and (ii) transfer emotional knowledge learnt from multiple audio emotion domains to one another to improve overall emotion recognition accuracy. To show the robustness of our proposed method, extensive experiments on facial expression-based emotion recognition and speech emotion recognition are carried out on three bench-marking data sets: SAVEE, EMODB, and eNTERFACE. Experimental results show that our proposed method achieves superior performance compared with existing transfer learning methods.

Keywords Transfer learning · Emotion recognition · Facial expression-based emotion recognition · Speech emotion recognition

1 Introduction

Emotions of human manifest in facial expressions, voices, gestures, and postures. An accurate emotion recognition system based on one or a combination of these modalities would be useful in many downstream applications including medical data analytics, robotics, human computer interaction, affective computing, and automobile safety [29, 30]. Literature has shown a strong focus on applying facial expression recognition to building reliable emotion recognition systems. However, this approach faces another challenging problem as very subtle emotional changes manifested in facial expression could go undetected [29]. Recently deep learning techniques have been applied to this research problem and achieved considerable progress [1, 9, 16].

In addition to facial expression-based emotion recognition stream, speech signals, which are regarded as one of the most natural media of human communication, carry both content of explicit linguistic and information of implicit paralinguistic expressed by speakers [46]. Due to the richness of this source of information, over the last two decades numerous studies and efforts have been devoted to automatic and accurate detection of human emotions from speech signals. Similarly to facial expression-based emotion recognition, early attempts in speech emotion recognition have utilised handcrafted acoustic features to describe paralinguistic information [19, 38]. These methods have been surpassed by deep learning techniques,

✉ Duc Thanh Nguyen
duc.nguyen@deakin.edu.au

[1] School of Information Technology, Deakin University, 75 Pigdons Road, Waurn Ponds, VIC 3216, Australia

[2] Speech, Audio, Image and Video Technology (SAIVT) Laboratory, Queensland University of Technology, 2 George Street, Brisbane, QLD 4000, Australia

[3] School of Food and Agricultural Sciences, The University of Queensland, 5391 Warrego Hwy, Gatton, QLD 4343, Australia

which are capable of automatically and directly learning features from training data.

Although deep learning-based approaches have made great contributions to progressing the emotion recognition research, those approaches strongly rely on training data sets, which are supposed to capture sufficient and diverse information about the problem domain. Furthermore, there is often a domain shift between a source domain and a target domain, which is unknown in practice. To adapt a pre-trained model to a target domain, fine-tuning is often adopted. To the best of our knowledge, fine-tuning is arguably the most widely exploited method for transfer learning while working with deep architectures. It begins with a pre-trained model that has been trained on a source domain and further refines the model on a target domain. Compared with training from scratch, fine-tuning a pre-trained off-the-shelf model on a target data set can considerably boost up the performance of the model, whereas lessening annotated data requirements on the target domain [13]. However, pre-trained models (even after being fine-tuned) are still limited in their generalisation capability and thus often perform poorly on novel test sets. This is due to irrelevant features learnt from source domain may sill retain in a pre-trained model while important features may be forgotten after fine-tuning.

To resolve these issues, the authors in [11] exploited a progressive method, originally proposed by [35], to transfer knowledge across three paralinguistic tasks: emotion, speaker, and gender recognition. Nevertheless, this method is computationally expensive as the number of task-dependent models keeps growing proportionally to the number of studied tasks [24]. Recently, Fernando et al. [10] proposed PathNet for transfer learning between various tasks. PathNet is a neural network in which pathways (also called agents) through different layers of the network are learnt to specific tasks. Agents also hold an accountability for determining which parameters to be updated for subsequent learning. Pathways for different tasks are selected using genetic algorithm [15].

Inspired by the success of PathNet in transfer learning in multi-task learning problem, in this paper, we propose a meta-transfer learning method for emotion recognition. Specifically, we first investigate the effectiveness of PathNet in transferring emotional knowledge between different visual data sets. We next investigate whether similar techniques can be used for transferring emotional knowledge from speech signal.

Meta-learning is a task-level learning approach with the goal of accumulating experience from learning multiple tasks. Model agnostic meta-learning (MAML) [39], a state-of-the-art representative of this technique, learns to find the optimal initialisation state to quickly adapt a base learner to a new task. Similarly to MAML, our transfer learning also acts as a meta-learner which learns an optimal pathway from a source domain. The parameters learnt from that optimal pathway then can be used as initialisation and transferred into different target domains.

In summary, we make the following contributions in our work.

- We introduce a novel transfer learning method for emotion recognition based on PathNet to deal with the problem of insufficient data and the catastrophic forgetting issue commonly experienced in traditional transfer learning techniques.
- We demonstrate the potential of our method in two case studies: transferring emotional knowledge from one visual/audio emotion data set into another visual/audio emotion data set and from multiple speech emotion data sets into a single speech emotion data set.
- We conducted extensive experiments on three commonly used benchmark emotion data sets: EMODB, eNTERFACE, and SAVEE. Experimental results show that our proposed method effectively transfers emotional knowledge across domains and significantly outperforms existing transfer learning schemes on all the data sets and case studies.

The remainder of this paper is organised as follows. Section 2 summarises related research. Section 3 presents our proposed method. Section 4 reports our experimental results, and Sect. 5 concludes the paper with remarks.

2 Related work

Facial expression-based emotion recognition has been a well-studied research topic. A recent literature review of deep learning for facial expression-based emotion recognition can be found in [26]. In this section, we limit our review to only deep learning-based speech emotion recognition techniques.

2.1 Deep learning for speech emotion recognition

Inspired by the success of deep learning in various fields, many deep learning-based methods have been developed for speech emotion recognition. For instance, Kim et al. [20] proposed an architecture for extracting local invariant features in spectral domain by combining long short-term memory (LSTM), fully convolutional neural network (FCN), and convolutional neural network (CNN). In this method, long-term dependencies can be well captured, thereby making utterance-level features discriminative. Moreover, by embedding identity skip connections in the architecture, this method could mitigate over-fitting.

In [22], three-dimensional convolutional neural networks were proposed to model spectro-temporal dynamics by simultaneously extracting short-term and long-term spectral features. Another manner to encode the temporal information in utterances for emotion recognition is the use of recurrent networks [44]. Recently, attention mechanism [40] has been applied to further improve deep learning-based emotion recognition [33, 44, 48, 49]. In [4], speech emotion recognition was paired with speech-to-text in a multi-task learning framework.

Spectrograms of speech signals can also be considered as images on which CNNs can be applied [2, 5, 46]. For instance, in [46], three channels of the Mel-spectrogram of speech signal including static, delta, and delta delta were treated as three channels of a colour image to be fed to a CNN. Combination of multiple feature types, e.g. spectrogram-based features, Mel-frequency cepstrum coefficients, and wave-based features, for emotion recognition has also been explored recently [49].

In order to handle the mismatch between training and test data, Kim et al. [21] formulated emotion recognition as a multi-task learning problem where gender and naturalness were considered as auxiliary tasks. Experimental results on within-corpus and cross-corpus scenarios showed that the multi-task learning approach could improve the generalisation ability of the speech emotion recognition system. Sahu et al. [36] exploited adversarial auto-encoders for (i) compressing high-dimensional emotional utterances into low-dimensional vectors without sacrificing discriminative power of the utterances and (ii) synthesising emotional features. This system mainly concentrates on detecting emotions at utterance-based level instead of frame-based level.

Multimodal information also show their capability of boosting up the performance of emotion recognition. For instance, the authors in [29, 30] proposed deep architectures for learning features from audio and video streams. In [31], auto-encoders were used for learning representative yet compact audio-visual features while LSTM was employed to model the sequential structure of these audio-visual features.

2.2 Transfer learning in speech emotion recognition

Deep learning-based emotion recognition is often hindered by the lack of large and diverse databases for training deep learning models. To address this issue, one often creates a pre-trained model on some generic data sets, e.g. ImageNet, and then fine-tunes the model on domain-specific, i.e. emotion, data sets. Examples of this approach include [18, 20–22, 28, 46, 47].

Transfer learning has also been adopted to address cross-corpus and cross-language scenarios. For instance, Latif et al. [23] proposed a transfer learning technique for deep belief networks (DBNs) to recognise emotion from various languages. Experimental results on five different corpora in three different languages demonstrate the robustness of the proposed method. These results also indicate that use of a large number of languages and a small part of target data during training could dramatically strengthen the recognition accuracy. However, that method was experimented independently on different data sets, each of which acquired a different set of emotions. Therefore, it is not clear if transfer learning could improve the overall recognition performance on all the experimented data sets.

Researchers have also explored whispered speech emotion recognition, where different feature transfer learning methods have been developed by utilising shared-hidden-layer auto-encoders, extreme learning machines auto-encoders, and denoising auto-encoders [6]. The key idea of this approach is to develop a transformation for automatically capturing useful features hidden in data, and for transferring the learnt features from a source domain-training (normal phonated speech) to a target domain-test (whispered speech). In another study, Deng et al. [7] pointed out that many speech emotion recognition systems usually demonstrate poor performance on speech data when there is significant difference between training and test speech arising from the variations in the linguistic content, speaker accents, and domain/environmental conditions. To overcome this issue, an unsupervised domain adaptation algorithm was introduced and trained by simultaneously learning discriminative information from labelled data and incorporating prior knowledge from unlabelled data.

Although transfer learning has been widely applied to emotion recognition, there still remain difficulties in adapting a pre-trained model from a source domain to a target domain. These difficulties include forgetting important knowledge while transferring irrelevant knowledge between the two domains.

3 Proposed method

We propose in this paper a meta-transfer learning method by adopting PathNet [10] to solve the problem of emotion recognition across domains. PathNet is suitable for insufficient data problems as its architecture, although being large and complex, allows effective learning on small data sets. This is because the learning is only applied to sub-networks (i.e. pathways) with smaller number of parameters. Our method includes two main components: an input

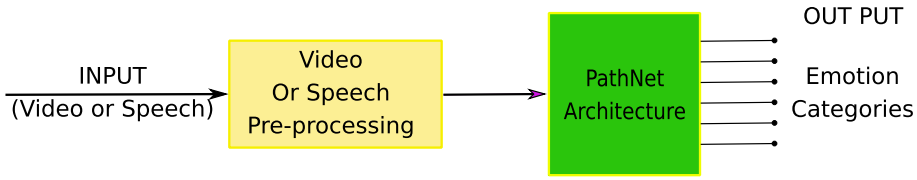


Fig. 1 Pipeline of our method

pre-processing component for video/speech processing, followed by a PathNet-based component for emotional knowledge transferring in emotion classification. We depict the pipeline of our method in Fig. 1.

For video stream, face detection is applied to extract all face regions (see Sect. 3.1). For audio stream, we initially extract three channels of the log Mel-spectrograms (static, delta, and delta delta) from audio segments over all utterances in the audio stream (see Sect. 3.2). Output of the pre-processing component is subsequently fed into our PathNet-based component (see Sect. 3.3) to classify a final facial expression or a speech emotion score. The PathNet-based component can automatically find the optimal pathway for the classification of emotion. In addition, pathways in this PathNet-based component are learnt to adapt to various target domains, and thus enabling the transferring of emotional knowledge across domains. To the best of our knowledge, our proposed method is the first attempt to investigate PathNet in dealing with the dearth of suitable data sets in emotion recognition. The procedures of feature learning and our PathNet architecture are described in more detail in the following subsections.

3.1 Video pre-processing

We assume that human emotions are obtained from facial expressions. We further assume that there is at least one human face in the input video stream. Therefore, we first apply the well-known face detection algorithm in [42] to extract all face regions from all image frames in the input video stream. Each detected face is delineated by a bounding box enclosing that detected face. In case where the face detection algorithm detects no faces, the face that has been most recently detected in previous frames in the video stream is used. If the face detection algorithm produces more than one face in a frame, the face with highest detection score is considered. This pre-processing step results in a sequence of human faces that will be passed into the PathNet-based component for feature learning and emotion classification. Figure 2a illustrates this pre-processing step.

As shown in the literature, there are other modern face detection algorithms built upon deep learning techniques, e.g. multi-task cascaded convolutional networks [45], OverFeat [37], RetinaFace [8]. However, since our video data includes relatively simple face images, e.g. plain background, one face per video sequence, we found that the cascaded face detector developed by Viola and Jones in [42] is accurate enough for our task, while performing very fast. In addition, the main focus of our work is the transfer learning algorithm, which can be applied to different face detectors.

3.2 Audio pre-processing

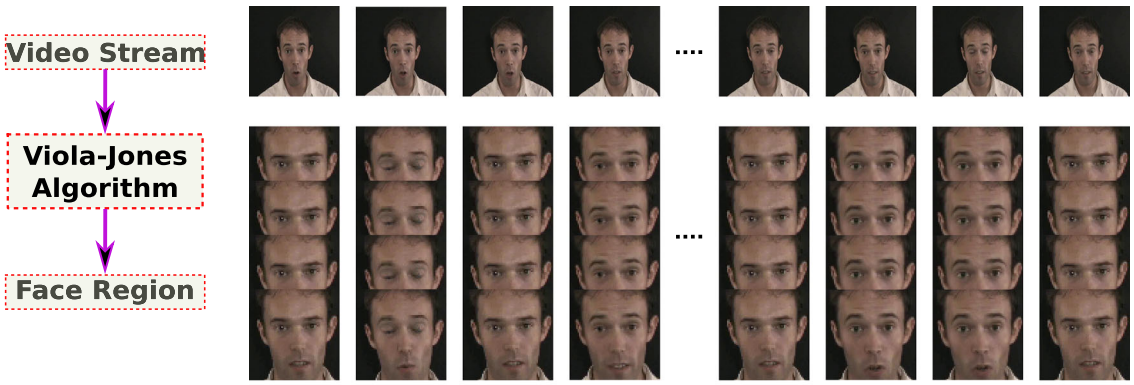
Zhang et al. [46] pointed out that the handcrafted features, such as RASTA-PLP [17], pitch frequency features, energy-related features [41], formant frequency [43], zero crossing rate (ZCR) [34], Mel-frequency cepstrum coefficients (MFCC) and its first derivative, linear prediction cepstrum coefficients (LPCC), linear prediction coefficients (LPC) [32, 38], are not discriminative enough for recognising subjective emotions. Therefore, in order to enhance the performance of speech emotion recognition system, instead of exploiting such hand-crafted features, we create images from the log Mel-spectrograms of audio segments over all utterances. As shown in the literature, Mel-frequency-based features that can be extracted from Mel-spectrograms have shown their capability of capturing important characteristics of human speech and thus have often been used in speech emotion recognition (e.g. [2, 5, 46]). In addition, Mel-spectrograms can be naturally shaped as 2D images, which are conventional data format for CNNs.

Specifically, given a speech utterance (i.e. 1D signal), $F = 64$ Mel-filter banks (from 20 to 8000 Hz) are first applied with a 25 ms Hamming window and 10ms overlapping to compute the log Mel-spectogram for the entire signal. The Mel-spectrogram is then segmented by a context window of $T = 64$ frames (corresponding to 10 ms $\times$ 63 + 25 ms = 655 ms) and 30 frame overlapping. On each window, $C = 3$ coefficients including static, delta, and delta-delta coefficients are extracted, resulting in an image in size $F \times T \times C$.

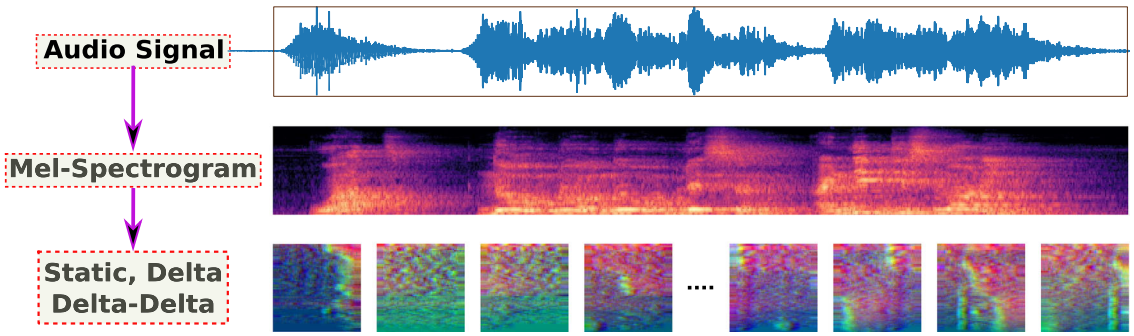
3.3 PathNet

3.3.1 Architecture

Our PathNet includes $L = 3$ layers, each layer contains $M = 20$ modules. Each module contains 20 neurons and functions as a neural network consisting of linear units,



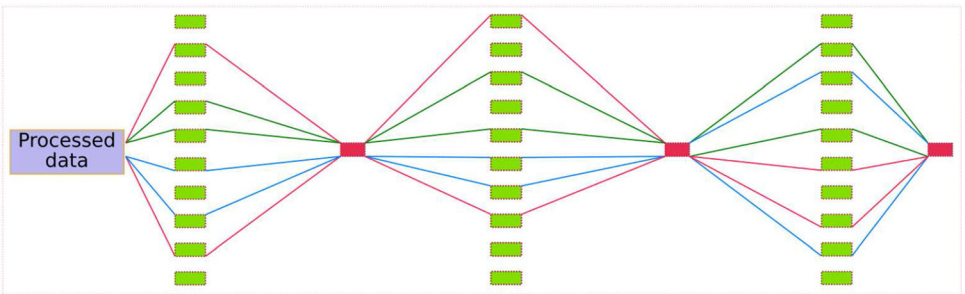
(a) Video pre-processing. All frames are initially extracted from all videos, face regions are then detected using the algorithm by Viola and Jones [42], before being resized to 64×64×3.



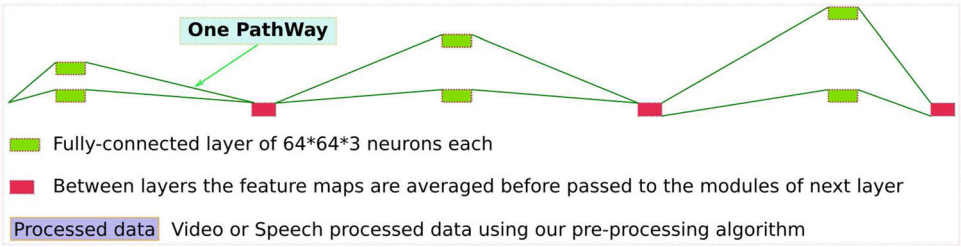
(b) Audio pre-processing. Mel-spectrograms from audio segments over all utterances are first extracted. Then, 3 channels of Mel-spectrograms with size $64 \times 64 \times 3$ ($F = 64$, $T = 64$, $C = 3$) corresponding to static, delta, and delta-delta coefficients are extracted

Fig. 2 Pre-processing steps for video and audio stream

Fig. 3 PathNet architecture



(a) PathNet architecture with 3 layers, ten modules per layer. Two modules are activated in each layer and included in a pathway, and a population of four random pathways are initialised.



(b) Description of a pathway

followed by a transfer function (rectified linear units). For each layer, the outputs of all modules in that layer are averaged before being fed into active modules of the subsequent layer. A module is active if it is shown in the path genotype and currently validated. A pathway is a path connecting active modules in all layers. A maximum of $N = 4$ distinct modules per layer are typically allowed in a pathway. The final layer is only used for the task which is being learnt and not shared with other tasks. Figure 3 illustrates our PathNet and its pathways. For the shake of simplicity in illustration, we simplify the visualisation of our PathNet's architecture in the figure by drawing only ten modules in each layer, up to two modules are activated in each layer and allowed to be included in a pathway, and a population of four random pathways are initialised.

Hyperparameters such as the number of layers, the number of pathways, make some impact on the convergence speed and accuracy of transfer learning. However, exploring all possible combinations of these hyperparameters is infeasible due to exploded number of combinations. In our experiments, we investigated these hyperparameters sequentially, starting with varying the number of layers (e.g. $L \in [3, 4]$), then the number of pathways (e.g. starting from 20). We empirically found that the transfer learning method worked stably and consistently in terms of accuracy in a small range for L (e.g. $L \in [3, 4]$), while performing best in terms of convergence speed, at our setting (i.e. 3 layers and 20 pathways).

3.3.2 Pathway evolution and transfer learning

Emotional knowledge from a source domain is managed by a pathway, that can be found by training the PathNet. Training the PathNet on a domain includes finding an optimal pathway for that domain, and, at the same time, optimising the optimal pathway's weights to fit that domain. It is hard to formulate this kind of training as a convex optimisation problem where conventional optimisation techniques, e.g. gradient descent, can be applied. To address this difficulty, genetic algorithms are often utilised. Genetic algorithms simulate the natural selection process and can be combined with other optimisation techniques (for sub-optimisation tasks). In this work, we adopt the binary tournament selection algorithm in [15] due to its proven efficacy by supporting parallel architectures, adaptivity of selection pressure to tournament size [27], and independency of the scale of the fitness function [12]. Figure 4 illustrates the selection process in the binary tournament selection algorithm. The entire algorithm and its results are described in Fig. 5. The algorithm includes three main steps as follows.

- **Step 1**: When the PathNet is trained on a training data set of the source domain, at the beginning, a population of S genotypes is randomly generated.
- **Step 2**: $K = 2$ pathways are randomly selected from the population. Each pathway is selected at a time and represented by a $N \times L$ matrix of integers in the range [1, 13]. These pathways are sequentially trained (i.e. their weights are learnt) using stochastic gradient descent for T epochs (a number of steps in each epoch equals the number of training samples divided by mini-batch size). The fitness of each pathway is calculated as the recognition accuracy (i.e. the ratio of the number of samples classified correctly and the total number of training samples) on the training set of the source domain.
- **Step 3**: Once the training of two pathways is completed, the pathway with the bad performance (called loser) is replaced by the pathway with the better performance (called winner). The loser is then mutated with a probability of $1/(N \times L)$ per each candidate of the

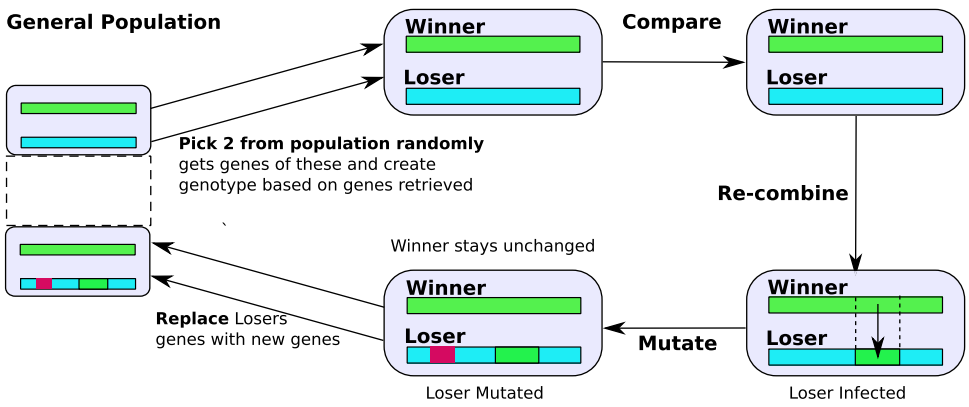


Fig. 4 Illustration of the binary tournament selection algorithm in [15]. The genotypes of the population are viewed as a pool of strings. One single cycle of the Microbial GA is operated by initially *randomly picking* two, and subsequently *compare* their fitnesses to determine *Winner*, *Loser*, and finally *recombine* where some proportion of Winner's genetic material infects the *Loser*, before *mutating* the revised version of *Loser*

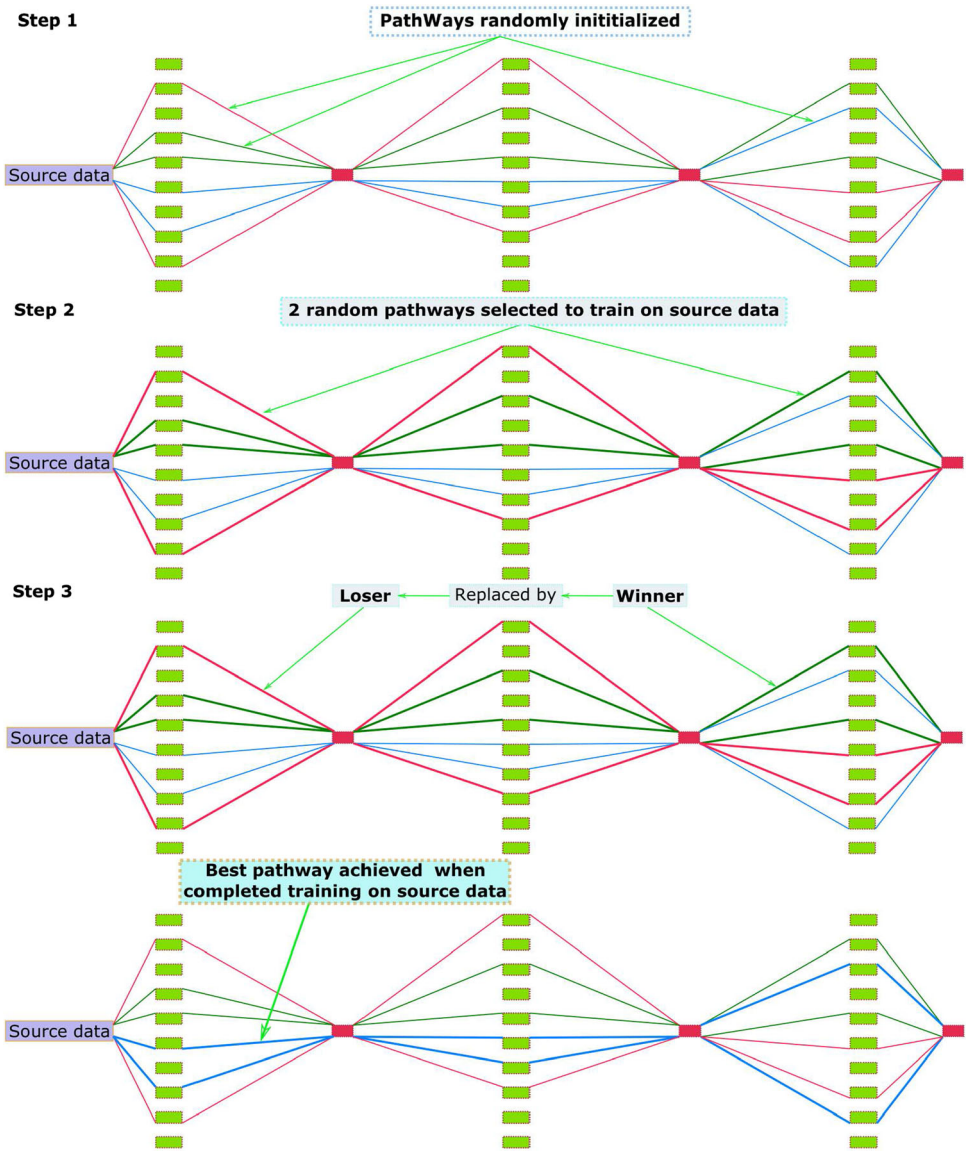


Fig. 5 Pre-training an emotion recognition system in a source domain using binary tournament selection algorithm

genotype by adding a random integer in the range $[-2, 2]$ to every element in winning pathway's matrix.

Step 2 and 3 are repeated in G iterations (i.e. generations). When the training process is completed, we achieve the best pathway for the PathNet in the source domain.

To transfer the emotional knowledge learnt from a source domain to a target domain, the best pathway learnt in the source domain is fixed, i.e. its parameters are no longer permitted to be modified. The remaining parameters, which are not shown in that pathway, are reinitialised, and are then again trained/evolved on a training set of the target domain. As shown in our experimental results, through this knowledge transferring mechanism, the emotion recognition system can learn new emotional knowledge from the target domain faster than learning from scratch or using fine-tuning approach (i.e. fine-tuning a source-domain pre-trained model in a target domain).

Training the PathNet in a target domain is done in a similar way with that in the source domain. The only difference is that the best pathway achieved in the source domain is always activated during the training process in the target domain. Once the training on the target domain is completed, we also achieve a best pathway, which best contains new knowledge learnt from the target domain. We illustrate the transfer learning process and its results in Fig. 6.

There is a concern on the number of parameters in our model as the PathNet architecture may stack up with a large number of parameters, consequently leading to a possibility that the model is prone to over-fitting. However, as explained above, in each generation, although two

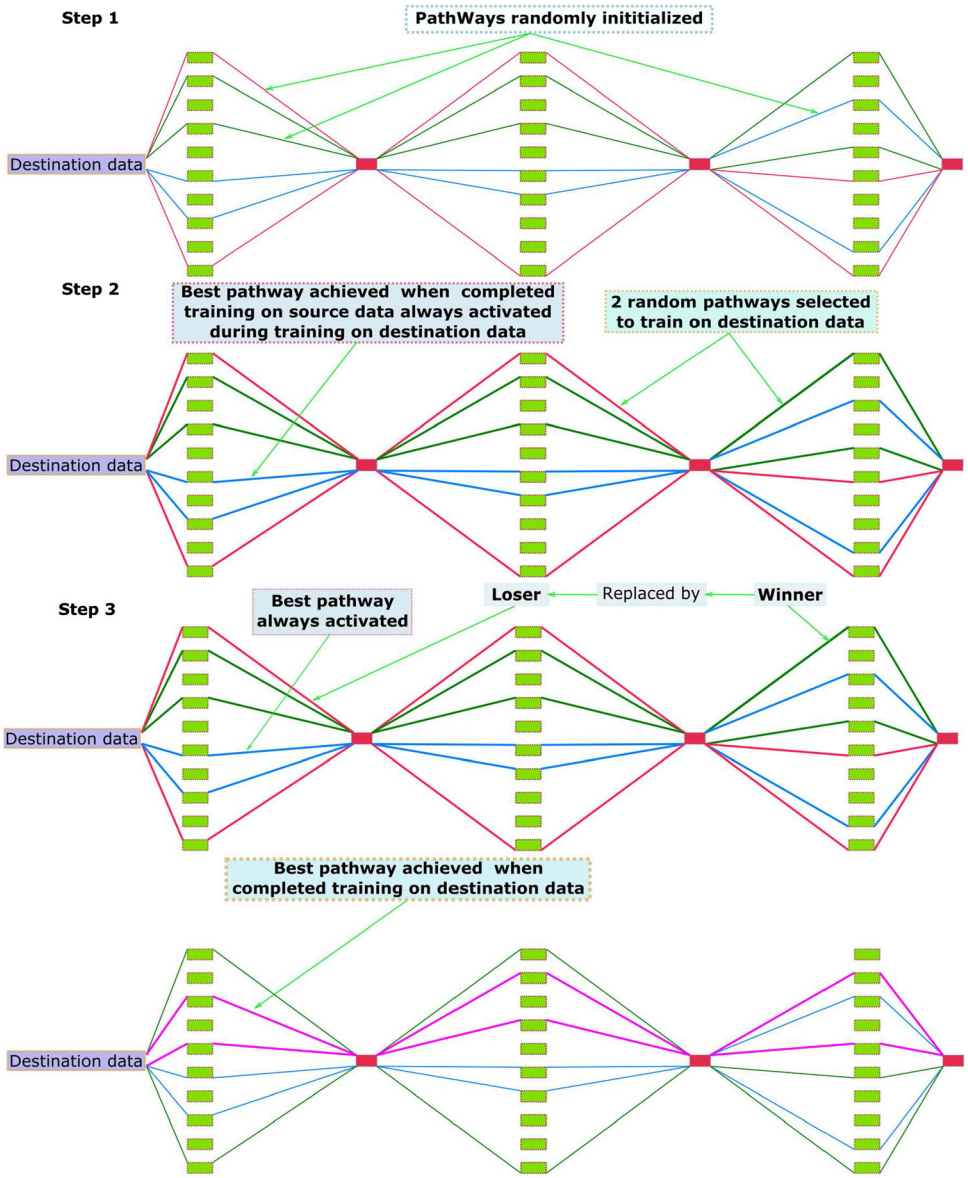


Fig. 6 Transfer learning of a pre-trained emotion recognition system in a target domain using binary tournament selection algorithm

pathways are involved in the binary tournament selection algorithm [15], only one pathway is trained at a time. In other words, the two pathways are trained sequentially.

The algorithmic complexity of the binary tournament selection algorithm relies on the population size S, the number of individuals K selected in each tournament (e.g. $K = 2$ in the binary setting), and the number of generations G. Selecting the winner requires ranking K individuals in a tournament, leading to a complexity in $\mathcal{O}(K)$. The loser is replaced by the winner and then mutated. This makes the population size unchanged, and therefore, technically the overall computational complexity of the binary tournament selection algorithm is $\mathcal{O}(S \times K \times G)$. However, we observed that the selection algorithm converged much faster in reality, as good individuals could be selected after every iteration.

4 Experiments

4.1 Data sets

We experimented our method on three benchmark data sets as follows.

- eNTERFACE [25] is an audio-visual data set. This data set includes 44 subjects and 1,293 video sequences with proportions for women and men are 23% and 77%, respectively. The subjects were asked to express 6 emotions including “anger”, “disgust”, “fear”, “happiness”, “sadness”, and “surprise”.
- SAVEE [14] is an audio-visual data set recorded by researchers (aged from 27 to 31 years) at the University of Surrey. This data set was made by 4 native male

British speakers. All of them were also required to speak and express 7 emotions including "anger", "disgust", "fear", "happiness", "sadness", "surprise", and "neutral". The data set comprises of 120 utterances per speaker, resulting in a total of 480 sentences.

- EMO-DB [3] is an acted speech corpus containing 535 emotional utterances with 7 different acted emotion classes listed as "disgust", "anger", "neutral", "sadness", "boredom", and "fear". These emotion classes were stimulated by 5 male and 5 female professional native German-speaking actors, generating 5 long and 5 short sentences German utterances used in daily communication. These actors were asked to read predefined sentences in the targeted 7 emotions. These audio files are on average around 3 s long. They were recorded using an an-echoic chamber with high-quality recording equipment at a sampling rate of 16 kHz, 16-bit resolution, and mono channel.

In our experiments, we consider eNTERFACE as a large-scale data set, and SAVEE and EMO-DB as small-scale sets. This setting fits well the purpose our study, i.e. addressing data scarcity in deep learning-based emotion recognition using transfer learning. We show that directly training our emotion recognition system on these small data sets results in poor performance. Since those data sets contain both audio and visual emotional information, we also use them for comparison of facial expression-based emotion recognition and speech emotion recognition.

4.2 Implementation details

To validate the ability of our transfer learning method, we initially trained our PathNet on a source data set (e.g. visual/audio eNTERFACE, visual/audio SAVEE) and then transferred it to a target data set. In each experiment, a population of 20 pathways were randomly generated on a source/target data set. The pathway evolution algorithm in Sect. 3.3.2 was applied with 200 generations. For each generation, two pathways were randomly picked and trained. A pathway was trained using stochastic gradient descent with a learning rate of 0.02, mini-batch size of 64, and T epochs; T was set to the number of training samples divided by mini-batch size. Recall that, in transfer learning, the best pathway, determined on a source data set, was fixed and remained active, and the rest of the PathNet's parameters were re-initialised for searching for a new best pathway on a target data set.

To evaluate our method and compare it with existing ones, we applied k-fold cross-validation with $k = 5$ on experimented data sets (i.e. each data set was randomly partitioned into k equal parts, one of which was used for validation while the other ones were used for training). We adopted emotion recognition accuracy and confusion matrix, which reflects the accuracy of each method on every emotion class, as performance metrics.

4.3 Results

We conducted various sets of experiments using our proposed system to examine how well facial expression-based emotion recognition and speech emotion recognition can be improved when emotional knowledge is transferred between different domains.

4.3.1 Facial expression-based emotion recognition

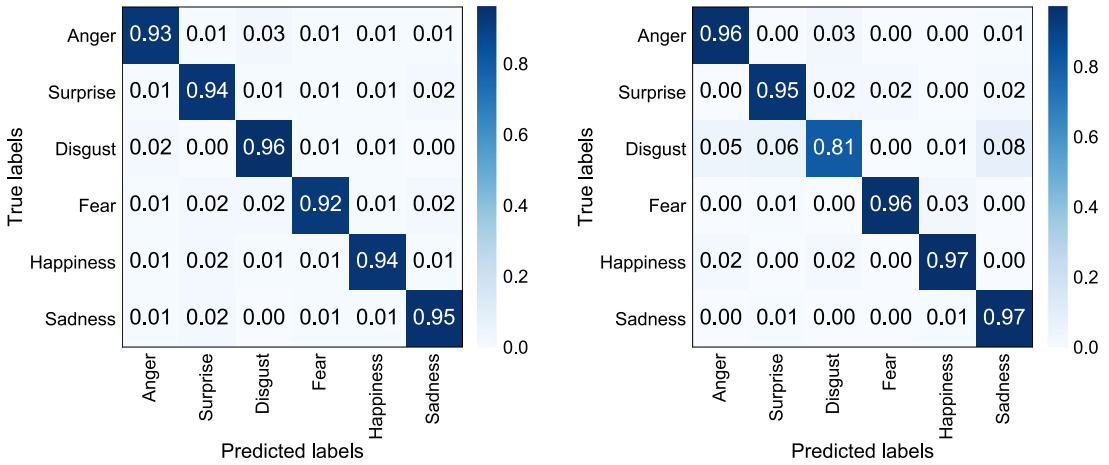
We first investigated our proposed meta-transfer learning in facial expression-based emotion recognition where emotional knowledge was transferred from one visual emotion data set (called source data set) to another visual emotion data set (called target data set). Specifically, in the first setting, we trained from scratch our PathNet on visual eNTERFACE (resulting in a model V_eNTER) and then evolved it on visual SAVEE (resulting in a model V_eNTER→ SAV). Similarly, in the second setting, we created a model V_SAV by training our PathNet on visual SAVEE from scratch and then transferred V_SAV on visual eNTERFACE to make a model V_SAV→ eNTER. Since SAVEE data set consists of an additional type of emotion ("neutral") compared to eNTERFACE data set, to make transfer learning compatible to these two data sets, only emotion classes commonly shared in both the datasets were considered. Those common emotion classes include "anger", "surprise", "disgust", "fear", "happiness", and "sadness".

We report the recognition accuracy and confusion matrices of our meta-transfer learning method in transferring facial expression-based emotional knowledge between eNTERFACE and SAVEE data set in both settings in Table 1 and Fig. 7. As shown in the results, in general, our transfer learning method works well across both settings with an overall accuracy of 94%. Although transfer learning from visual SAVEE to visual eNTERFACE

Table 1 Results of transfer learning in facial expression-based emotion recognition on visual eNTERFACE and visual SAVEE

Model	Ang	Sur	Dis	Fea	Hap	Sad	Overall
$V_{eNTER \rightarrow SAV}$	0.93	0.94	0.96	0.92	0.94	0.95	0.94
$V_{SAV \rightarrow eNTER}$	0.96	0.95	0.81	0.96	0.97	0.97	0.94

V_eNTER→ SAV denotes the PathNet model initially trained on visual eNTERFACE and then transferred to visual SAVEE. V_SAV→ eNTER denotes the PathNet model initially trained on visual SAVEE and then transferred to visual eNTERFACE



(a) Confusion matrix of facial expression-based emotion recognition when transferring emotional knowledge from visual eNTERFACE to visual SAVEE

(b) Confusion matrix of facial expression-based emotion recognition when transferring emotional knowledge from visual SAVEE to visual eNTERFACE

Fig. 7 Confusion matrices of our proposed meta-transfer learning method applied to facial expression-based emotion recognition

achieves impressive results on several emotion classes (e.g. "happiness", "sadness"), transfer learning from visual eNTERFACE to visual SAVEE performs more consistently across all emotion classes.

To show the effectiveness of our proposed meta-transfer learning method, we compared it with other baselines. Each baseline follows conventional training/testing setting on a single domain, i.e. to train and test the emotion recognition system on a data set. Specifically, we evaluated our emotion recognition system when it was trained and tested on either visual eNTERFACE or visual SAVEE. We denote the PathNet model trained on visual eNTERFACE as V_eNTER and the PathNet model trained on visual SAVEE as V_SAV. We also compared our approach with the work in [46] that applied fine-tuning for transfer learning. Table 2 presents the results of this experiment. As illustrated in Table 2, our meta-transfer learning method achieves an overall accuracy of 94% for facial expression-based emotion recognition in both settings (i.e. transfer learning from visual eNTERFACE to visual SAVEE and vice versa). This is the highest performance on both data sets. The results show a significant improvement gained by our transfer learning method (up to 6%) compared with training a pre-trained model from scratch. Our transfer learning also shows its superiority (up to 9% higher) over the commonly used fine-tuning approach in emotion recognition [46].

4.3.2 Speech emotion recognition

Next we evaluated our proposed meta-transfer learning in speech emotion recognition. Similarly to facial expression-based emotion recognition, we applied our method in

Table 2 Comparison of our meta-transfer learning method with other baselines and methods in facial expression-based emotion recognition

Method	Overall Accuracy
(a) Evaluation results on visual SAVEE	
Fine-tuning [46] (trained on visual eNTERFACE)	0.85
V_{SAV}	0.89
$V_{eNTER \rightarrow SAV}$	**0.94**
(b) Evaluation results on visual eNTERFACE	
Fine-tuning [46] (trained on visual SAVEE)	0.88
V_{eNTER}	0.88
$V_{SAV \rightarrow eNTER}$	**0.94**

V_eNTER and V_SAV denote the baselines trained from scratch on visual eNTERFACE and on visual SAVEE, respectively. V_eNTER→ SAV and V_SAV→ eNTER denote the PathNet models initially trained on visual eNTERFACE and transferred to visual SAVEE, and initially trained on visual SAVEE and transferred to visual eNTERFACE, respectively. Best performances are highlighted

transferring emotional knowledge in speech signals from a source data set to a target data set.

In this experiment, we chose audio eNTERFACE as the source data set and audio SAVEE as the target data set. However, unlike the case study of facial expression-based emotion recognition, we did not conduct transfer learning from audio SAVEE to audio eNTERFACE. This is because audio eNTERFACE is much larger than audio SAVEE and, as shown in our empirical results, transfer learning of a model from a smaller-scale data set to a larger-scale one does not help to improve the model on the larger-scale data set. In addition, this transfer learning strategy may even perform worse than directly training the model from scratch on the larger-scale data set. For instance, we observed a decrease of 4% and 6% in the overall accuracy when transfer learning was applied from audio SAVEE to audio eNTERFACE, compared with direct use of A_SAV and with A_eNTER (trained from scratch) on audio eNTERFACE.

We also experimented our meta-transfer learning when transferring emotion knowledge from multiple audio emotion domains (audio eNTERFACE and audio SAVEE) to one audio emotion domain (EMO-DB).

We present the recognition accuracy and confusion matrices of our meta-transfer learning in speech emotion recognition in Table 3 and Fig. 8. In general, compared with facial expression-based emotion recognition on visual SAVEE, speech emotion recognition on audio SAVEE is more challenging, especially in recognising "disgust" emotion class. In contrast, for the setting of transfer learning from audio eNTERFACE and audio SAVEE to EMO-DB, the emotion recognition system performs best at "disgust" with perfect accuracy (100%), while demonstrating relatively high performance on the other emotion classes, e.g. the lowest accuracy is 94%, that is the recognition accuracy of "fear" and "happiness".

We visualise the transfer learning process from audio eNTERFACE to audio SAVEE in Fig. 9 (PathNet initially trained on audio eNTERFACE) and Fig. 10 (PathNet transferred to audio SAVEE).

Table 3 Results of meta-transfer learning in speech emotion recognition

Method	Ang	Sur	Dis	Fea	Hap	Sad	Overall
$A_{eNTER \rightarrow SAV}$	0.73	0.70	0.56	0.72	0.80	0.77	0.71
$A_{eNTER+SAV \rightarrow EMO}$	0.99	0.99	1.0	0.94	0.94	0.96	0.97

A_eNTER→ SAV denotes the PathNet model initially trained on audio eNTERFACE and then transferred to audio SAVEE. A_eNTER+SAV→ EMO denotes the PathNet model initially trained on both audio eNTERFACE and audio SAVEE, and then transferred to EMO-DB. Note that EMO-DB contains only speech data

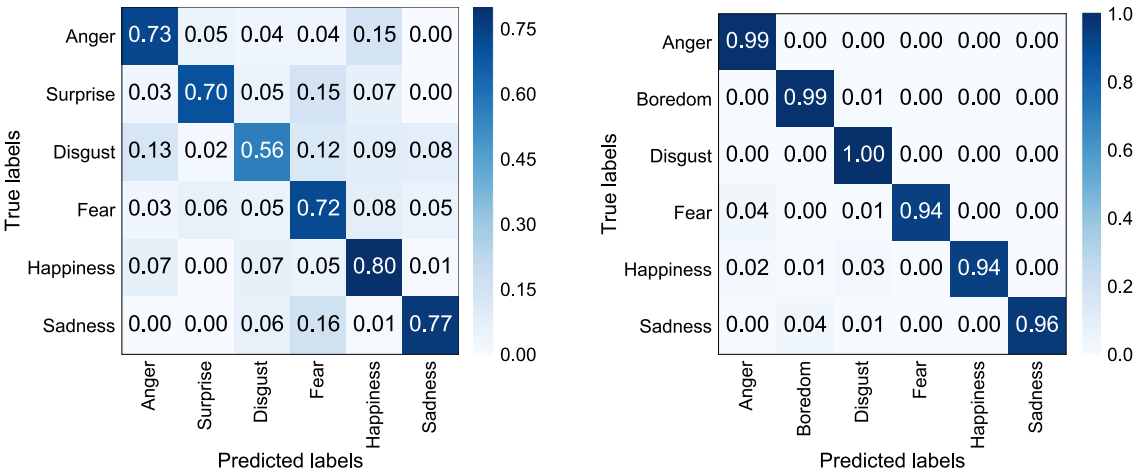
Our proposed meta-transfer learning also significantly outperforms conventional training/testing settings (i.e. training and testing the emotion recognition system on the same domain). To prove this, we trained and tested the emotion recognition system on audio SAVEE. We refer this baseline to as A_SAV. We made another baseline called A_eNTER+SAV→ EMO by training the emotion recognition system on both audio eNTERFACE and audio SAVEE, and then testing it on EMO-DB data set. For each training/testing setting, we also compared our approach with the fine-tuning approach in [46]. To the best of our knowledge, the work in [46] is the current best baseline in speech emotion recognition on experimented data sets. We present comparison results in Table 4.

The results in Table 4 indicate several insights. First, compared with the fine-tuning approach in [46], which is also the current state of the art in emotion speech recognition, our meta-transfer learning makes a significant improvement (up to 16% of overall accuracy). This improvement is consistent over different training/testing settings. Second, transfer learning from a larger-scale and possibly richer source data set to a smaller-scale target data set (e.g. from audio eNTERFACE to audio SAVEE) enhances the speech emotion recognition system on the target data set (up to 8%), compared with training the emotion recognition system from scratch on the target data set.

The reason that our transfer learning approach significantly outperforms the fine-tuning approach in [46] probably because emotional knowledge presented in the best pathway achieved on a source data set can be reused to initialise a new pathway on a target data set. In addition, this best pathway is always involved in the learning phase of PathNet on any target data set. However, when PathNet is initially trained on an emotion data set from scratch, emotional knowledge is randomly initialised and learnt with only one set of pathways to fit with that emotion data set.

5 Discussion and conclusion

Progress in emotion recognition research has been hindered by lack of large amount of labelled emotion data. To overcome this issue, existing studies have explored the use of transfer learning to enhance emotion recognition methods from various but limited emotion data sources. However, there are unsolved issues in the current transfer learning approach such as discarding learnt knowledge, retaining irrelevant knowledge. In this paper, we propose an alternative transfer learning technique based on PathNet, a network architecture that uses pathways to learn knowledge from different tasks/domains. The PathNet

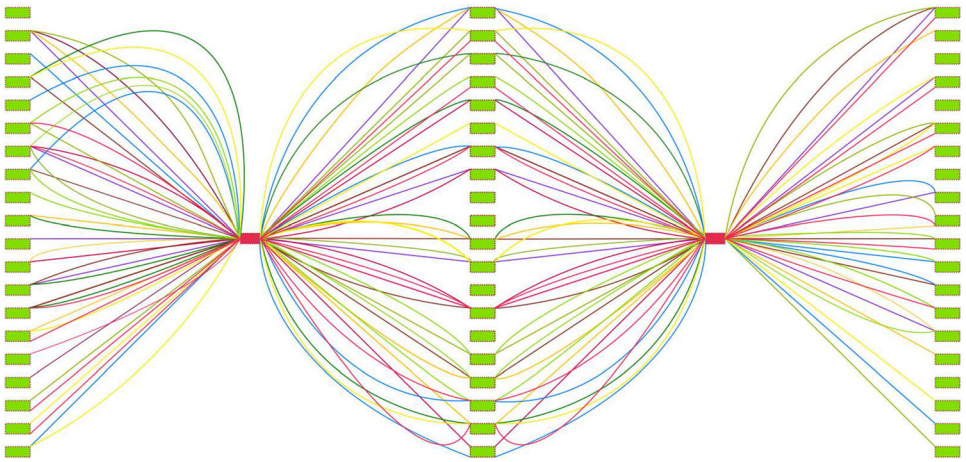


(a) Confusion matrix of speech emotion recognition when transferring emotional knowledge from audio eNTERFACE to audio SAVEE

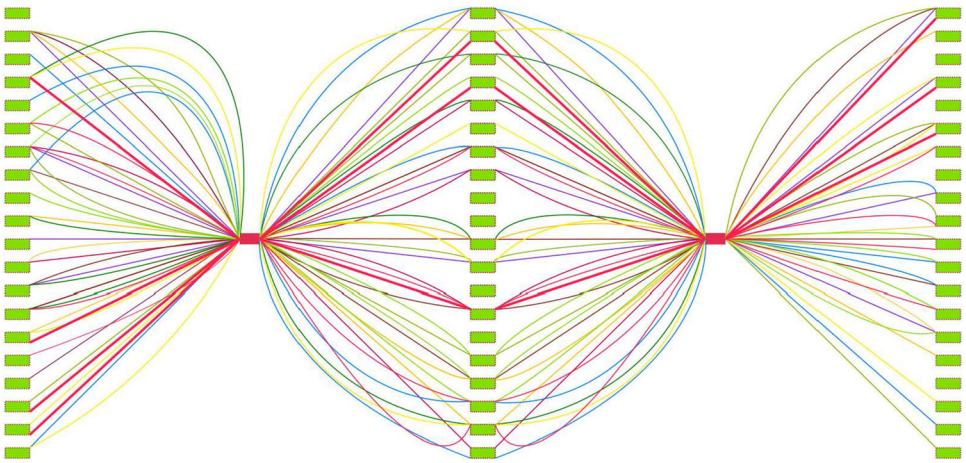
(b) Confusion matrix of speech emotion recognition when transferring emotional knowledge from audio eNTERFACE and audio SAVEE to EMODB

Fig. 8 Illustrates confusion matrix of our proposed system evaluated on audio SAVEE and audio EMO-DB

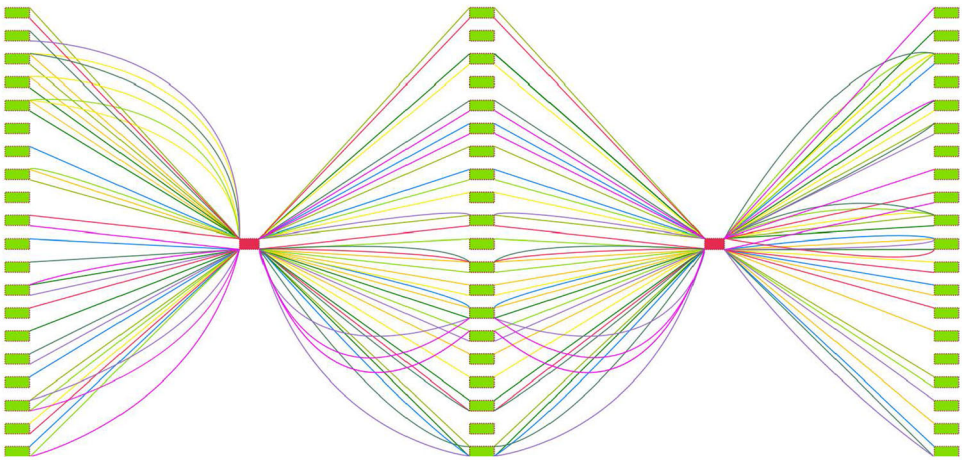
Fig. 9 Visualisation of our PathNet trained from scratch on audio eNTERFACE (source data set)



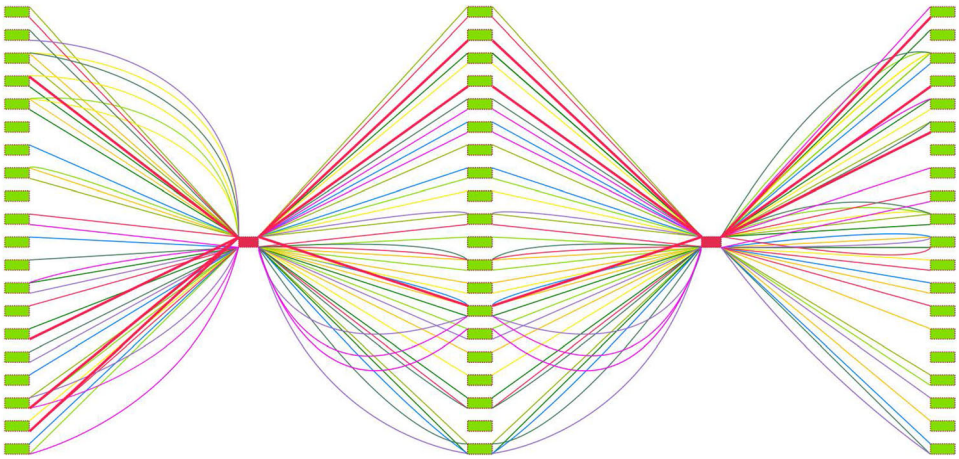
(a) A population of 20 pathways are randomly initialised on audio eNTERFACE.



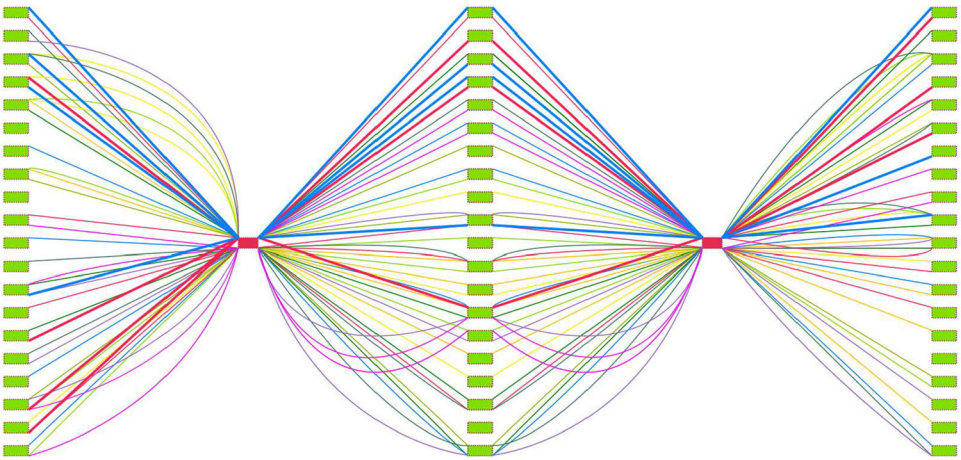
(b) Optimal pathways (highlighted in thick red lines) and their parameters are learnt on audio eNTERFACE.



(a) A new population of pathways are generated on audio SAVEE.



(b) Transferred pathways (in red) are always activated but their parameters are always fixed during transfer learning on audio SAVEE.



(c) New optimal pathways (highlighted in thick blue lines) and their parameters are learnt on audio SAVEE.

Fig. 10 Visualisation of our PathNet trained on audio eNTERFACE (source data set) and transferred to audio SAVEE (target data set)

architecture can learn to discover which pathways to reuse for new tasks/domains, leading to successfully addressing the aforementioned challenges. To validate our proposed method, we conducted extensive experiments including transfer learning from one emotion data set to another emotion data set, in both visual and audio modality, and transfer learning from multiple emotion data sets to one emotion data set. Experimental results on benchmark data

Table 4 Comparison of our meta-transfer learning method with other baselines and methods in speech emotion recognition

Method	Overall accuracy
(a) Evaluation results on audio SAVEE	
Fine-tuning [46] (trained on audio eNTERFACE)	0.69
A_{SAV}	0.81
$A_{eNTER \rightarrow SAV}$	**0.85**
(b) Evaluation results on EMO-DB	
Fine-tuning [46] (trained on audio eNTERFACE+SAVEE)	0.83
A_{EMO}	0.89
$A_{eNTER+SAV \rightarrow EMO}$	**0.97**

A_SAV and A_EMO denote the baselines trained from scratch on audio SAVEE and on EMO-DB, respectively. A_eNTER→ SAV and A_eNTER+SAV→ EMO denote the PathNet models initially trained on audio eNTERFACE and transferred to audio SAVEE, and initially trained on the both audio eNTERFACE and SAVEE and transferred to EMO-DB, respectively. Best performances are highlighted

sets verified the effectiveness of our method in transferring emotional knowledge between different domains, and its superiority over the current transfer learning technique in emotion recognition.

In this paper, we investigated transfer learning between different domains/data sets within the same datatype (i.e. visual-to-visual or audio-to-audio data sets). It is also worthwhile to explore transfer learning across different datatypes. We consider this as our future work.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data availability The data sets generated during and/or analysed during the current study are available in the eNTERFACE [25], SAVEE [14], and EMO-DB [3] repository. eNTERFACE [25]: http://www.enterface.net/enterface05/docs/results/databases/project2_database.zip. SAVEE [14]: https://www.kaggle.com/datasets/barelydedicated/savee-database. EMO-DB [3]: https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb.

Declarations

Conflict of interest We have no conflict of interests to declare.

References

1. Abbasnejad I, Sridharan S, Nguyen D et al (2017) Using synthetic data to improve facial expression analysis with 3D convolutional networks. In: IEEE international conference on computer vision workshops, pp 1609–1618
2. Badshah AM, Ahmad J, Rahim N et al (2017) Speech emotion recognition from spectrograms with deep convolutional neural network. In: International conference on platform technology and service, pp 1–5
3. Burkhardt F, Paeschke A, Rolfes M et al (2005) A database of German emotional speech. In: INTERSPEECH, pp 1517–1520
4. Cai X, Yuan J, Zheng R et al (2021) Speech emotion recognition with multi-task learning. In: INTERSPEECH, pp 4508–4512
5. Chang J, Scherer S (2017) Learning representations of emotional speech with deep convolutional generative adversarial networks. In: IEEE international conference on acoustics, speech and signal processing, pp 2746–2750
6. Deng J, Frühholz S, Zhang Z et al (2017) Recognizing emotions from whispered speech based on acoustic feature transfer learning. IEEE Access 5:5235–5246
7. Deng J, Xu X, Zhang Z et al (2017) Universum autoencoder-based domain adaptation for speech emotion recognition. IEEE Signal Process Lett 24(4):500–504
8. Deng J, Guo J, Zhou Y et al (2019) Retinaface: Single-stage dense face localisation in the wild. CoRR arXiv:abs/1905.00641
9. Fan Y, Lu X, Li D et al (2016) Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In: ACM international conference on multimodal interaction, pp 445–450
10. Fernando C, Banarse D, Blundell C et al (2017) Pathnet: evolution channels gradient descent in super neural networks. CoRR arXiv:abs/1701.08734
11. Gideon J, Khorram S, Aldeneh Z et al (2017) Progressive neural networks for transfer learning in emotion recognition. In: INTERSPEECH, pp 1098–1102
12. Goldberg DE, Deb K (1990) A comparative analysis of selection schemes used in genetic algorithms. In: Foundations of genetic algorithms, pp 69–93
13. Guo Y, Shi H, Kumar A et al (2019) Spottune: transfer learning through adaptive fine-tuning. In: IEEE/CVF conference on computer vision and pattern recognition, pp 4800–4809
14. Haq S, Jackson P (2009) Speaker-dependent audio-visual emotion recognition. In: International conference on auditory-visual speech processing, pp 53–58

15. Harvey I (2009) The microbial genetic algorithm. In: European conference on artificial life, pp 126–133
16. Hasani B, Mahoor MH (2017) Facial expression recognition using enhanced deep 3D convolutional neural networks. In: IEEE international conference on computer vision and pattern recognition workshops, pp 2278–2288
17. Hermansky H, Morgan N, Bayya A et al (1992) RASTA-PLP speech analysis technique. In: IEEE international conference on acoustics, speech and signal processing, pp 121–124
18. Kaya H, Gürpınar F, Salah AA (2017) Video-based emotion recognition in the wild using deep transfer learning and score fusion. Image Vis Comput 65:66–75
19. Kerkeni L, Serrestou Y, Raoof K et al (2019) Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO. Speech Commun 114:22–35
20. Kim J, Englebienne G, Truong KP et al (2017) Deep temporal models using identity skip-connections for speech emotion recognition. In: ACM international conference on multimedia, pp 1006–1013
21. Kim J, Englebienne G, Truong KP et al (2017) Towards speech emotion recognition 'in the wild' using aggregated corpora and deep multi-task learning. In: INTERSPEECH, pp 1113–1117
22. Kim J, Truong K, Englebienne G et al (2017) Learning spectro-temporal features with 3DCNNs for speech emotion recognition. In: International conference on affective computing and intelligent interaction, pp 383–388
23. Latif S, Rana R, Younis S et al (2018) Cross corpus speech emotion classification: an effective transfer learning technique. CoRR arXiv:abs/1801.06353
24. Lee SW, Kim JH, Jun J et al (2017) Overcoming catastrophic forgetting by incremental moment matching. In: Advances in neural information processing systems, pp 4655–4665
25. Martin O, Kotsia I, Macq B et al (2006) The eNTERFACE' 05 audio-visual emotion database. In: International conference on data engineering workshops, pp 1–8
26. Mellourk W, Handouzi W (2020) Facial emotion recognition using deep learning: review and insights. Procedia Comput Sci 175:689–694
27. Miller BL, Goldberg DE (1996) Genetic algorithms, selection schemes, and the varying effects of noise. Evolut Comput 4(2):113–131
28. Ng HW, Nguyen VD, Vonikakis V et al (2015) Deep learning for emotion recognition on small datasets using transfer learning. In: ACM international conference on multimodal interaction, pp 443–449
29. Nguyen D, Nguyen K, Sridharan S et al (2017) Deep spatio-temporal features for multimodal emotion recognition. In: IEEE winter conference on applications of computer vision, pp 1215–1223
30. Nguyen D, Nguyen K, Sridharan S et al (2018) Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition. Comput Vis Image Underst 174:33–42
31. Nguyen D, Nguyen DT, Zeng R et al (2022) Deep auto-encoders with sequential learning for multimodal dimensional emotion recognition. IEEE Trans Multimedia 24:1313–1324
32. Pao TL, Chen YT, Yeh JH et al (2006) Mandarin emotional speech recognition based on SVM and NN. In: International conference on pattern recognition, pp 1096–1100
33. Peng Z, Lu Y, Pan S et al (2021) Efficient speech emotion recognition using multi-scale CNN and attention. In: IEEE international conference on acoustics, speech and signal processing, pp 3020–3024
34. Rabiner LR, Sambur MR (1975) An algorithm for determining the endpoints of isolated utterances. Bell Syst Tech J 54(2):297–315
35. Rusu AA, Rabinowitz NC, Desjardins G et al (2016) Progressive neural networks. CoRR arXiv:abs/1606.04671
36. Sahu S, Gupta R, Sivaraman G et al (2017) Adversarial auto-encoders for speech based emotion recognition. In: INTERSPEECH, pp 1243–1247
37. Sermanet P, Eigen D, Zhang X et al (2014) Overfeat: Integrated recognition, localization and detection using convolutional networks. In: International Conference on Learning Representations
38. Shen P, Changjun Z, Chen X (2011) Automatic speech emotion recognition using support vector machine. In: International conference on electronic mechanical engineering and information technology, pp 621–625
39. Sun Q, Liu Y, Chua T et al (2019) Meta-transfer learning for few-shot learning. In: IEEE/CVF conference on computer vision and pattern recognition, pp 403–412
40. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
41. Ververidis D, Kotropoulos C, Pitas I (2004) Automatic emotional speech classification. In: IEEE international conference on acoustics, speech, and signal processing, pp 593–596
42. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154
43. Xiao Z, Dellandrea E, Dou W et al (2005) Features extraction and selection for emotional speech classification. In: IEEE conference on advanced video and signal based surveillance, pp 411–416
44. Yoon S, Byun S, Dey S et al (2019) Speech emotion recognition using multi-hop attention mechanism. In: IEEE international conference on acoustics, speech and signal processing, pp 2822–2826
45. Zhang K, Zhang Z, Li Z et al (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503
46. Zhang S, Huang T, Gao W (2018) Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. IEEE Trans Multimedia 20(6):1576–1590
47. Zhang S, Huang T, Gao W et al (2018) Learning affective features with a hybrid deep model for audio-visual emotion recognition. IEEE Trans Circuits Syst Video Technol 28(10):3030–3043
48. Zhu W, Li X (2022) Speech emotion recognition with global-aware fusion on multi-scale feature representation. In: IEEE international conference on acoustics, speech and signal processing, pp 6437–6441
49. Zou H, Si Y, Chen C et al (2022) Speech emotion recognition with co-attention based multi-level acoustic information. In: IEEE international conference on acoustics, speech and signal processing, pp 7367–7371