Treinamento Dinâmico de Conscientização da Dificuldade para Previsão de Emoção Contínua

Zixing Zhang, Jing Han, Eduardo Coutinho e Bjorn Schuller ¨

Resumo—A previsão de emoções contínuas no tempo tornou-se uma tarefa cada vez mais atraente no aprendizado de máquina. Esforços consideráveis foram feitos para melhorar o desempenho desses sistemas. No entanto, o foco principal tem sido o desenvolvimento de modelos mais sofisticados e a incorporação de diferentes modalidades expressivas (por exemplo, fala, face e fisiologia). Neste artigo, motivados pelo benefício da consciência de dificuldade em um procedimento de aprendizagem humana, propomos uma nova estrutura de aprendizado de máquina, ou seja, Treinamento de Conscientização de Dificuldade Dinâmica (DDAT), que lança uma nova luz sobre a pesquisa - explorando diretamente as dificuldades na aprendizagem para impulsionar o processo de aprendizado de máquina. A estrutura DDAT consiste em duas etapas: recuperação de informações e exploração de informações. Na primeira etapa, fazemos uso do erro de reconstrução de características de entrada ou da incerteza de anotação para estimar a dificuldade de aprender informações específicas. O nível de dificuldade obtido é então usado em conjunto com os recursos originais para atualizar a entrada do modelo em um segundo estágio de aprendizado com a expectativa de que o modelo possa aprender a focar nas regiões de alta dificuldade do processo de aprendizado. Realizamos extensos experimentos em um banco de dados de referência (RECOLA) para avaliar a eficácia da estrutura proposta. Os resultados experimentais mostram que nossa abordagem supera as linhas de base relacionadas, bem como outros sistemas bem estabelecidos de previsão de emoções contínuas no tempo, o que sugere que integrar dinamicamente as informações de dificuldade para redes neurais pode ajudar a melhorar o processo de aprendizado.

Termos de indexação — predição de emoções, aprendizado de percepção de dificuldade, aprendizado dinâmico

I. INTRODUÇÃO

Os sistemas de previsão de emoções contínuas no tempo receberam amplo interesse na comunidade de aprendizado de máquina (ML) na última década [1]–[3]. Uma das principais razões para esse interesse é o fato de que as previsões de emoções contínuas no tempo podem analisar estados afetivos sutis e complexos de humanos ao longo do tempo e desempenham um papel central em agentes conversacionais inteligentes que visam alcançar uma interação natural e intuitiva entre humanos e máquinas [2], [4]–[7]. Grandes esforços têm sido feitos neste campo, e a maioria deles geralmente pode ser classificada em duas vertentes. Uma vertente se concentra principalmente em projetar ou implementar cada vez mais sofisticados e modelos de previsão robustos, como redes neurais recorrentes (RNNs) baseadas em memória de curto prazo (LSTM) [1], [8], redes neurais convolucionais (CNNs) [9]–[13] e aprendizado de ponta a ponta quadros [14]. Outra vertente concentra-se principalmente na integração de múltiplas modalidades (por exemplo, , áudio e vídeo) e técnicas de modelagem [16], [15],

Além desses estudos, outra pesquisa descobriu recentemente que os dados de treinamento emocional podem ser aprendidos de forma prática em diferentes graus [17], [18]. Ou seja, alguns dados podem ser facilmente aprendidos dado um modelo específico, enquanto alguns dados são relativamente difíceis. Diante disso, algumas abordagens promissoras foram propostas em aprendizado de máquina para otimizar o procedimento de aprendizado. Por exemplo, a abordagem mais convencional está associada ao boosting [19], [20], que atualiza dinamicamente os pesos das amostras que são difíceis de reconhecer ou mesmo falsamente reconhecidas. Adicionalmente, uma abordagem mais recente e promissora refere-se à aprendizagem curricular, que foi introduzida pela primeira vez em [21]. O aprendizado curricular apresenta os dados de fácil a difícil durante o processo de treinamento para que o modelo possa evitar ser pego em mínimos locais na presença de critérios de treinamento não convexos. A aprendizagem curricular tornou-se ainda mais popular com o avanço da aprendizagem profunda. Para a previsão de emoções, um punhado de estudos relacionados foram relatados muito recentemente [18], [22], [23], que mostraram a eficiência do aprendizado curricular.

No entanto, uma das principais desvantagens dessas abordagens é sua falta de facilidade para tarefas de reconhecimento de padrões baseadas em sequência, como a que estamos enfrentando. Ou seja, no processo de aprendizagem, as amostras, apresentadas ou não dentro de uma sequência, são consideradas de forma individual e independente. A informação de contexto ignorada, no entanto, de fato desempenha um papel vital no reconhecimento de padrão baseado em sequência [24]. Para este fim, propomos uma nova estrutura de aprendizagem, Dynamic Difficulty Awareness Training (DDAT), para previsão de emoção contínua no tempo neste artigo. Em contraste com as abordagens anteriores, como o aumento mencionado e o aprendizado curricular, o DDAT proposto pode ser bem integrado a modelos sensíveis ao contexto convencionais (por exemplo, LSTM-RNNs), permitindo que os modelos explorem as informações do contexto. Até onde sabemos, este é o primeiro esforço para explorar a informação de dificuldade no reconhecimento de padrões baseado em sequência, como o presente caso de previsão de emoção contínua no tempo.

A suposição subjacente do DDAT é que um modelo é capaz de oferecer melhor desempenho se deixarmos explicitamente o modelo saber a dificuldade de aprendizado das amostras ao longo do tempo. Essa suposição está de acordo com a descoberta de que os humanos normalmente prestam mais atenção às tarefas que são inerentemente

Z. Zhang está com GLAM - o Grupo de Linguagem, Áudio e Música, Imperial College London (Reino Unido). E-mail: zixing.zhang@imperial.ac.uk.

J. Han (autor correspondente) é titular da ZD.B Chair of Embedded Intel ligence for Health Care and Wellbeing, University of Augsburg (Alemanha). E-mail: jing.han@informatik.uni-augsburg.de.

E. Coutinho trabalha no Departamento de Música da Universidade de Liverpool (Reino Unido) e na ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg (Alemanha). E-mail: e.coutinho@liverpool.ac.uk.

B. Schuller está com o GLAM – o Grupo de Linguagem, Áudio e Música, Imperial College London (Reino Unido) e a ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg (Alemanha). E-mail: bjoern.schuller@imperial.ac.uk.

difícil para melhor desempenho [25], [26].

Para implementar o DDAT, consideramos duas estratégias, ou seja, utilizar o Erro de Reconstrução (RE) dos dados de entrada ou o nível de incerteza de percepção (PU) das emoções como indicadores dinâmicos da dificuldade para conduzir o processo de aprendizagem.
Em seguida, integramos o indicador de dificuldade com dados originais para aprendizado posterior, de forma que dote os modelos com uma consciência de aprendizado de dificuldade. Este processo também é parcialmente inspirado nas técnicas de awareness propostas para o reconhecimento robusto de fala [27], [28], onde os tipos de ruído são considerados informações auxiliares para modelagem acústica.

Em ML, RE normalmente serve como uma função objetiva de um autoencoder (AE) ao extrair representações de alto nível.
Um AE bem projetado é considerado para reconstruir bem a entrada de suas representações de alto nível aprendidas [29]. Recentemente, o RE também foi explorado para tarefas, como detecção de anomalias [30], [31] e classificação [32]. Para a detecção de anomalias, um AE é previamente treinado em amostras normais para servir como um novo detector de eventos. Ao alimentar uma nova amostra no AE, o RE obtido em comparação com um limite predefinido decide se é anormal [30], [31].
Para classificação, vários AEs específicos da classe são pré-treinados separadamente. Ao alimentar uma amostra desconhecida nesses AEs simultaneamente, os valores do RE correspondente são interpretados como indicadores de associação de classe [32]. Notavelmente, todos esses trabalhos levantam a hipótese de que dados com o mesmo rótulo têm distribuições de dados semelhantes.
Ou seja, os dados incompatíveis resultam potencialmente em REs maiores do que os dados correspondentes. Isso nos motiva a empregar o RE como um índice de dificuldade de aprendizagem porque é bem conhecido em ML que dados incompatíveis promovem severamente a complexidade da modelagem [33].

Além disso, PU é um termo empregado em tarefas subjetivas de reconhecimento de padrões para se referir ao nível de discordância entre anotadores ao calcular um padrão-ouro em um processo de anotação [34]. Para predição de emoções, tem sido frequentemente determinado que o PU tem uma alta correlação com a dificuldade de aprendizagem de um modelo de reconhecimento. Por exemplo, o trabalho relatado em [35] e [36] descobriu que os sistemas de predição de emoções funcionam melhor em regiões de baixa incerteza do que em regiões de alta incerteza. Da mesma forma, os achados em [17] mostraram que a eliminação das amostras rotuladas com alta incerteza do conjunto de treinamento leva a um melhor modelo de predição de emoções. Essa constatação nos provoca a utilizar a UP como mais um índice de dificuldade de aprendizagem. Também é importante notar que o princípio da estratégia baseada em PU restringe sua aplicação a tarefas subjetivas de reconhecimento de padrões. Apesar de o conceito de 'incerteza' ter sido empregado em trabalhos anteriores de previsão de emoções, ele foi calculado entre múltiplas previsões de sistemas variáveis [37], o que difere significativamente da definição de PU neste artigo, ou meramente utilizado para aprendizagem multitarefa [38] (cf. Secção II).

Motivados pela análise acima e seguindo nosso trabalho experimental anterior [39], onde apenas o RE foi investigado para predição de emoções na fala, demonstramos neste artigo que a estrutura DDAT proposta pode auxiliar os modelos ML na detecção de 'momentos' no processo de aprendizagem que são de maior dificuldade no contexto da predição audiovisual de emoções contínuas no tempo. Mais especificamente, o

As contribuições do presente artigo incluem o seguinte: (i) propor uma nova estrutura que explora o conhecimento sobre a dificuldade de aprendizagem das amostras durante o processo de aprendizagem para predição de emoções contínuas no tempo; (ii) apresentar e analisar duas estratégias específicas (ou seja, baseadas em RE ou PU) para implementar esta estrutura; (iii) apresentar uma abordagem de ajuste dinâmico para ajustar ainda mais dinamicamente as previsões; e (iv) avaliar de forma abrangente a eficácia da estrutura proposta em um banco de dados de previsão de emoção audiovisual referenciado.

O restante deste artigo está organizado da seguinte forma. Na Seção II, revisamos brevemente estudos anteriores e relacionados. Na Seção III, apresentamos uma descrição detalhada da estrutura e algoritmo do framework DDAT proposto. Então, na Seção IV, oferecemos um extenso conjunto de experimentos conduzidos para exemplificar a eficácia e robustez da estrutura DDAT junto com uma discussão. Finalmente, apresentamos nossas conclusões e direções futuras de pesquisa na Seção V.

II. TRABALHO RELACIONADO

Para a previsão de emoções contínuas, muitas abordagens novas foram propostas e investigadas na última década. Algumas abordagens esperam projetar ou implementar um modelo de previsão mais sofisticado e robusto [1], [8]–[11], [14]. Dado que a informação de contexto é crucial para estimar padrões sequenciais (previsão de emoção contínua em nosso caso), redes neurais recorrentes (RNNs), especialmente aquelas implementadas com células de memória de longo prazo (LSTM), foram introduzidas [1], e são ainda entre os modelos atuais de última geração [40]. Uma das principais vantagens das RNNs LSTM é que elas podem modelar dependências de longo alcance entre sequências [24], [41] e, portanto, são eficientes na captura da informação temporal da expressão emocional [1].

Mais recentemente, a chamada arquitetura de rede end-to-end tem emergido como uma estrutura de rede promissora, que pode derivar automaticamente representações diretamente de dados brutos (não processados), em vez de extrair manualmente recursos artesanais. Por exemplo, em [14], Tzirakis et al. treinaram conjuntamente as CNNs no front-end e os LSTM-RNNs no back-end, onde as CNNs se encarregam principalmente de extrair representações de sinais de áudio brutos e os LSTM-RNNs concatenados são responsáveis por capturar as informações temporais. Uma estrutura similar também foi mostrada em [42].

Enquanto isso, algumas outras abordagens tentam superar as desvantagens de modelos individuais por meio da integração de várias modalidades ou modelos diferentes em uma estratégia de conjunto [15], [16]. Uma abordagem comum ao considerar várias modalidades é a fusão inicial (também conhecida como nível de recurso) de informações unimodais. Isso é normalmente obtido pela concatenação de todos os recursos de várias modalidades em um vetor de recursos combinado, que é então usado como informação de entrada para os modelos [16], [43]–[45]. Um benefício da fusão precoce é que ela pode fornecer melhor capacidade discriminativa ao modelo, explorando as informações complementares existentes entre as diferentes modalidades. Por exemplo, características acústicas superam empiricamente características visuais para estimativa de excitação, enquanto o oposto ocorre para estimativa de valência [43]. Outro frequentemente

A abordagem empregada é a fusão tardia (também conhecida como nível de decisão), que envolve a combinação de previsões obtidas de diversos alunos (modelos) para determinar a previsão final. Para construir os diversos alunos, Wei et al. [46] criaram um conjunto de alunos LSTM-RNN que foram treinados em diferentes modalidades (por exemplo, áudio e vídeo), enquanto Qiu et al. [47] desenvolveram uma variedade de estruturas de topologia de redes de crenças profundas (DBN).
Para combinar as previsões de vários alunos, uma abordagem direta aplica a média (não) ponderada, como a regressão linear simples (SLR) [45], [48]. Outra abordagem comum é o empilhamento, em que as previsões de diferentes alunos são empilhadas e usadas como entradas de um modelo não linear subsequente que é treinado para tomar uma decisão final [46], [47], [49]. A fim de aproveitar as vantagens individuais de diferentes modelos, Han et al. [16] propuseram ainda uma estrutura de modelagem de força que concatena dois modelos diferentes em uma arquitetura hierárquica. Nesta abordagem, a predição gerada pelo primeiro modelo é concatenada com os recursos de entrada originais, e esse vetor de recursos expandido é definido como a entrada para o próximo modelo.

Todas as abordagens descritas acima se concentram apenas em estender a capacidade ou superar as desvantagens do modelo de aprendizado. Informações sobre dificuldades no processo de aprendizagem, no entanto, raramente foram exploradas até o momento, até onde sabemos.

Além disso, o DDAT também se relaciona com o aprendizado multitarefa (MTL) [6], [38], [50], [51]. Em [51], Deng et al. reconstruíram as entradas com um AE como uma tarefa auxiliar para a previsão de emoções de maneira semi-supervisionada e demonstraram que o AE pode destilar recursos representativos de alto nível a partir de dados não rotulados em grande escala. Em [38], Han et al. propuseram a utilização do PU como uma tarefa auxiliar para predição contínua e dimensional de emoções e descobriram que essa informação ajuda a melhorar o desempenho. Em [50], Nicolaou et al. introduziu uma estrutura associativa de saída para aprender as correlações e padrões entre diferentes dimensões emocionais (ou seja, excitação e valência). Nesta estrutura, as previsões de excitação e valência de modelos independentes são fundidas e alimentadas em um modelo consequencial para uma previsão final (ou seja, excitação ou valência). A eficácia desta abordagem foi replicada em [52], [53].

Análoga ao MTL, a atual estrutura DDAT considera as tarefas de reconstrução de entradas ou previsão de incerteza de percepção como tarefas auxiliares. No entanto, o RE e o PU são assumidos como os indicadores de dificuldade de aprendizagem, e as entradas do modelo são atualizadas dinamicamente para dotar o modelo de uma capacidade de aprendizagem baseada na dificuldade.

III. TREINAMENTO DE CONSCIÊNCIA DA DIFICULDADE DINÂMICA

Nesta seção, descrevemos a estrutura DDAT. Seja x ÿ X o vetor de características no espaço de características de entrada, e y ÿ Y, o rótulo no espaço de rótulos de emoções. Para uma tarefa de reconhecimento de padrão sequencial em nosso caso, xt indica um vetor de características no t-ésimo quadro extraído de um enunciado.

A. Visão geral do sistema

O pseudocódigo que descreve o algoritmo proposto é apresentado no Algoritmo 1. Ele consiste em duas etapas principais: (i) recuperar informações de dificuldade e (ii) explorar informações de dificuldade. Na primeira etapa, para extrair e indicar as informações relacionadas à dificuldade do processo de aprendizagem, propomos duas estratégias diferentes: ontologia e estratégias de conteúdo.

A estratégia baseada em ontologia foca no próprio modelo. Especificamente, determinamos a dificuldade da tarefa por meio da reconstrução das informações de entrada, assumindo que o RE é um proxy para sua capacidade de aprendizado em um determinado momento.

Pelo contrário, a estratégia orientada por conteúdo se concentra nos dados e assume que diferentes dados podem ser aprendidos em diferentes graus. Ou seja, alguns dados podem ser facilmente aprendidos com um modelo específico, enquanto outros dados podem ser difíceis.
Esta abordagem decorre parcialmente da aprendizagem curricular [21], que demonstrou que cada dado não pode ser igualmente aprendido de forma a ser bem organizado para o treinamento do modelo. No campo da predição de emoções, alguns estudos mostraram que o nível de dificuldade dos dados a serem aprendidos está intimamente relacionado ao seu PU [35], [36], conforme discutido na Seção I. Inspirados por esses estudos, empregamos o PU para representar a dificuldade e complexidade das amostras.

Na segunda etapa, concatenamos os recursos originais xt com o vetor de dificuldade dt recuperado por uma das duas estratégias mencionadas, atualizamos as entradas via [xt, dt] e treinamos novamente o modelo de regressão para previsão contínua de emoções. Devido ao fato de que dt varia ao longo do tempo, o vetor de dificuldade estendido fornece percepção dinâmica ao modelar x em um contínuo.

B. Aprendizagem Multitarefa

MTL é um processo de aprendizagem de múltiplas tarefas simultaneamente. Normalmente, há uma tarefa principal e uma ou mais tarefas auxiliares. Ao tentar modelar as tarefas auxiliares junto com a tarefa principal, o modelo aprende informações compartilhadas entre as tarefas, o que pode ser benéfico para o aprendizado da tarefa principal. Matematicamente, a função objetivo em MTL pode ser formatada como:

$$J(\ddot{y}0) = \sum_{m=1}^{M} wmLm(x, ym; \ddot{y}m) + \ddot{y}R(\ddot{y}0), \quad (1)$$

onde M denota o número de tarefas e Lm(·) representa a função de perda da tarefa m, que é ponderada por wm. ÿ0 e ÿm representam, respectivamente, os parâmetros gerais do modelo e os específicos da tarefa m, e ÿ é um hiperparâmetro que controla a importância do termo de regularização R(ÿ0).

Neste artigo, para inferir a dificuldade da informação que está sendo modelada no primeiro estágio do framework DDAT, usamos uma estrutura MTL para aprender conjuntamente a predição de emoções contínuas juntamente com a reconstrução dos recursos de entrada ou a predição de PU. A justificativa é dupla: por um lado, o modelo faz melhor uso do MTL para predição contínua de emoções. O benefício do MTL foi demonstrado por vários estudos para a previsão de emoções, conforme descrito na Seção II. Por outro lado, o modelo usa uma rede, em vez de duas [39], para explorar a dificuldade do processo de aprendizagem.

Algoritmo 1: Treinamento Dinâmico de Conscientização da Dificuldade

Inicializar:
h: redes neurais; x: vetor
de características, x = [x1, x2, . . . , xr] I, J: épocas de treinamento predefinidas

1 se orientado por ontologia
2 **então** tarefa auxiliar T ÿ entrada de reconstrução;
3 else se orientado a conteúdo, **então**
4 tarefa auxiliar T ÿ predição da incerteza da percepção; **5 fim**

6 % recuperando informação de dificuldade estágio **7**
para i = 1, ..., eu **otimizar** a função
8 Lemt(·) + w2 ÿ Laux(·) + ÿR(ÿ0), onde w1 e w2 regulam as contribuições da predição da emoção Lemt(·) e da predição da tarefa auxiliar Laux(·); avalie h no conjunto de desenvolvimento para predição de emoções: CCCval,i; **se** CCCval,i > CCCval,iÿ1 **então** salve h; **end 13 end 14** obter a dificuldade atenção d com
9 base na tarefa auxiliar T escolhida;
10
11
12

15 % explorando o estágio de informação de dificuldade
16 para j = 1, ..., J **atualizar** a função
17 minimizando a função de perda J (ÿemt) para predição
18 de emoção; avalie h no conjunto de desenvolvimento para predição de emoção: CCCval,j ; **se** CCCval,j > CCCval,jÿ1 **então** salve h;
19 **fim 23 fim**
20
21
22

C. Recuperação de Informação de Dificuldade Baseada em Ontologia

A Figura 1 ilustra a estrutura do DDAT baseado em RE, onde o indicador de dificuldade d é gerado a partir do processo de reconstrução dos insumos. Conforme descrito na Seção III-B, a rede empregada é treinada em um contexto MTL e, portanto, a saída inclui dois caminhos – o caminho de predição de emoção e o caminho AE. O primeiro é treinado de forma supervisionada, enquanto o último é treinado de forma não supervisionada.

Assim, há duas tarefas a serem realizadas no treinamento da rede, ou seja, prever emoções e reconstruir entradas.
Especificamente, dada uma sequência de tempo como entrada x, a rede é otimizada minimizando a função de perda como

$$J(\ddot{y}0) = w1 \ddot{y} Lemt(\cdot) + w2 \ddot{y} Lre(\cdot) + \ddot{y}R(\ddot{y}0), \quad (2)$$

onde Lemt(·) e Lre(·) denotam as funções de perda para previsão de emoção e reconstrução de entrada, respectivamente. Para calculá-los, tomamos o erro quadrático médio (MSE) para ambos os caminhos de aprendizagem, ou seja, para previsão de emoções,

$$Lemt(\cdot) = \sum_{t=1}^{T} |\hat{y}t \ddot{y} yt||2 ; \quad (3)$$

e para a reconstrução de entrada,

$$Lre(\cdot) = \sum_{t=1}^{T} |\hat{x}t \ddot{y} xt||2 , \quad (4)$$

onde xt e yt são uma amostra e sua anotação no tempo t de uma sequência de entrada com um período de tempo T, respectivamente. xˆt e yˆt denotam as previsões da rede para reconstruir suas entradas xt e estimar as emoções yt, respectivamente.

Espera-se que Lre(·) ÿ 0 se o modelo for suficientemente poderoso e robusto. No entanto, experimentos empíricos mostraram que os resultados estão longe dessa expectativa. Descobertas anteriores freqüentemente indicam que uma maior incompatibilidade de distribuição entre os dados fornecidos e todo o conjunto de dados de treinamento tende a produzir uma RE mais alta [30], [31], [54], [55]. Portanto, o RE de certa forma implica no grau de dificuldade do modelo em aprender tais dados ou, em outras palavras, reflete a dificuldade dos dados em serem aprendidos pelo modelo.

Uma vez treinado o modelo, a dificuldade do processo de aprendizado (d) pode ser obtida computando-se a distância entre a entrada x e sua correspondente reconstrução xˆ. A distância pode ser um vetor e calculado por,

$$d = e = x \ddot{y} \hat{x}, \quad (5)$$

ou um escalar E somado sobre todos os atributos, ou seja,

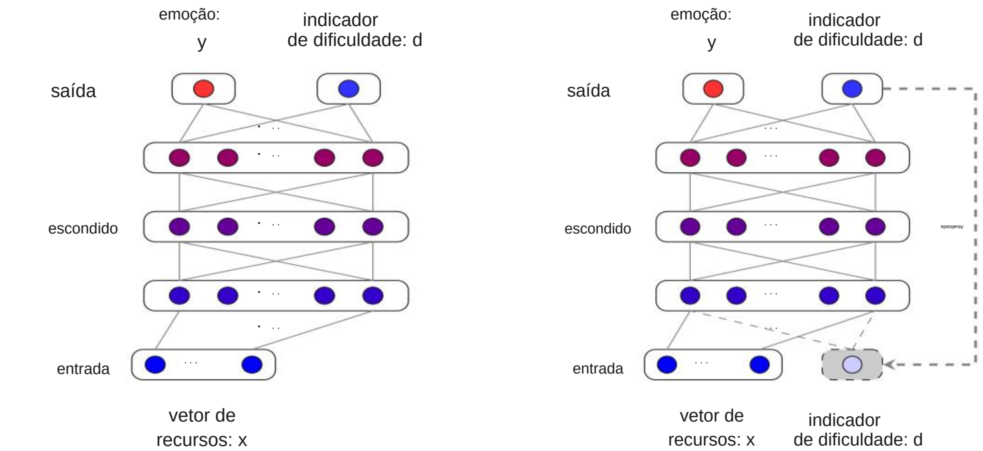
$$d = [E] = [r \sum_{i=1} (xi \ddot{y} \hat{x}i)], \quad (6)$$

onde x = [x1, x2, . . . , xr] e r é a dimensão do vetor de recursos.

Na etapa de exploração da dificuldade, atualizamos a entrada do modelo com o novo vetor, ou seja, x = [x, e] ou x = [x, E].
Ao fazer isso, os vetores de recursos de entrada são de 2r ou r + 1 dimensões ao retroalimentar um vetor de erro ou escalar.

D. Recuperação de informações de dificuldade baseada em conteúdo

Conforme mencionado anteriormente, UP é um indicador do nível de incerteza da percepção de um estado emocional para uma determinada amostra observada. No contexto da computação afetiva, consideramos que a previsão de emoções é uma tarefa subjetiva que difere de muitas outras tarefas objetivas de reconhecimento de padrões, como o reconhecimento facial, que possuem uma verdade básica [56]. A fim de obter um padrão-ouro para uma tarefa subjetiva, é necessário que um número suficiente de avaliadores observe a mesma amostra e que suas classificações sejam reduzidas para eliminar, tanto quanto possível, as variações individuais na percepção e classificação. Nesse caso, uma forma possível de inferir a incerteza é calculando o nível de discordância entre avaliadores, que assume que, para cada amostra, a UP pessoal está altamente correlacionada com o nível de discordância entre avaliadores [38], [57].



(a) Estágio de recuperação de informações de dificuldade (b) Estágio de exploração de informações de dificuldade

Fig. 1. O Treinamento Dinâmico de Conscientização da Dificuldade (DDAT) inclui informações sobre a dificuldade (a) estágio de recuperação e (b) estágio de exploração. A informação de dificuldade é indicada pelo erro de reconstrução de entrada (ou seja, um vetor de erro ou a soma de todos os erros) ou pelas incertezas de percepção de emoção.

Neste estudo, o PU $u^{(i)}$ i ∈ {excitação, valência}, é representado pelo desvio padrão das anotações K

Como

$$u_n^{(i)} = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(y_{n,k}^{(i)} - \bar{y}_n^{(i)}\right)^2}, \quad (7)$$

onde $\bar{y}^{(i)}$ denota o valor médio dado as anotações K:

$$\bar{y}_n^{(i)} = \frac{1}{K}\sum_{k=1}^{K} y_{n,k}^{(i)}. \quad (8)$$

A estrutura do DDAT baseado em PU também é ilustrada na Fig. 1, onde o indicador de dificuldade d é determinado pela incerteza da percepção. A rede projetada inclui um caminho de predição de emoção e um caminho de predição de PU, ambos treinados conjuntamente de maneira supervisionada. Portanto, a função objetivo da Eq. (1) pode ser reformulado como

$$J(\theta) = w_1 \cdot L_{emt}(\theta) + w_2 \cdot L_{pu}(\theta) + \lambda \cdot \mathbf{R}(\theta). \quad (9)$$

$L_{pu}(\theta)$ representa as funções de perda para previsão de PU e é expressa por

$$L_{pu}(\theta) = \sum_{t=1}^{T} ||\hat{u}_t - u_t||, \quad (10)$$

onde u_t é um valor PU para a amostra no tempo t das sequências de entrada com o tempo T.

Uma vez que a rede é otimizada no primeiro estágio de aprendizado, sua entrada evoluirá para x = [x, u] com r + 1 dimensões no segundo estágio de aprendizado.

E. Late Fusion e Dynamic Tuning

Conforme discutido na Seção II, as abordagens de fusão tardia têm se mostrado frequentemente eficazes para a previsão de emoções contínuas [15], [16], [48] devido ao fato de que informações complementares podem ser fornecidas pelas várias modalidades ou modelos [15] , [16], [48]. Sob esta luz, realizamos uma fusão tardia para combinar as previsões de emoções de diferentes modalidades, modelos de aprendizagem ou uma combinação dos mesmos. A fusão tardia é realizada com uma abordagem SLR:

$$y = \beta + \sum_i \beta_i \cdot y_i, \quad (11)$$

onde y_i denota a predição original com a modalidade (ou seja, áudio ou vídeo) ou modelo i (ou seja, DDAT baseado em RE ou PU), e β_i são os parâmetros estimados no conjunto de desenvolvimento e y é a predição fundida.

Apesar da eficácia do SLR, esta abordagem de fusão convencional simplesmente assume que as previsões $y_{i,t}$ em um contínuo são consideradas igualmente importantes para cada previsão. Isso significa que o parâmetro de β_i é, portanto, um constante no tempo, dado um conjunto de fluxo y_i. Esta abordagem ignora as mudanças na confiabilidade das previsões. Para resolver este problema, propomos ainda uma estratégia de ajuste dinâmico de acordo com a confiabilidade das previsões no tempo.

Matematicamente, aplicamos um SLR adicional na previsão original $y_{i,t}$ e o indicador de dificuldade correspondente $d_{i,t}$ no tempo t:

$$y'_{i,t} = \gamma + \gamma_i \cdot y_{i,t} + \gamma_d \cdot d_{i,t}, \quad (12)$$

onde $d_{i,t}$ é representado por E_t para os sistemas DDAT baseados em RE ou u_t para os sistemas DDAT baseados em PU. Intuitivamente, a previsão é ajustada dinamicamente pela informação de dificuldade.

4. EXPERIMENTOS E RESULTADOS

Para avaliar a eficácia dos métodos propostos, realizamos extensos experimentos com o banco de dados de referência do AudioVisual Emotion Challenges (AVEC) de 2015 [58] e 2016 [48].

A. Bancos de dados e recursos

O corpus multimodal REmote COLlaborative and Af fective interações (RECOLA) [59] (um banco de dados padrão de

os desafios AVEC para previsão de emoção contínua no tempo audiovisual [48], [58]) foi selecionado para nossos experimentos devido ao seu uso generalizado nesta área. Esta base de dados foi criada para estudar comportamentos socioafetivos a partir de dados multimodais no contexto de tarefas colaborativas remotas. Inclui gravações audiovisuais (e fisiológicas) de interações espontâneas e naturais de 27 participantes francófonos durante a resolução de uma tarefa colaborativa realizada em díades por meio de videoconferência. O corpus é composto por gravações de áudio, vídeo e fisiologia periférica que foram obtidas de forma síncrona e contínua ao longo do tempo.

A fim de garantir a independência do locutor para experimentos de ML, o corpus foi dividido em três partições – treinamento, desenvolvimento (validação) e teste – com cada partição contendo nove sessões colaborativas. Esta divisão é equilibrada em termos de gênero, idade e língua materna dos participantes.
O corpus contém anotações contínuas de valor e tempo de duas dimensões afetivas – excitação e valência – que foram obtidas de seis avaliadores falantes de francês (três mulheres) nos primeiros cinco minutos de cada gravação audiovisual. Os rótulos obtidos foram reamostrados a uma taxa de quadros constante de 40 ms e a média calculada sobre todos os avaliadores para criar um 'padrão ouro' para cada instância. Desentendimentos entre avaliadores também foram computados para todas as instâncias [59]. Para nossos experimentos, utilizamos apenas sinais de áudio e vídeo.

As características acústicas e visuais empregadas em nossos experimentos são os mesmos conjuntos usados para calcular as linhas de base AVEC 2015 e 2016 para comparação justa com outros métodos. Os recursos acústicos consistem no Conjunto de Parâmetros Acústicos Minimalistas de Genebra estendido (eGeMAPS [60]). Uma vez que o banco de dados RECOLA contém sinais e anotações contínuas de tempo longo, dois funcionais (média aritmética e derivação padrão) foram aplicados sobre os descritores sequenciais de baixo nível (LLDs, por exemplo, pitch, loudness, energia, Mel Frequency Cepstral Coeficientes, jitter , e shimmer) em uma janela fixa de 8 s com um passo de 40 ms. Isso resultou em um conjunto de 88 características acústicas por segmento.

Em relação aos recursos visuais, utilizamos tanto a aparência quanto os recursos padrão geométricos dos desafios AVEC. As características de aparência foram computadas usando padrões binários de Gabor locais de três planos ortogonais através da divisão do vídeo em volumes de vídeo espaço-temporais. Uma redução de característica foi então realizada aplicando uma análise de componentes principais de uma aproximação de classificação baixa (até a classificação 500), levando a 84 características representando 98% da variância. Para extrair as características geométricas, primeiro foram extraídos 49 marcos faciais de cada quadro e, em seguida, alinhados com uma forma média a partir de pontos estáveis (localizados nos cantos dos olhos e na região do nariz). Isso resultou em 316 feições por quadro: ou seja, 196 feições foram obtidas calculando a diferença entre as coordenadas dos marcos alinhados e aquelas da forma média e entre os locais dos marcos alinhados no quadro anterior e atual, e 71 foram obtidos por calculando as distâncias euclidianas (norma L2) e os ângulos (em radianos) entre os pontos em três grupos diferentes. Um adicional de 49 recursos correspondem à distância euclidiana entre a mediana dos pontos de referência estáveis e cada ponto de referência alinhado em um vídeo quadro.

Semelhante às características acústicas, a média aritmética e a derivação padrão foram calculadas sobre as características visuais sequenciais de cada quadro usando uma janela deslizante de 8 s com um tamanho de passo de 40 ms. Este processo levou a 168 aparências e 632 características visuais geométricas.

Para detalhes completos sobre o banco de dados e conjuntos de recursos, consulte [48], [58]. Observe que obtivemos 67,5 k segmentos extraídos no total para cada partição (treinamento, desenvolvimento ou teste).

B. Configuração Experimental e Métricas de Avaliação

A estrutura DDAT implementada em nossos experimentos consiste em uma RNN profunda (DRNN) equipada com unidades recorrentes fechadas (GRUs) [61]. As GRUs são uma alternativa às unidades de memória de longo prazo, que também podem capturar as dependências de longo prazo em tarefas baseadas em sequência e mitigar os efeitos do problema do gradiente de fuga [61]. Em comparação com as unidades LSTM, as GRUs têm menos parâmetros devido ao fato de não possuírem células de memória e portas de saída separadas, o que resulta em um processo de treinamento mais rápido e uma demanda menor de dados de treinamento para obter uma boa generalização. Mais importante ainda, muitas avaliações empíricas [62] indicaram que as GRUs têm um desempenho tão competitivo quanto as unidades LSTM.

A estrutura DRNN foi otimizada em termos de número de camadas ocultas e número de GRUs por camada na fase de desenvolvimento. Aplicamos uma grade de pesquisa composta por {1, 3, 5, 7, 9} camadas ocultas e {40, 80, 120} unidades ocultas por camada. Para cada estratégia de aprendizado, sempre escolhemos a estrutura de rede com melhor desempenho, a fim de aliviar o impacto da variação das estruturas de rede no desempenho do sistema. O treinamento das DRNNs foi realizado utilizando o algoritmo de otimização de Adam [63] com uma taxa de aprendizado inicial de 0,001. Para facilitar o processo de treinamento, definimos o tamanho do minilote para quatro. Além disso, uma padronização online foi aplicada aos dados de entrada usando as médias e variações do conjunto de treinamento.

Além disso, conforme sugerido em [48], a compensação de atraso de anotação foi empregada para compensar o atraso temporal entre as pistas observáveis e as anotações correspondentes relatadas pelos anotadores [64]. Identificamos esse atraso em 2,4 s, de acordo com uma série de avaliações experimentais em [65], e mudamos o padrão-ouro de volta no tempo com relação aos recursos para todas as modalidades e tarefas em nossos experimentos.

Para avaliar o desempenho dos modelos, usamos a métrica oficial dos desafios AVEC 2015 e 2016 – o Coeficiente de Correlação de Concordância (CCC) [58]:

$$rc = \frac{2r\ddot{y}_x\ddot{y}_y}{2\ddot{y}_x + \ddot{y}_y^2 + (\mu_x - \mu_y)^2}, \quad (13)$$

onde r representa o coeficiente de correlação de Pearson entre duas séries temporais (por exemplo, previsão e padrão-ouro), μ_x e μ_y denotam a média de cada série temporal e $\ddot{y}$ representam as variações $\ddot{y}_x^2$ e $\ddot{y}_y^2$ correspondentes. Comparado com o PCC, o CCC considera não apenas a semelhança de forma entre duas séries, mas também a precisão do valor. Isso é especialmente relevante para estimar o desempenho de modelos de previsão de emoções contínuas no tempo, pois tanto as tendências quanto as absolutas

os valores de previsão são relevantes para descrever o desempenho de um modelo. A métrica CCC cai na faixa de [-1, 1], onde +1 representa concordância perfeita, -1 discordância total e 0 nenhuma concordância.

Para refinar a previsão obtida, realizamos ainda uma cadeia de pós-processamento, incluindo filtragem mediana, centralização, dimensionamento e mudança de tempo, conforme sugerido em [48], [58]. O tamanho da janela de filtragem W (variando de 0,12 s a 0,44 s a uma taxa de 0,08 s) e o atraso de deslocamento de tempo D (variando de 0,04 s a 0,60 s em um passo de 0,04 s) foram otimizados usando um método de busca em grade. Todos os parâmetros de pós-processamento foram otimizados no conjunto de desenvolvimento e então aplicados ao conjunto de teste. Portanto, esses parâmetros de pós-processamento tinham várias configurações para diferentes tarefas.

Para comparar a abordagem DDAT proposta com outras abordagens relacionadas e de última geração, conduzimos o aprendizado curricular, conforme apresentado na Seção I. Selecionamos particularmente o critério de 'discordância entre anotadores' (ou seja, PU neste artigo) como exemplo porque é apropriado para a tarefa em questão e também superior a outros critérios [23].
Para manter as configurações otimizadas, continuamos usando as redes neurais profundas (DNNs) equipadas com duas camadas ocultas (1.024 nós por camada) e dividimos todo o conjunto de treinamento em cinco partes com base nos níveis de PU. Além disso, nós o implementamos também com GRU-RNNs para uma comparação de desempenho justa entre o aprendizado curricular e o DDAT proposto.

Finalmente, para comparar estatisticamente os vários experimentos conduzidos com as linhas de base dos desafios AVEC, realizamos a transformação r-to-z de Fisher [66]. Em detalhe, dadas duas distribuições X e Y [os pares (Xi , Yi) ÿ iid] que têm uma distribuição normal bivariada com correlação, a transformação de Fisher z é aproximadamente normalmente distribuída com

$$\text{significar} \quad m = \frac{1}{2}\ln\frac{1+r}{1-r} = \operatorname{arctanh}(r), \tag{14}$$

e erro padrão

$$\sigma = \frac{1}{\sqrt{N-3}}, \tag{15}$$

onde N é o tamanho da amostra e r é o verdadeiro coeficiente de correlação. Teoricamente, a transformação de Fisher é excepcionalmente eficiente para tamanhos de amostra pequenos porque a distribuição de amostragem da correlação de Pearson normalmente é altamente distorcida.

Após a transformação de Fisher, um teste unicaudal foi realizado para comparar duas distribuições. Um valor p inferior a 0,05 indica uma diferença significativa. Note-se que r é substituído por rc (CCC) devido à eficiência de rc neste artigo.

C. Previsão de emoções com treinamento dinâmico de consciência de dificuldade

O desempenho dos sistemas avaliados antes e depois do pós-processamento das predições para os alvos de excitação e valência é apresentado nas Tabelas I e II, respectivamente.
Para investigar a estrutura DDAT proposta, não apenas conduzimos o aprendizado de tarefa única tradicional, mas também o MTL para comparação, com três conjuntos de recursos diferentes - um conjunto de recursos acústicos (eGeMAPS) e dois conjuntos de recursos visuais (aparência e características geométricas), conforme descrito na Seção IV-A. Vale ressaltar que a estrutura de rede empregada para cada modalidade e abordagem de aprendizagem foi otimizada respectivamente no espaço de parâmetros restrito, conforme mencionado na Seção IV-B.
Em seguida, as estruturas de rede com melhor desempenho foram empregadas para comparação de desempenho. Isso alivia amplamente o impacto inconsistente no desempenho do sistema devido à variação das estruturas de rede. Da comparação das Tabelas I e II, pode-se ver que o pós-processamento das previsões do modelo geralmente leva a um melhor desempenho.
Por exemplo, as melhores linhas de base para ativação e valência são respectivamente aumentadas de 0,617 a 0,652 CCC com recursos acústicos (eGeMAPS) e de 0,403 a 0,417 CCC com recursos visuais (geométricos). Observações semelhantes também podem ser obtidas nos sistemas MTL e nos sistemas DDAT propostos; por exemplo, para os sistemas MTL, os CCCs são aumentados de 0,613 para 0,654 com o conjunto de recursos eGeMAPS para excitação e de 0,487 para 0,488 com o conjunto de recursos geométricos para valência. Diante desses resultados, focamos doravante na análise dos experimentos com a etapa de pós-processamento (cf. Tabela II).

Para o sistema de linha de base, os resultados obtidos são competitivos ou até melhores do que o benchmark do subdesafio de previsão de emoções no AVEC 2016 [48] em três fluxos de informações e duas tarefas de previsão. Esses resultados apóiam descobertas anteriores, mostrando que as GRUs podem oferecer desempenho competitivo quando comparadas às unidades LSTM [61], [62].

Ao treinar as redes em conjunto com a reconstrução de entrada (MTL baseado em RE) ou previsão de incerteza de percepção (MTL baseado em PU), pode-se observar que os sistemas executam ligeiramente os sistemas de linha de base em nove dos doze casos no conjunto de teste. Isso indica que existe uma relação substancial entre as duas tarefas aprendidas em conjunto. Para ser mais específico, as representações da última camada oculta da rede neural, que são aprendidas de forma síncrona a partir da previsão da emoção e outras tarefas auxiliares (ou seja, reconstruir a entrada ou prever a incerteza da percepção), potencialmente beneficiam ainda mais a previsão da emoção.

Implementamos ainda mais a abordagem de aprendizagem curricular, bem como sua linha de base por meio de DNNs [23] e GRU RNNs. A partir da Tabela 1, pode-se observar que as DNNs têm um desempenho surpreendentemente pior do que as GRU-RNNs, principalmente devido à sua capacidade limitada de capturar as informações de contexto [24].
Ao alimentar os dados para o modelo de treinamento de um nível de dificuldade baixo para um nível de dificuldade alto, o desempenho dos modelos é notavelmente aprimorado em todos os cenários. No entanto, ainda não é competitivo com os modelos DDAT na maioria dos casos.
Além disso, observa-se que o aprendizado curricular baseado em GRU-RNN supera o sistema baseado em DNN principalmente devido à capacidade de aprendizado dos GRUs.

O desempenho dos sistemas MTL é aprimorado ainda mais pela estrutura DDAT proposta, conforme mostrado na Tabela II. Em particular, o desempenho do sistema DDAT para regressões de excitação e valência, respectivamente, atinge valores de CCC de 0,694 e 0,422 com o conjunto de recursos de áudio-eGeMAPS, 0,438 e 0,457 com o conjunto de recursos de aparência de vídeo e 0,400 e 0,501 com o conjunto de recursos de vídeo geométrico conjunto de características. Esses resultados demonstram que os sistemas DDAT superam significativamente

TABELA I

DESEMPENHO DO SISTEMA (COEFICIENTE DE CORRELAÇÃO DE CONCORDÂNCIA ; CCC) **antes** do PÓS-PROCESSAMENTO DAS PREDIÇÕES DO MODELO PARA A ESTRUTURA CONVENCIONAL DE APRENDIZAGEM DE TAREFAS ÚNICA (LINHA DE BASE) , A ESTRUTURA DE APRENDIZAGEM MULTI-TAREFAS (MTL) E O TREINAMENTO DE CONSCIÊNCIA DE DIFICULDADE DINÂMICA PROPOSTO (DDAT) ESTRUTURA UTILIZANDO ERRO DE RECONSTRUÇÃO (RE, VETOR OU ESCALAR DE SOMA) E VARIANTES DE INCERTEZA DE PERCEPÇÃO (PU) . ESTES RESULTADOS RELACIONADOS AOS EXPERIMENTOS CONDUZIDOS NO DESENVOLVIMENTO E PARTIÇÕES DE TESTE PARA AMBOS OS ALVOS DE AROUSAL E DE VALÊNCIA. TRÊS CONJUNTOS DE RECURSOS (ÁUDIO-EGEMAPS, VÍDEO-APARÊNCIA E VÍDEO-GEOMÉTRICO) FORAM EMPREGADOS PARA AVALIAR TODAS AS ABORDAGENS. OS CASOS EM QUE DDAT TEM SIGNIFICADO ESTATÍSTICO DE MELHORIA DE DESEMPENHO SOBRE MTL SÃO MARCADO PELO "" SÍMBOLO.

CCC	áudio-eGeMAPS				vídeo-aparição				videogeométrico			
	aro		val		aro		val		aro		val	
	desenvolvedor	teste	desenvolvedor	teste	desenvolvedor	teste	desenvolvedor	teste	desenvolvedor	teste	desenvolvedor	teste
linha de base	0,743	.617	.460	.380	.501	.416	0,481	.391	.407	.256	.598	.403
Aprendizagem multitarefa (MTL)												
baseado em RE	.743	.590	.513	.298	.485	.472	.434	.512	.351	.429	.283	.632	.487	.630
baseado em PU	.727	.613	.417			.459	.426	.444	.342	.442	.284	.444		
Treinamento de Conscientização da Dificuldade Dinâmica Proposta (DDAT)												
Baseado em RE (vetor) 0,745 0,605 Baseado em RE (soma) 0,783 0,671 Baseado em PU 0,769 0,623			.485	.374	.495	.473	.429	.482	.326	.509	.299	.627	.464	.633
			.410	.493	.397	.487	.464	.507	.460	.478	.359	.467	.629	.500
						.478	.457	.476	.412	.450	.336			
Outro estado da arte												
DNS [23]	.216	.362	.003	.004	.018	.265	.173	.294	.174	.017	.011	.193	.128	.271
Aprendizagem Curricular (DNN) [23]	.445	.533	.006	.494	.335	.402	.292	.437	.444	.199	.116	.245	.582	.497
Aprendizagem Curricular (GRU-RNN)	.711	.600				.406	.366	.547	.479	.392	.291			

TABELA II

DESEMPENHO DO SISTEMA (COEFICIENTE DE CORRELAÇÃO DE CONCORDÂNCIA ; CCC) **após** o PÓS-PROCESSAMENTO DAS PREDIÇÕES DO MODELO PARA A ESTRUTURA CONVENCIONAL DE APRENDIZAGEM DE TAREFAS ÚNICA (LINHA DE BASE) , A ESTRUTURA DE APRENDIZAGEM MULTI-TAREFAS (MTL) E O TREINAMENTO DE CONSCIÊNCIA DE DIFICULDADE DINÂMICA PROPOSTO (DDAT) ESTRUTURA UTILIZANDO ERRO DE RECONSTRUÇÃO (RE, VETOR OU ESCALAR DE SOMA) E VARIANTES DE INCERTEZA DE PERCEPÇÃO (PU) . ESTES RESULTADOS RELACIONADOS AOS EXPERIMENTOS CONDUZIDOS NO DESENVOLVIMENTO E PARTIÇÕES DE TESTE PARA AMBOS OS ALVOS DE AROUSAL E DE VALÊNCIA. TRÊS CONJUNTOS DE RECURSOS (ÁUDIO-EGEMAPS, VÍDEO-APARÊNCIA E VÍDEO-GEOMÉTRICO) FORAM EMPREGADOS PARA AVALIAR TODAS AS ABORDAGENS. OS MELHORES RESULTADOS ALCANÇADOS NO CONJUNTO DE TESTE ESTÃO EM NEGRITO. OS CASOS ONDE DDAT TEM UM ESTATÍSTICO SIGNIFICADO DA MELHORIA DE DESEMPENHO SOBRE MTL É MARCADO PELO "" SÍMBOLO.

CCC	áudio-eGeMAPS				vídeo-aparição				videogeométrico			
	aro		val		aro		val		aro		val	
	desenvolvedor	teste	desenvolvedor	teste val	desenvolvedor	teste	desenvolvedor	teste	desenvolvedor	teste	desenvolvedor	teste
linha de base	0,783	0,652	.473	.400	.528	.403	0,493	.404	.523	.314	.620	.417		
Aprendizagem multitarefa (MTL)												
baseado em RE	.788	.629	.519	.331	.506	.512	.425	.529	.366	.502	.324	.632	.488	.643
baseado em PU	.803	.654	.416			.502	.406	.468	.418	.508	.327	.452		
Treinamento de Conscientização da Dificuldade Dinâmica Proposta (DDAT)												
Baseado em RE (vetor) .806 .517 .378 Baseado em RE (soma) .807 .676 **.508 .422 Baseado** em PU .811 .498 .407		**.694**				.533	.434	.520	.329	.559	.355	.634	.473	.639
		.664				.539	.437	.528	**.457**	.544	**.400**	.471	.632	**.501**
						.518	**.438**	.514	.431	.513	.397			
Outras DNNs de última geração [23] .573 Aprendizagem curricular (DNN) [23] .687 Aprendizagem curricular (GRU-RNN) .754 Modelagem de força [16] .755 Ponta a ponta [14] .786 b Seleção de recursos + deslocamento [49] 0,800 SVR + deslocamento [48] 0,796 SC + CNN + LSTM [40] 0,846		.517	.129	.044	.159	.387	.220	.306	.206	.312	.296	.362	.216	.300
		.591	.174	.501	.357	.417	.343	.446	.419	.394	.267	.269	.609	.500
		.611	.476	.364	.428	.491	.391	.557	.492	.444	.336			
		.666	.369	.398	– .455	.350	.196	.592	.464	–	–	–	–	
		.715	.375	.450	–	.371	.435	.637	.620	–	–	–	–	
[a]		–				.587	–	.441	–	.173	–	.441	–	.612
[c]		.648				.483	.343	.474	.486	.379	.272	.507		
[d]		–				.346	–	.511	–	–	–	–	–	

Nota: "–" indica que o CCC correspondente não foi fornecido.

[a] características acústicas e visuais extraídas automaticamente por modelos de redes neurais profundas

[b] Método vencedor do desafio AVEC '15

[c] Método baseline AVEC '16

[d] Método vencedor do desafio AVEC '16

(p < 0,05 via transformação Fisher r-to-z) o método de linha de base, bem como a abordagem MTL (exceto no caso de regressão de valência com o conjunto de recursos de áudio-eGeMAPS).

Além disso, os sistemas que usam a estrutura DDAT proposta superam consistentemente a abordagem de aprendizado curricular e são competitivos e, em alguns casos, até superiores. para, a maioria dos outros métodos de ponta, como a modelagem de força [16] e os sistemas 'codificação esparsa (SC) + CNN + LSTM' (vencedor do AVEC 2016) [40]. Apesar do fato de que os sistemas propostos são um pouco piores do que o sistema fim-a-fim, que extrai automaticamente as representações dos sinais de áudio e vídeo brutos que retêm o padrão completo

TABELA III
PCCS OBTIDO ENTRE SI ENTRE A MELHORIA DE DESEMPENHO (ÿc), O ERRO DE RECONSTRUÇÃO (), E A INCERTEZA DE PERCEPÇÃO (μ).

	aro		val	
	desenvolvedor	teste	desenvolvedor	teste
PCC(, ÿc) áudio-eGeMAPS vídeo-	.128	.180	.078	.114
aparência vídeo-	.139	.371	.304	.263
geométrico	.140	.155	.044	.090
PCC(μ, ÿc) áudio-eGeMAPS vídeo-	.150	.181	.150	.181
aparência vídeo-	.205	.173	.383	.440
geométrico	.101	-.104	.310	.103
PCC(, μ) áudio-eGeMAPS vídeo-	.150	0,072	0,040	-0,024
aparência vídeo-	-.077	.060	-.077	.059
geométrico	.127	0,071	0,050	0,021

informações, a estrutura DDAT pode ser incorporada ao sistema de ponta a ponta no futuro.

Ao comparar as duas abordagens usadas nos experimentos DDAT baseados em RE, descobrimos que adicionar a soma geral do erro [cf. Fig. 1 (b)] leva a um melhor desempenho do que adicionar o vetor de erro [cf. Fig. 1 (a)]. Isso pode ser atribuído à dimensionalidade redundante do vetor de erro, que entretanto gera muito ruído no treinamento da rede. Ao comparar o DDAT baseado em RE e o DDAT baseado em PU, percebe-se que as duas abordagens têm desempenho semelhante.

Isso sugere que ambas as abordagens atingem o mesmo objetivo, mas de maneiras diferentes. Ou seja, ambas as abordagens exploram com sucesso a informação de dificuldade no processo de aprendizado de padrões, enquanto as abordagens DDAT baseada em RE e baseada em PU medem a informação de dificuldade pela capacidade de reconstrução dos dados e pela incerteza-percepção dos dados, respectivamente.
Além disso, vale a pena mencionar que a abordagem DDAT baseada em RE, em contraste com a DDAT baseada em PU, não apenas se encaixa nas tarefas subjetivas de reconhecimento de padrões (por exemplo, predição de emoções neste trabalho), mas também possui o potencial para ser aplicada a tarefas objetivas. tarefas (por exemplo, predição de fonemas).

Para investigar a contribuição da informação de dificuldade extraída para a melhoria de desempenho do sistema, calculamos ainda a correlação (em termos de PCC) entre os valores do indicador de dificuldade (ou seja, o RE obtido ou o PU) entre a melhoria de desempenho. Especificamente, a melhoria de desempenho ÿc foi calculada como ÿc = |yˆbsÿy|ÿ|yˆDDAT ÿy|, dado o alvo (padrão ouro) y e a previsão do sistema DDAT yˆDDAT (ou o sistema de linha de base yˆbs).

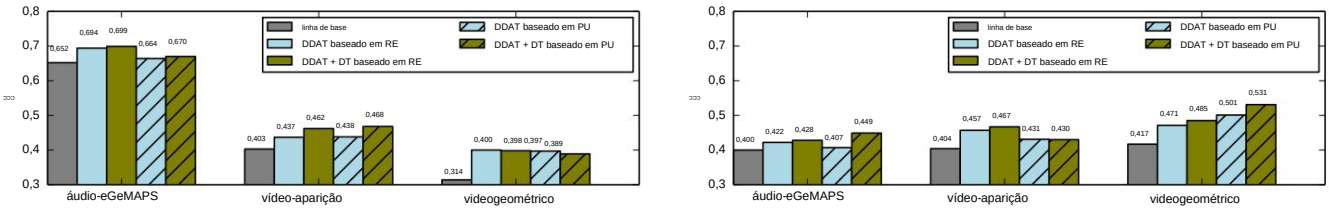
As três primeiras linhas da Tabela III mostram os PCCs obtidos entre o RE e a melhoria de desempenho [ie, PCC(, ÿc)]. Esses PCCs positivos sugerem que a informação da dificuldade pode ajudar a melhorar o desempenho do modelo no processo de aprendizagem. Esta conclusão confirma nossas descobertas anteriores em [39]. Observe que, em [39], a base de dados selecionada possui assuntos diferentes deste artigo. Observações semelhantes podem ser encontradas ao calcular os PCCs entre o PU e a melhoria de desempenho [isto é, PCC(μ, ÿc), como mostrado nas segundas três linhas]. Os PCCs são aumentados para 0,384 e 0,440 nos conjuntos de desenvolvimento e teste no caso de valência ao usar recursos visuais baseados em aparência. Mais detalhadamente, pode-se ver que, ao usar a abordagem DDAT baseada em RE, os PCCs alcançados para previsão de excitação são relativamente maiores do que os de previsão de valência na maioria dos casos. No entanto, uma observação oposta é feita ao usar a abordagem DDAT baseada em PU. Isso provavelmente se deve ao fato de que o arousal é mais sensível do que a valência à força de expressão ou escala que potencialmente resulta em maior RE, enquanto a valência está mais associada às variações sutis que facilmente enganam o julgamento dos anotadores [15], [56].

Além disso, calculamos os PCCs entre o RE e o PU obtidos, conforme mostrado nas três últimas linhas da Tabela III.
De um modo geral, a maioria desses PCCs está em torno de zero, indicando que o RE obtido é amplamente independente do PU.
Isso implica ainda que as estratégias propostas de DDAT baseadas em RE e baseadas em PU capturam os diferentes fenômenos subjacentes.
Assim, espera-se que a combinação das duas abordagens proporcione melhor desempenho. Os experimentos relacionados e os resultados correspondentes são apresentados na Seção IV-D.

D. Ajuste dinâmico e fusão tardia A Figura 2

ilustra os desempenhos dos modelos DDAT com e sem ajuste dinâmico das previsões. Em comparação com as previsões sem ajuste dinâmico, as previsões ajustadas dinamicamente geram ganhos na maioria dos casos. Por exemplo, o melhor CCC alcançado para predição de excitação aumentou de 0,684 para 0,699, usando o sistema DDAT baseado em RE com o conjunto de recursos de áudio-eGeMAPS, enquanto para predição de valência aumentou de 0,511 para 0,531, usando o sistema DDAT baseado em PU com o conjunto de recursos geométricos de vídeo. As exceções incluem as previsões de excitação para sistemas DDAT baseados em RE e PU usando o conjunto de recursos geométricos de vídeo e as previsões de valência para o sistema DDAT baseado em PU usando o conjunto de recursos de aparência de vídeo. Em ambos os casos, as diferenças permanecem mínimas e insignificantes por meio do teste estatístico acima mencionado da transformação z-para-r de Fisher.

Em seguida, realizamos um conjunto de fusões tardias nas previsões individuais produzidas pelo uso de diferentes modalidades e modelos.
A Tabela IV lista todos os cenários (combinações) considerados em nossos experimentos, bem como o respectivo desempenho. Como pode ser visto na tabela, o melhor desempenho no conjunto de teste para ativação e valência foi obtido ao fundir as previsões de todas as modalidades e modelos. Nesse contexto, os melhores resultados no conjunto de teste foram alcançados em 0,766 CCC para excitação e 0,660 CCC para valência. Esses resultados superam a maioria dos últimos resultados relatados dos mesmos dados e estão próximos do melhor resultado apresentado no AVEC 2016 [40] (ou seja, 0,770 e 0,687 de CCCs para previsão de despertar e valência), apesar de esse sistema também utilizar um modalidade adicional (características fisiológicas). Uma ilustração do desempenho do melhor sistema DDAT em comparação com o sistema de linha de base e o padrão-ouro é mostrada na Fig. 3 (dados de um sujeito aleatório da partição de teste). De modo geral, percebe-se que nossas previsões estão mais próximas do padrão-ouro, principalmente na região que possui valores relativos de pico.



(a) excitação (b) valência

Fig. 2. Comparação de desempenho (CCC) entre o aprendizado de tarefa única, a abordagem de treinamento de conscientização de dificuldade dinâmica proposta com base no erro de reconstrução (RE) ou na incerteza de percepção (PU) e suas versões sintonizadas dinamicamente (DT). Os resultados referem-se à partição de teste para alvos de excitação (a) e valência (b) usando três conjuntos de recursos (áudio-eGeMAPS, vídeo-aparência e vídeo-geométrico).

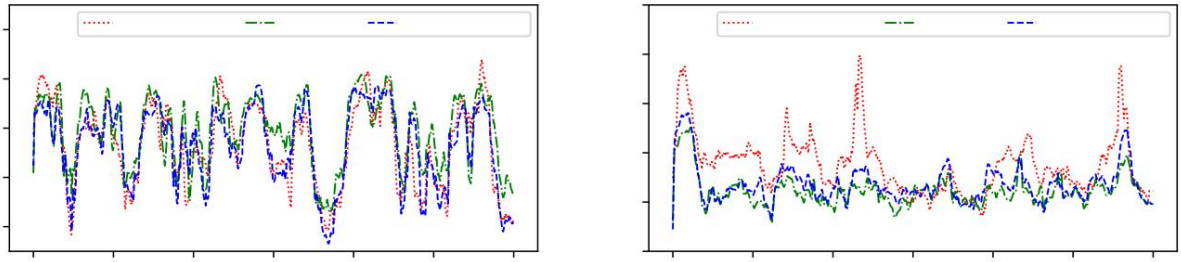
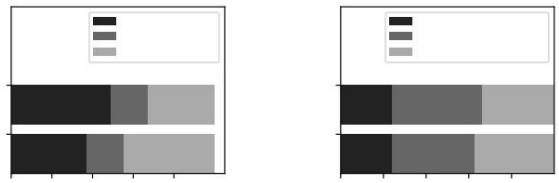


Fig. 3. Previsão automática de excitação (a) e valência (b) via sinais audiovisuais obtidos com o melhor modelo de fusão tardia para um sujeito aleatório (# 9) da partição de teste.



(a) fluxo de informações (b) modelo

Fig. 4. Porcentagem da contribuição de cada fluxo de informação (a) ou modelo (b) para alcançar as melhores previsões de excitação ou valência.

Para analisar a importância de cada modalidade e modelo, calculamos suas contribuições para as previsões de ativação e valência dos respectivos modelos de melhor desempenho.
A Fig. 4 mostra suas contribuições. Para predição de excitação, as características acústicas desempenham um papel mais importante do que as características visuais, enquanto o oposto acontece para a predição de valência. Também é esperado que os sistemas DDAT baseados em RE e PU contribuam mais do que os sistemas de linha de base para as previsões finais. Além disso, o sistema DDAT baseado em PU é um pouco mais importante para a previsão de valência do que para a previsão de excitação. Isso pode ser devido ao fato de que a previsão da valência emocional é muito mais difícil do que a excitação para a modalidade de áudio [2], [67], [68].

V. CONCLUSÕES E TRABALHOS FUTUROS

Em contraste com estudos anteriores que visavam explorar a 'força' ou superar a 'fraqueza' da modelagem, pela primeira vez investigamos a exploração da informação de dificuldade (fraqueza) diretamente no processo de aprendizagem para a previsão contínua de emoções. Para extrair as informações de dificuldade, propusemos duas estratégias baseadas na ontologia da modelagem ou no conteúdo a ser modelado. Os dois tipos de informação medem separadamente a dificuldade de aprendizagem de um modelo, reconstruindo sua entrada, ou a 'dureza' dos dados a serem aprendidos, prevendo sua incerteza de percepção. Essas informações indicadas por um índice foram então concatenadas nos recursos originais para atualizar as entradas.

Os métodos propostos foram sistematicamente avaliados em um banco de dados de referência RECOLA [48]. Os resultados experimentais demonstraram que os métodos propostos melhoram claramente o desempenho de predição de um modelo, envolvendo a informação de dificuldade em seu processo de aprendizagem.

Indo além da aprendizagem curricular tradicional e abordagens de reforço que são especificamente projetadas para tarefas de reconhecimento de padrões discretos, as abordagens propostas de Treinamento de Conscientização Dinâmica da Dificuldade (DDAT) podem aprender particularmente bem o padrão sequencial, como a previsão contínua de emoção neste artigo. Ao envolver as informações de erro de reconstrução de entrada ou as informações de incerteza de percepção de emoção, descobrimos que as redes neurais podem ter um desempenho melhor. No entanto, vale a pena notar que a incerteza da percepção é meramente definida para um padrão subjetivo

TABELA IV

DESEMPENHO DE FUSÃO TARDIA (CCC) EM DIFERENTES ESTRATÉGIAS DE FUSÃO (I.E., BASEADO EM MODALIDADE , BASEADO EM MODALIDADE E MODELO E MODALIDADE DINAMICAMENTE SINTONIZADO E BASEADO EM MODELO) PARA O DESENVOLVIMENTO E PARTIÇÕES DE TESTE DE REGRESSÃO DE AROUSAL E DE VALÊNCIA . AS PREDIÇÕES SÃO GERADAS A PARTIR DA ESTRUTURA DDAT BASEADA EM ERROS DE RECONSTRUÇÃO (Pré) OU A ESTRUTURA DDAT BASEADA EM PERCEPÇÃO E INCERTEZA (Ppu); SUAS VERSÕES AJUSTADAS DINAMICAMENTE (Pre,dt OU Ppu,dt); OU O MODELO DE LINHA DE BASE (Pbs). OS MELHORES RESULTADOS ALCANÇADOS NO CONJUNTO DE TESTE ESTÃO EM NEGRITO. OBSERVE QUE Pre, Pre,dt , Ppu, Ppu,dt E Pbs SÃO AS PREDIÇÕES FUNDIDAS DE DIVERSOS FLUXOS DE INFORMAÇÃO (I. E., AUDIO-EGEMAPS, VIDEO-APARÊNCIA E VIDEO-GEOMÉTRICO). AS 1ª-3ª, 4ª-5ª e 6ª-8ª LINHAS DE RESULTADOS SÃO OBTIDAS RESPECTIVAMENTE DA MODALIDADE, MODALIDADE E MODELO E MODALIDADE dinamicamente sintonizada- E MODELO-BASE DE FUSÃO LATINA ESTRATÉGIAS.

várias abordagens de fusão tardia	aro			val		
	desenvolvedor		teste	desenvolvedor		teste
Pre Pre,dt Ppu Ppu,dt Pbs dev						
baseado em modalidade	.822	.690	.705 .584	.853	.763	.738 .615 .838 .715 .738 .615
baseado em modalidade e modelo	.860	.761	.755 .639	.864	.752	.766 .653
baseado em modalidade e modelo (ajustado dinamicamente)	.853	.761	.739 .621	.819	.721	.733 .631 .856 .766 **.756** .651 .863 .754 .766 **.660**
modelagem de						
força de última geração [16]	.808	.685	.671 .554	.731	.714	.502 .612
de ponta a ponta [14]						
estado da arte (+ fisiologia)						
seleção de recurso + deslocamento [49] –	.824	.747	.688 .609	.820	.702	.682 .638 .862 .770 .750 .687
SVR + compensação [48] [b]						
SC + CNN + LSTM [40] [c]						

– Vencedor AVEC '15
[b] linha de base AVEC '16
[c] Vencedor AVEC '16

tarefa de reconhecimento. Para uma tarefa objetiva, pode ser razoável empregar alternativamente a incerteza de previsão.

No futuro, continuaremos investigando a eficiência do DDAT proposto em predições de padrões discretos. Além disso, investigaremos as abordagens para as quais as informações de dificuldade poderiam ser usadas como pesos de predição. Além disso, estruturas ponta a ponta projetadas para extrair representações automaticamente têm atraído cada vez mais atenção e estão começando a mostrar um desempenho promissor. Portanto, uma estrutura avançada de ponta a ponta também será considerada em nosso sistema. Mais detalhadamente, com relação ao sistema DDAT fim-a-fim baseado em incerteza de percepção, podemos simplesmente substituir os GRN-RNNs por uma rede fim-a-fim enquanto todas as outras entradas e saídas permanecem. Com relação ao sistema ponta a ponta baseado em erro de reconstrução, podemos considerar reconstruir as representações de alto nível em vez dos sinais brutos ao extrair as informações de erro de reconstrução (ou seja, o indicador de dificuldade).

RECONHECIMENTO

Este trabalho foi apoiado por uma plataforma TransAtlantic Subsídio de colaboração “Digging into Data” (ACLEW: Analyzing Experiências de linguagem infantil em todo o mundo), com o apoio do Conselho de Pesquisa Econômica e Social do Reino Unido por meio da bolsa de pesquisa nº HJ-253479 (ACLEW) e do Programa Horizonte 2020 da União Europeia por meio da ação de pesquisa e inovação nº 645094 (SEWA) e nº 645378 (ARIA-VALUSPA).

REFERÊNCIAS

[1] M. Wollmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie e R. Cowie, “Abandoning Emoction Classs – Rumo ao reconhecimento contínuo de emoções com modelagem de dependências de longo alcance ”, no Proc. INTERSPEECH, Brisbane, Austrália, 2008, pp. 597–600.

[2] H. Gunes e M. Pantic, “Reconhecimento automático, dimensional e contínuo de emoções,” International Journal of Synthetic Emotions, vol. 1, não. 1, pp. 68–99, janeiro de 2010.

[3] H. Gunes, B. Schuller, M. Pantic e R. Cowie, “Representação, análise e síntese de emoções no espaço contínuo: uma pesquisa”, em Proc. Conferência Internacional sobre Reconhecimento Automático de Gestos e Faces (FG), Santa Bárbara, CA, 2011, pp. 827–834.

[4] RW Picard, Computação afetiva. Cambridge, MA: Imprensa do MIT, 1997.

[5] Y.-H. Yang e J.-Y. Liu, “Estudo quantitativo do comportamento de ouvir música em um contexto social e afetivo”, IEEE Transactions on Multimedia, vol. 15, não. 6, pp. 1304–1315, outubro de 2013.

[6] S. Zhao, H. Yao, Y. Gao, R. Ji e G. Ding, "Previsão de distribuição de probabilidade contínua de emoções de imagem por meio de regressão esparsa compartilhada multitarefa", IEEE Transactions on Multimedia, vol. 19, não. 3, pp. 632– 645, março de 2017.

[7] Z. Zhang, J. Han, X. Xu, J. Deng, F. Ringeval e B. Schuller, “Alavancando dados não rotulados para reconhecimento de emoções com aprendizado semi-supervisionado colaborativo aprimorado,” IEEE Access, vol. 6, não. 1, pp. 22 196–22 209, dezembro de 2018.

[8] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie e M. Pantic, “AVEC 2011 – O primeiro desafio internacional de emoção audiovisual”, em Proc. a 4ª Conferência Internacional sobre Computação Afetiva e Interação Inteligente (ACII), Memphis, TN, 2011, pp. 415–424.

[9] Q. Mao, M. Dong, Z. Huang e Y. Zhan, “Aprendendo recursos salientes para reconhecimento de emoções de fala usando redes neurais convolucionais,” IEEE Transactions on Multimedia, vol. 16, não. 8, pp. 2203–2213, dezembro de 2014.

[10] H. Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng e J. Cos mas, “Rede neural com atraso de tempo para previsão de dimensão emocional contínua a partir de sequências de expressão facial,” IEEE Transactions on Cybernetics, vol . 46, nº. 4, pp. 916–929, abril de 2016.

[11] R. Xia e Y. Liu, “Uma estrutura de aprendizado multitarefa para reconhecimento de emoções usando espaço contínuo 2D,” IEEE Transactions on Affective Computing, vol. 8, não. 1, pp. 3–14, janeiro de 2017.

[12] S. Zhang, S. Zhang, T. Huang e W. Gao, “Reconhecimento de emoção de fala usando rede neural convolucional profunda e correspondência de pirâmide temporal discriminante,” IEEE Transactions on Multimedia, vol. PP, não. 99, pp. 1–1, 2017.

[13] H. Li, J. Sun, Z. Xu e L. Chen, “Reconhecimento de expressão facial multimodal 2D+3D com rede neural convolucional de fusão profunda,” IEEE Transactions on Multimedia, vol. 19, não. 12, pp. 2816–2831, dezembro de 2017.

[14] P. Tzirakis, G. Trigeorgis, MA Nicolaou, B. Schuller e S. Zafeiriou, "Reconhecimento de emoção multimodal de ponta a ponta usando redes neurais profundas", IEEE Journal of Selected Topics in Signal Processing, edição especial sobre processamento de fala e linguagem de ponta a ponta, vol. 11, não. 8, pp. 1301–1309, dezembro de 2017.

[15] Z. Zeng, M. Pantic, GI Roisman e TS Huang, “Uma pesquisa de métodos de reconhecimento de afeto: expressões de áudio, visuais e espontâneas,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, nº. 1, pp. 39–58, março de 2009.

[16] J. Han, Z. Zhang, N. Cummins, F. Ringeval e B. Schuller, "Modelagem de força para reconhecimento automático de afeto contínuo do mundo real a partir de sinais audiovisuais", Image and Vision Computing, vol. 65, pp. 76–86, setembro de 2017.

[17] Z. Zhang, F. Eyben, J. Deng e B. Schuller, “Uma seleção de instâncias de aprendizagem baseada em concordância e esparsidade e sua aplicação a fenômenos de fala subjetivos”, em Proc. o 5º Workshop Internacional sobre Sinais Sociais de Emoção, Sentimento e Dados Abertos Vinculados, satélite do LREC, Reykjavik, Islândia, 2014, pp. 21–26.

[18] L. Gui, T. Baltruaitis e LP Morency, "Aprendizagem curricular para reconhecimento de expressões faciais", em Proc. 12ª Conferência Internacional IEEE sobre Reconhecimento Automático de Gestos Faciais (FG), Washington, DC, 2017, pp. 505–511.

[19] P. Liu, S. Han, Z. Meng e Y. Tong, “Reconhecimento de expressão facial por meio de uma rede de crença profunda impulsionada”, em Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, 2014, pp. 1805–1812.

[20] LL Presti e ML Cascia, “Boosting Hankel Matrixes for Face Emoction Recognization and Pain Detection,” Computer Vision and Image Understanding, vol. 156, pp. 19–33, março de 2017.

[21] Y. Bengio, J. Louradour, R. Collobert e J. Weston, “Curriculum learning,” in Proc. 26ª Conferência Internacional Anual sobre Machine Learning (ICML), Montreal, Canadá, 2009, pp. 41–48.

[22] S. Braun, D. Neil e S. Liu, “Um método de aprendizado curricular para robustez de ruído aprimorada no reconhecimento automático de fala”, em Proc. 25ª Conferência Europeia de Processamento de Sinais, (EUSIPCO), Kos, Grécia, 2017, pp. 548–552.

[23] R. Lotfian e C. Busso, “Aprendizagem curricular para reconhecimento de emoções de fala a partir de rótulos de crowdsourcing”, arXiv preprint arXiv:1805.10339, maio de 2018.

[24] A. Graves, Rotulagem de sequência supervisionada com redes neurais recorrentes. Berlim/Heidelberg, Alemanha: Springer, 2012.

[25] MI Posner e SE Petersen, “O sistema de atenção do cérebro humano,” Revisão Anual de Neurociência, vol. 13, não. 1, pp. 25–42, 1990.

[26] DA Washburn e R. Putney, “Atenção e dificuldade da tarefa: quando o desempenho é facilitado?” Aprendizagem e Motivação, vol. 32, nº. 1, pp. 36–47, fevereiro de 2001.

[27] ML Seltzer, D. Yu e Y. Wang, "Uma investigação de redes neurais profundas funciona para reconhecimento de fala robusto a ruído", em Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canadá, 2013, pp. 7398–7402.

[28] P. Karanasou, Y. Wang, MJ Gales e PC Woodland, "Adaptação de modelos acústicos de redes neurais profundas usando i-vetores fatorados". em Proc. INTERSPEECH, Cingapura, 2014, pp. 2180–2184.

[29] P. Vincent, H. Larochelle, Y. Bengio e P. Manzagol, "Extraindo e compondo recursos robustos com codificadores automáticos de redução de ruído", em Proc. Conferência Internacional sobre Machine Learning (ICML), Helsinki, Finlândia, 2008, pp. 1096–1103.

[30] E. Marchi, F. Vesperini, F. Eyben, S. Squartini e B. Schuller, "Uma nova abordagem para detecção acústica automática de novidade usando um codificador automático de redução de ruído com redes neurais LSTM bidirecionais", em Proc. Conferência Internacional IEEE sobre Processamento de Áudio, Fala e Sinal (ICASSP), Brisbane, Austrália, 2015, pp. 1996–2000.

[31] H. Yang, B. Wang, S. Lin, D. Wipf, M. Guo e B. Guo, “Extração não supervisionada de destaques de vídeo por meio de codificadores automáticos recorrentes robustos”, em Proc. IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 4633–4641.

[32] S. Petridis e M. Pantic, "Fusão audiovisual baseada em previsão para classificação de vocalizações não lingüísticas," IEEE Transactions on Affective Computing, vol. 7, não. 1, pp. 45–58, janeiro de 2016.

[33] Z. Zhang, N. Cummins e B. Schuller, "Exploração avançada de dados para análise de fala - uma visão geral", IEEE Signal Processing Magazine, vol. 34, nº. 4, pp. 107–129, julho de 2017.

[34] L. Devillers, L. Vidrascu e L. Lamel, "Desafios na anotação de emoções da vida real e detecção baseada em aprendizado de máquina," Redes neurais, vol. 18, não. 4, pp. 407 – 422, maio de 2005.

[35] J. Deng, W. Han e B. Schuller, "Medidas de confiança para reconhecimento de emoções de fala: um começo", em Proc. a 10ª Conferência ITG sobre Comunicação de Fala, Braunschweig, Alemanha, 2012, pp. 1–4.

[36] T. Dang, V. Sethu, J. Epps e E. Ambikairajah, "Uma investigação da incerteza de previsão de emoção usando regressão de mistura gaussiana", em Proc. INTERSPEECH, Estocolmo, Suécia, 2017, pp. 1248–1252.

[37] T. Dang, B. Stasak, Z. Huang, S. Jayawardena, M. Atcheson, M. Hayat, P. Le, V. Sethu, R. Goecke e J. Epps, “Investigando palavras afetam recursos e fusão de previsões probabilísticas incorporando incerteza no AVEC 2017”, no Proc. 7º Workshop Anual sobre Audio/Visual Emotion Challenge (AVEC), Mountain View, CA, 2017, pp. 27–35.

[38] J. Han, Z. Zhang, M. Schmitt e B. Schuller, "From hard to soft: Rumo a um reconhecimento de emoções mais humano, modelando a incerteza da percepção", em Proc. ACM International Conference on Multimedia (MM), Mountain View, CA, 2017, pp. 890–897.

[39] J. Han, Z. Zhang, F. Ringeval e B. Schuller, "Aprendizagem baseada em erros de reconstrução para reconhecimento contínuo de emoções na fala", em Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 2367–2371.

[40] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, WM Campbell, CK Dagli e TS Huang, “Aprendizagem multimodal de áudio, vídeo e sensor fisiológico para previsão contínua de emoções”, em Proc. o 6º Workshop Internacional sobre Audio/Visual Emotion Challenge (AVEC), Amsterdã, Holanda, 2016, pp. 97–104.

[41] S. Hochreiter e J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, não. 8, pp. 1735–1780, novembro de 1997.

[42] X. Ma, H. Yang, Q. Chen, D. Huang e Y. Wang, "DepAudioNet: um modelo profundo eficiente para classificação de depressão baseada em áudio", em Proc. o 6º Workshop Internacional sobre Audio/Visual Emotion Challenge (AVEC), Amsterdã, Holanda, 2016, pp. 35–42.

[43] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J. Thiran, T. Ebrahimi, D. Lalanne e B. Schuller, "Previsão de classificações de emoções dimensionais assíncronas a partir de dados audiovisuais e fisiológicos", Cartas de Reconhecimento de Padrões, vol. 66, pp. 22–30, novembro de 2015.

[44] L. Chao, J. Tao, M. Yang, Y. Li e Z. Wen, “Reconhecimento de emoções multimodais multimodais baseadas em rede neural recorrente de memória de longo prazo”, em Proc. o 5º Workshop Internacional sobre Audio/Visual Emotion Challenge (AVEC), Brisbane, Austrália, 2015, pp. 65–72.

[45] Z. Huang, T. Dang, N. Cummins, B. Stasak, P. Le, V. Sethu e J. Epps, "Uma investigação da compensação de atraso de anotação e fusão associativa de saída para previsão de emoção contínua multimodal," em Proc. the 5th International Workshop on Audio/Visual Emotion Chal lenge (AVEC), Brisbane, Austrália, 2015, pp. 41–48.

[46] J. Wei, E. Pei, D. Jiang, H. Sahli, L. Xie e Z. Fu, “Reconhecimento de afeto contínuo multimodal baseado em LSTM e aprendizado de kernel múltiplo”, em Proc. Cúpula e Conferência Anual da Associação de Processamento de Sinais e Informações da Ásia-Pacífico (APSIPA), Siem Reap, Camboja, 2014, pp. 1–4.

[47] X. Qiu, L. Zhang, Y. Ren, PN Suganthan e G. Amaratunga, "Ensemble deep learning para regressão e previsão de séries temporais", em Proc. Computational Intelligence in Ensemble Learning (CIEL), Orlando, FL, 2014, pp. 1–6.

[48] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Tor res Torres, S. Scherer, G. Stratou, R. Cowie e M. Pantic, “AVEC 2016: Oficina e desafio de reconhecimento de depressão, humor e emoção”, in Proc. o 6º Workshop Internacional sobre Audio/Visual Emotion Challenge (AVEC), Amsterdã, Holanda, 2016, pp. 3–10.

[49] L. He, D. Jiang, L. Yang, E. Pei, P. Wu e H. Sahli, "Previsão de dimensão afetiva multimodal usando redes neurais recorrentes de memória de curto prazo bidirecional profunda", em Proc. o 5º Workshop Internacional sobre Audio/Visual Emotion Challenge (AVEC), Brisbane, Austrália, 2015, pp. 73–80.

[50] M. Nicolaou, H. Gunes e M. Pantic, "Previsão contínua de afeto espontâneo a partir de várias pistas e modalidades no espaço de ativação de valência", IEEE Transactions on Affective Computing, vol. 2, não. 2, pp. 92–105, abril de 2011.

[51] J. Deng, X. Xu, Z. Zhang, S. Fruhholz, e B. Schuller, "Semi-supervisioned autoencoders for Speech Emoction Recognition," IEEE/ACM Transaction on Audio, Speech, and Language Processing, vol. 26, não. 1, pp. 31–43, janeiro de 2018.

[52] MA Nicolaou, H. Gunes e M. Pantic, "Regressão RVM associativa de saída para previsão de emoção dimensional e contínua," Image and Vision Computing, vol. 30, não. 3, pp. 186–196, março de 2012.

[53] S. Parthasarathy e C. Busso, "Prevendo conjuntamente excitação, valência e dominância com aprendizado multitarefa", em Proc. INTERSPEECH, Estocolmo, Suécia, 2017, pp. 1103–1107.

[54] Y. Xia, X. Cao, F. Wen, G. Hua e J. Sun, “Aprendendo reconstruções discriminativas para remoção de valores discrepantes não supervisionados”, em Proc. IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1511–1519.

[55] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, e B. Schuller, “Facing realism in espontâneo espontaneo reconhecimento de fala: Aprimoramento de recursos por autoencoder com redes neurais LSTM,” em Proc. INTERSPEECH, San Francisco, CA, 2016, pp. 3593–3597.

[56] B. Schuller e A. Batliner, Paralinguística computacional: emoção, afeto e personalidade no processamento de fala e linguagem. Hoboken, NJ: John Wiley & Sons, 2013.

[57] IB Mauss e MD Robinson, “Medidas de emoção: uma revisão,” Cognição e Emoção, vol. 23, não. 2, pp. 209–237, fevereiro de 2009.

[58] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie e M. Pantic, “AV+EC 2015: O primeiro desafio de reconhecimento de afeto fazendo a ponte entre o áudio , vídeo e dados fisiológicos”, no Proc. o 5º Workshop Internacional sobre Audio/Visual Emotion Challenge (AVEC), Brisbane, Austrália, 2015, pp. 3–8.

[59] F. Ringeval, A. Sonderegger, JS Sauer e D. Lalanne, “Apresentando o corpus multimodal RECOLA de colaboração remota e afetividade

interações interativas”, em Proc. 10ª Conferência e Workshops Internacionais IEEE sobre Reconhecimento Automático de Gestos e Faces (FG), Xangai, China, 2013, pp. 1–8.

[60] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andre, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan e K. Truong, “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) para pesquisa de voz e computação afetiva,” IEEE Transactions on Affective Computing, vol. 7, não. 2, pp. 190–202, abril de 2016.

[61] K. Cho, B. Van Merrienboer, D. Bahdanau e Y. Bengio, “Sobre as propriedades da tradução automática neural: abordagens do codificador-decodificador”, em Proc. Workshop sobre Sintaxe, Semântica e Estrutura na Tradução Estatística (SSST), Doha, Catar, 2014, pp. 103–111.

[62] R. Jozefowicz, W. Zaremba e I. Sutskever, "Uma exploração empírica de arquiteturas de rede recorrentes", em Proc. Conferência Internacional sobre Machine Learning (ICML), Lille, França, 2015, pp. 2342–2350.

[63] DP Kingma e J. Ba, "Adam: Um método para otimização estocástica", em Proc. Conferência Internacional sobre Representações de Aprendizagem (ICLR), San Diego, CA, 2015, 15 páginas.

[64] S. Mariooryad e C. Busso, "Corrigindo rótulos emocionais contínuos no tempo modelando o atraso de reação dos avaliadores", IEEE Transactions on Affective Computing, vol. 6, não. 2, pp. 97–108, abril de 2015.

[65] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmi e M. Pantic, “AVEC 2017–Real life depressão e oficina de reconhecimento de afeto e desafio”, em Proc. o 7º Workshop Internacional sobre Audio/Visual Emotion Challenge (AVEC), Mountain View, CA, 2017, pp. 3–10.

[66] J. Cohen, P. Cohen, SG West e LS Aiken, Análise de regressão/correlação múltipla aplicada para as ciências comportamentais. Abingdon, Reino Unido: Routledge, 2013.

[67] Z. Zhang, E. Coutinho, J. Deng e B. Schuller, “Aprendizado cooperativo e sua aplicação ao reconhecimento de emoções a partir da fala,” IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 23, não. 1, pp. 115–126, janeiro de 2015.

[68] E. Coutinho e B. Schuller, “Códigos acústicos compartilhados fundamentam a comunicação emocional na música e na fala – Evidências da aprendizagem de transferência profunda,” PloS One, vol. 13, não. 1, pág. e0191754, 2018.

Zixing Zhang (M'15) concluiu seu mestrado em eletrônica física pela Universidade de Correios e Telecomunicações de Pequim (BUPT), China, em 2010, e seu doutorado em engenharia da computação pela Universidade Técnica de Munique (TUM), Alemanha, em 2015. Atualmente é pesquisador associado do Departamento de Computação do Imperial College London (ICL), Reino Unido, desde 2017.

Antes disso, ele foi pesquisador de pós-doutorado na Universidade de Passau, Alemanha, de 2015 a 2017. Ele é autor de cerca de setenta publicações em livros revisados por pares, periódicos e anais de conferências até o momento e organizou sessões especiais, como em o IEEE 7th Affective Computing and Intelligent Interaction (ACII) em 2017 e na 43ª IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) em 2018.

O Dr. Zhang atua como revisor de periódicos líderes em suas áreas, como IEEE T-NNLS, IEEE T-CYB, IEEE T-AC, IEEE T-MM, IEEE T-ASLP, Speech Communication e Computer Speech & Language. Seus interesses de pesquisa estão em aprendizado profundo, aprendizado supervisionado semanalmente e aprendizado de transferência para análise de fala inteligente e robusta, como reconhecimento de emoções.

Jing Han (S'16) recebeu seu diploma de bacharel (2011) em engenharia eletrônica e da informação pela Harbin Engineering University (HEU), China, e seu mestrado (2014) pela Nanyang Technological University, Cingapura. Ela agora está trabalhando como estudante de doutorado com a ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing na Universidade de Augsburg, Alemanha, envolvida no programa Horizon 2020 da UE SEWA. Ela revisa regularmente IEEE Transactions on Cybernetics e IEEE Signal Processing Letters. Seus interesses de pesquisa estão relacionados ao aprendizado profundo para computação afetiva multimodal e saúde

Cuidado.

Eduardo Coutinho graduou-se pela Universidade do Porto (Portugal) em 2003 e concluiu seu doutorado em Ciências da Computação e Afetivas em 2009. Desde então, Coutinho trabalhou nas áreas interdisciplinares de psicologia musical e ciências afetivas na Universidade de Sheffield (Reino Unido), o Centro Suíço de Ciências Afetivas (Suíça), a Universidade Técnica de Munique (Alemanha) e o Imperial Col lege London (Reino Unido). Atualmente, Coutinho é professor de Psicologia Musical na Universidade de Liverpool (Reino Unido) e Pesquisador Associado na ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing na Universidade de Augsburg (Alemanha). Na sua investigação, Coutinho centra-se no estudo das experiências emocionais com a música e nas ligações entre a comunicação da emoção pela música e o tom de voz. Atualmente, também se dedica ao desenvolvimento de métodos e ferramentas que permitem o uso da música para melhorar diversos aspectos do bem-estar no dia a dia. Em 2013, ele recebeu o Prêmio de Transferência de Conhecimento do Centro Nacional Suíço de Competência em Pesquisa em Ciências Afetivas e, em 2014, recebeu o Prêmio Jovem Investigador da International Neural Network Society.

Bjorn Schuller ¨ (M'05-SM'15-F'18) recebeu seu diploma em 1999, seu doutorado por seu estudo sobre Reconhecimento Automático de Fala e Emoção em 2006, e sua habilitação e Professor Adjunto na área de Signal Processing and Machine Intelligence em 2012, tudo em engenharia elétrica e tecnologia da informação pela TUM em Munique/Alemanha. É professor de Machine Learning no Departamento de Computação do Imperial College London/UK, onde dirige o GLAM – the Group on Language, Audio & Music, Professor Titular e chefe da ZD.B Chair of Embedded Intelligence for Health Care and Bem-estar pela Universidade de Augsburg/ Alemanha e CEO da audEERING. Anteriormente, foi professor titular e chefe da Cátedra de Sistemas Complexos e Inteligentes da Universidade de Passau/Alemanha. O professor Schuller é presidente emérito da Associação para o Avanço da Computação Afetiva (AAAC), membro eleito do Comitê Técnico de Processamento de Fala e Linguagem do IEEE, membro do IEEE e membro sênior do ACM. Ele (co-)autor de 5 livros e mais de 700 publicações em livros revisados por pares, periódicos e anais de conferências, levando a mais de 19.000 citações (h-index = 66). Schuller é co-presidente do programa Interspeech 2019, presidente de área repetido do ICASSP e editor-chefe do IEEE Transactions on Affective Computing, além de uma infinidade de outras funções e funções de editor associado e convidado em comitês técnicos e organizacionais.