Rumo à observação automatizada da sala de aula: máquina multimodal Aprendendo a estimar o clima positivo e o clima negativo da CLASSE

Anand Ramakrishnan1, Brian Zylich1 , Erin Ottmar1, Jennifer LoCasale-Crouch2 , e Jacob Whitehill1

[1] Instituto Politécnico de Worcester, Worcester, MA, EUA

[2] Universidade da Virgínia, Charlottesville, VA, EUA

Resumo—Neste trabalho apresentamos um sistema multimodal baseado em aprendizado de máquina, que chamamos de ACORN, para analisar vídeos de salas de aula escolares para as dimensões Clima Positivo (PC) e Clima Negativo (NC) do protocolo de observação CLASS [1] que é amplamente utilizado em pesquisas educacionais. ACORN usa redes neurais convolucionais para analisar características espectrais de áudio, os rostos de professores e alunos e os pixels de cada quadro de imagem e, em seguida, integra essas informações ao longo do tempo usando Redes Convolucionais Temporais.

As previsões audiovisuais de PC e NC da ACORN têm correlações de Pearson de 0,55 e 0,63 com pontuações de verdade fornecidas por codificadores CLASS especializados no conjunto de dados UVA Toddler (validação cruzada em n = 300 segmentos de vídeo de 15 minutos), e uma previsão ACORN puramente auditiva PC e NC com correlações de 0,36 e 0,41 no conjunto de dados MET (conjunto de teste de n = 2.000 segmentos de vídeos). Esses números são semelhantes à confiabilidade entre codificadores de codificadores humanos. Finalmente, usando Redes Convolucionais de Gráficos, damos passos iniciais (AUC = 0,70) para prever os momentos específicos (clipes de 45 a 90 segundos) quando o PC está particularmente fraco/forte. Nossas descobertas informam o projeto de observação automática de sala de aula e também sistemas mais gerais de reconhecimento de atividades de vídeo e reconhecimento de resumo.

Termos **de Indexação** —observação automática de sala de aula, Sistema de Pontuação de Avaliação de Sala de Aula, reconhecimento de expressão facial, análise auditiva

ÿ

1. INTRODUÇÃO

A qualidade das interações professor-aluno e aluno-aluno nas salas de aula prevê e impacta os resultados de aprendizagem dos alunos. Numerosos estudos correlacionais [2], [3], [4], [5], [6] e alguns causais em grande escala [7], [8] demonstraram a ligação entre o apoio emocional e instrucional na sala de aula e o apoio posterior das crianças. habilidades cognitivas, sociais e emocionais. Para caracterizar com precisão as interações em sala de aula, os pesquisadores educacionais desenvolveram uma variedade de protocolos de observação em sala de aula. Um dos protocolos mais utilizados é o Classroom Assessment Scoring System [1] (CLASS). Uma sessão típica de observação CLASSE requer anotadores humanos – que podem ser professores, investigadores educacionais ou administradores escolares – para examinar características específicas dos estados, ações e interações entre os alunos e professores durante a observação ao vivo ou vídeos gravados.

Embora a codificação CLASS seja uma ferramenta valiosa para a investigação educacional e a formação de professores, a sua utilidade é limitada pelas dificuldades da codificação manual: a codificação humana das pontuações CLASS requer formação significativa, é lenta e dispendiosa e pode sofrer de uma variabilidade significativa entre codificadores. . Por outro lado, o sucesso dos métodos contemporâneos de aprendizagem profunda para reconhecimento de objetos, reconhecimento de emoções e análise de fala, bem como métodos multimodais para reconhecimento de atividades e análise de vídeo, levanta a questão: aspectos específicos da observação de sala de aula poderiam ser realizados por uma máquina? e/ou as ferramentas perceptivas automatizadas poderiam ajudar os anotadores humanos na codificação de vídeos em sala de aula?

Aprendizado de máquina para medição educacional: Nos últimos dez anos, houve um aumento no interesse em aproveitar o aprendizado de máquina para desenvolver novas ferramentas para medição educacional (veja a seção Trabalhos Relacionados abaixo). A maior parte deste trabalho concentrou-se na análise do envolvimento e das emoções individuais dos alunos [9], [10], [11], [12], ou na classificação das ações pedagógicas dos professores a partir de seu discurso [13]. Durante os últimos anos, tem havido um interesse crescente na análise de toda a sala de aula a partir de vídeos [14], [15], [16]. Neste artigo, desenvolvemos nosso trabalho piloto [15] e exploramos uma variedade de métodos de aprendizagem profunda multimodais (visão, audição, linguagem) para estimar automaticamente as dimensões do Clima Positivo (PC) e do Clima Negativo (NC) da CLASSE a partir de vídeos de sala de aula. Esses vídeos (ver Figura 2, por exemplo) apresentam numerosos e graves desafios tanto para a visão computacional quanto para a análise de áudio, incluindo fala ruidosa e sobreposta, crianças muito pequenas cuja fala é pronunciada de forma imprecisa, postura extrema da cabeça, oclusão visual, iluminação descontrolada, e fundos visualmente complicados. Dados esses desafios, identificamos arquiteturas promissoras para percepção de baixo nível de características visuais e auditivas, bem como projetos de integração temporal de alto nível para estimar pontuações CLASS. Chamamos nosso sistema final de ACORN (Automatic Classroom Observation Recognition Network) e o validamos em dois conjuntos de dados codificados por CLASS (UVA Toddler e MET). Embora o foco de aplicação do nosso artigo seja a medição educacional, nossos resultados também têm implicações para outros problemas de computação afetiva, análise de vídeo e reconhecimento de atividades, especialmente quando a variável alvo é semanticamente de “alto nível”, como em nosso ambiente.

No início deste projeto de pesquisa (WPI IRB #17-151), não estava claro para nós se construções semanticamente de alto nível como CLASSE Clima Positivo e Clima Negativo poderiam ser estimadas por uma máquina com algum grau de precisão. Através de um processo de design iterativo, aproveitando técnicas contemporâneas de visão computacional e análise de fala, e projetando novas arquiteturas de integração de informações e procedimentos de treinamento, conseguimos aumentar a precisão de forma constante para igualar (e possivelmente exceder) a confiabilidade intercodificador dos codificadores CLASS humanos. Este artigo compartilha

Este material é baseado em trabalho apoiado pela National Science Foundation sob as bolsas nº 1551594 e nº 1822768, e pela Spencer Small Research Grant nº nº 201800131.

muitos dos insights de aprendizado de máquina multimodal que aprendemos ao longo do caminho.

1.1 Contribuições técnicas e novidades Estimativa

automatizada de CLASSE: O ACORN apresentado aqui é o primeiro sistema totalmente automatizado para estimar a partir de vídeos de sala de aula as dimensões da CLASSE e atinge uma precisão semelhante à de codificadores humanos. A análise de vídeos de sala de aula é um problema de reconhecimento de atividades de vídeo altamente desafiador: em contraste com grande parte da literatura anterior sobre reconhecimento de atividades [17], [18], [19], em que o intervalo temporal das atividades é geralmente de apenas alguns minutos (ou segundos) e são fáceis de serem percebidos por humanos comuns (por exemplo, “tirar da geladeira”), em nosso ambiente cada segmento de vídeo dura 15 minutos, e a tarefa perceptiva requer treinamento significativo (geralmente pelo menos várias semanas de prática para se tornar competente na codificação CLASSE). A nossa ACORN pode potencialmente servir como um instrumento científico para fornecer feedback aos professores e facilitar a investigação educacional. O trabalho aqui amplia significativamente nosso artigo anterior [15], aproveitando arquiteturas perceptivas audiovisuais mais poderosas para obter maior precisão (0,55 vs. 0,40 para PC, 0,63 vs. 0,51 para NC), identificando automaticamente os momentos mais importantes em um vídeo de sala de aula e avaliando em conjuntos de dados cada vez maiores. Nosso trabalho também se distingue de sistemas relacionados para análise automatizada de sala de aula de várias maneiras: Desenvolvemos modelos perceptivos para vídeo e áudio (em vez de apenas áudio [13]); nosso sistema estima um julgamento semântico de alto nível da dinâmica da sala de aula durante um longo período de tempo (em vez de focar no reconhecimento de comportamentos individuais de baixo nível [14]) e usamos redes neurais mais complexas em comparação com [20].

Reconhecimento de atividade humana a partir de vídeo – que detalhes importam?: Como um exemplo de pesquisa de reconhecimento de atividade humana em grupo, nosso trabalho fornece insights, por meio de uma sequência de experimentos controlados (modelos #1-#21 treinados e testados usando cruzamentos duplos em sala de aula). validação), em quais detalhes arquitetônicos (backbone CNN, atenção, integrador temporal, engenharia de recursos baseada em teoria) são importantes para capturar atributos semanticamente de alto nível em escalas de tempo relativamente longas (15min).
Este tipo de estudo empírico aprofunda a compreensão de quais melhorias de precisão dentro de um sistema complexo de aprendizado de máquina multimodal são aditivas e quais são incluídas em outras.
As tendências que identificamos são amplamente consistentes para as dimensões PC e NC da CLASSE e persistem mesmo após a reorganização aleatória das dobras de validação cruzada.

Convolução de grafos para detecção de eventos-chave - importância da topologia: Desenvolvemos uma abordagem de visão computacional - baseada em redes de convolução de grafos (GCN) [21] sobre um gráfico induzido pelas posições 2-d de faces detectadas, combinadas com atenção de gráfico e recorrente redes neurais – para identificar os momentos mais marcantes em um vídeo de sala de aula; este é um tipo de tarefa de detecção de eventos-chave e resumo de vídeo. Nossos resultados sugerem que esta abordagem oferece maior precisão do que vários outros métodos (por exemplo, o recentemente proposto modelo de destaque de vídeo siamês [22]). Além disso, conduzimos novas análises de ablação na topologia do gráfico para verificar se o benefício de nossa camada GCN deriva das interações entre alunos e professores vizinhos, e não apenas de ter outra camada não linear + pooling dentro de uma rede maior.

2. TRABALHO RELACIONADO

Pesquisadores da ciência da computação, ciência cognitiva e psicologia exploraram como usar o aprendizado de máquina para perceber alunos, professores e salas de aula há mais de 20 anos [23]. Este trabalho varia em diversas dimensões, incluindo o atributo alvo a ser previsto, sensores usados como entradas e abordagem algorítmica.

Atributos alvo: A maioria dos trabalhos na interseção da percepção da máquina e da educação tem se concentrado na caracterização automática dos estados afetivos dos alunos individuais, incluindo envolvimento [11], [24], [25], [26], concentração [27], [28], [29], frustração [25], [26], [30] e outras emoções de realização [31]. Isso pode ser útil para fornecer aos professores feedback em tempo real ou post-hoc sobre como os alunos respondem às suas instruções, ou como um sinal de recompensa em tempo real para sistemas de tutoria inteligentes [12], [32], [33] ou tutores robôs [24]. Alguns pesquisadores investigaram como identificar os comportamentos e estratégias pedagógicas dos professores [13], [16], [34], [35]. Finalmente, durante os últimos anos, também surgiram alguns projetos (incluindo o nosso) que analisam a dinâmica de uma sala de aula inteira, seja como uma coleção de alunos individuais [14] ou como uma medida agregada de muitos participantes interagindo [15], [36].

Isso pode fornecer dados brutos para painéis de professores e também facilitar a codificação automatizada de observação em sala de aula.

Sensores: Muitas abordagens usam visão computacional para analisar as expressões faciais, movimentos da cabeça e postura corporal dos alunos [11], [14], [20], [23], [24], [25], [26], [37]; esta linha de pesquisa decorre em grande parte do reconhecimento facial e de gestos, do aprendizado de máquina multimodal e das comunidades de computação afetiva. Outros analisam áudio e fala [13], [16], [34], [35], que são indiscutivelmente menos invasivos à privacidade do que a visão, para caracterizar o tipo de instrução usada em uma sala de aula em cada momento. Lá também são abordagens “sem sensores” [27], [28], [29], [38], muitas vezes lideradas por pesquisadores da comunidade de mineração de dados educacionais, que prevêem comportamentos ou emoções futuras dos alunos, analisando os arquivos de log gerados a partir de sistemas de tutoria inteligentes e cursos on-line abertos e massivos. Esses arquivos de log normalmente contêm um registro de todas as decisões que os alunos tomam (por exemplo, abrir um determinado módulo) ou respostas que eles dão em resposta a questões práticas. Finalmente, existem também alguns estudos que utilizam apenas texto [31], por exemplo, de fóruns de discussão online, para julgar as emoções dos alunos.

Algoritmos: Durante os últimos anos, surgiu uma série de ferramentas de software prontas para uso de alta qualidade para percepção visual automática, como OpenPose [39], OpenFace [40], bem como serviços baseados em nuvem para visão e análise de fala, como Amazon Rekognition e Google Cloud Speech. Esses sistemas geralmente são baseados em algoritmos de aprendizado profundo e presumivelmente treinados em conjuntos de dados muito grandes para produzir alta precisão. Portanto, é natural usá-los como motores de percepção de baixo nível que podem então ser processados para estimar atributos de nível superior [13], [14], [20], [24], [25], [37] . Por outro lado, tais sistemas e serviços não são adaptados à aprendizagem dos alunos ou aos ambientes de sala de aula, e é possível que modelos personalizados, treinados especificamente na população-alvo, possam funcionar melhor.

Conseqüentemente, muitos pesquisadores treinaram seus próprios sistemas de percepção personalizados [11], [15], [20], [34], [38].

2.1 Percepção da máquina nas salas de aula

Aqui resumimos brevemente os sistemas perceptivos baseados em aprendizado de máquina que analisam salas de aula inteiras. D'Mello et al. [35], [41] exploraram como segmentar e reconhecer a fala de alunos e professores em salas de aula sem restrições com base em diferentes configurações de microfone. Wang et al. [42] segmentaram a fala dos professores implantando pequenos dispositivos de gravação vestíveis em salas de aula de matemática. Ahuja, et al. [14] desenvolveram um hardware combinado

e um kit de ferramentas de software chamado EduSense, que detecta automaticamente os movimentos corporais e faciais dos alunos. Seu sistema usa OpenPose [39], bem como vários classificadores (florestas aleatórias, máquinas de vetores de suporte, perceptrons multicamadas) treinados em cima de suas saídas, para rastrear cada aluno em cada quadro de vídeo, bem como sua postura corporal, gestos manuais. e expressões faciais. Ele também analisa recursos de áudio gravados em diferentes microfones para determinar se a fala foi produzida pelos alunos ou pelos instrutores.

A arquitetura de aprendizado de máquina em [16] é baseada em um conjunto de árvores de decisão que analisam o volume e o desvio padrão do som da sala de aula em intervalos de 15 segundos, onde o objetivo é classificar as diferentes atividades da sala de aula.

Devido à sua popularidade na pesquisa educacional, recentemente alguns pesquisadores computacionais desenvolveram métodos para automatizar aspectos do Classroom Assessment Scoring System (CLASS).
O primeiro trabalho nesse sentido foi o de Qiao e Beling [36], que desenvolveram um sistema de visão computacional, otimizado dentro de uma estrutura de aprendizagem de múltiplas instâncias, para estimar quais clipes de 3 minutos em vídeos de sala de aula eram mais relevantes para os codificadores CLASS codificarem manualmente. .
Observe que 3 minutos é significativamente menor que o intervalo de anotação de 15 a 20 minutos prescrito pelo manual CLASS (consulte a Seção 3); isso ocorre porque seu sistema foi projetado para identificar os momentos-chave que justificavam uma inspeção humana mais detalhada, em vez de estimar as próprias pontuações CLASS.

James e outros. [37], [43] buscaram uma arquitetura semelhante ao nosso trabalho anterior [15] para reconhecimento automático de pontuações climáticas CLASS. No entanto, em contraste com a definição CLASS, que define o Clima Positivo e o Clima Negativo como dimensões independentes, o seu trabalho trata-os como dois lados de um espectro. [15] exploraram BiLSTMs que analisam recursos de expressão facial, bem como CNNs que analisam recursos de áudio de baixo nível. Comparado com [15], o presente trabalho explora arquiteturas mais poderosas e usa um conjunto de dados maior para obter uma precisão de previsão CLASS substancialmente maior.

3 SISTEMA DE PONTUAÇÃO DE AVALIAÇÃO EM SALA DE AULA

O Classroom Assessment Scoring System (CLASS) [1] é um protocolo de observação validado e amplamente utilizado [44] para medir a qualidade do ensino nas salas de aula escolares. Ao realizar a codificação CLASS, observadores humanos analisam as interações em sala de aula entre professores e alunos, e entre alunos e seus colegas, ao longo de 8 a 12 (o número varia dependendo da faixa etária) dimensões que são divididas em 2 a 4 domínios.

Por exemplo, para salas de aula para crianças pequenas, existem dois domínios: (1) Apoio Emocional e Comportamental, com 5 dimensões: Clima Positivo, Clima Negativo, Sensibilidade do Professor, Respeito pelas Perspectivas da Criança e Orientação Comportamental; e (2) Apoio Engajado à Aprendizagem, com 3 dimensões: Facilitação da Aprendizagem e Desenvolvimento, Qualidade do Feedback e Modelagem da Linguagem. Uma única pontuação em uma escala inteira de 1 a 7 é atribuída a cada dimensão com base na observação de uma porção de 15 minutos de instrução em sala de aula.
As pontuações CLASS de codificadores humanos especialistas demonstraram prever uma variedade de resultados educacionais e sociocomportamentais posteriores [2], [3], [4], [44].

Dentro do domínio de apoio emocional da CLASSE, duas dimensões são o **Clima Positivo (CP)** que mede o “calor, respeito e prazer comunicados por interações verbais e não-verbais” entre alunos e professores; e o **Clima Negativo (NC)** que mede o “nível geral de ex-

Clima Positivo	
Indicadores	**Marcadores Comportamentais**
Relacionamentos	Proximidade física, afeto correspondente
Afeto positivo	Sorrir, rir, elogios apropriados
Respeito	Contato visual, voz calorosa, linguagem de apoio

Clima Negativo	
Indicadores	**Marcadores Comportamentais**
Afeto Negativo	Irritabilidade, voz áspera, raiva
Controle Punitivo	Gritos, ameaças
Professor Negatividade	Voz sarcástica, humilhação
Negatividade infantil	Vitimização, bullying

TABELA 1: A CLASSE Clima Positivo e Negativo conforme apresentado em [1]. Cada Clima é subdefinido em termos de indicadores, cada um dos quais possui múltiplos marcadores comportamentais.

negatividade pressionada na sala de aula ”[1]. O foco do nosso artigo é reconhecer essas duas dimensões automaticamente.

3.1 Diretrizes de codificação O

manual CLASS para cada faixa etária (crianças, jardim de infância, ensino fundamental, etc.) fornece diretrizes sobre como pontuar cada dimensão. As pontuações são normalmente atribuídas para cada dimensão uma vez a cada 15 minutos (e às vezes até 20 minutos [45]); esta escala de tempo permite tempo suficiente para que sejam feitos julgamentos significativos sobre a qualidade das interações em sala de aula. Cada julgamento é baseado na presença ou ausência de marcadores comportamentais que pertencem a um indicador específico de uma determinada dimensão de CLASSE; nesse sentido, CLASS é organizado hierarquicamente.

Os marcadores comportamentais podem abranger dimensões auditivas, visuais, linguísticas e pedagógicas. Por exemplo, ao avaliar o Clima Positivo, os codificadores do CLASS são instruídos a considerar a frequência com que os sorrisos são exibidos pelos participantes da sala de aula; se o professor chama os filhos pelo nome e os olha nos olhos; se as emoções entre professores e alunos são congruentes; etc.

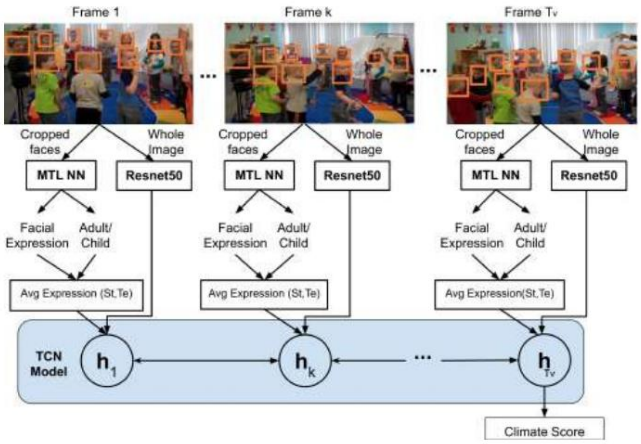
O Clima Negativo pode ser significado quando um professor levanta a voz com raiva de um aluno; faz ameaças de puni-los caso não se comportem; Embora esses comportamentos específicos possam servir como pontos de ancoragem para a codificação, a pontuação CLASS para cada dimensão é um julgamento holístico baseado em todo o segmento de vídeo de 15 minutos.
A Tabela 1 mostra um pequeno subconjunto de marcadores comportamentais aos quais

Os codificadores de CLASSE devem comparecer para Clima Positivo e Clima Negativo. É importante ressaltar que o Clima Negativo não é apenas a ausência de um Clima Positivo. Pelo contrário, o primeiro é caracterizado pela presença de comportamentos negativos evidentes, tais como ameaças e controlo punitivo. Uma sala de aula com um Clima Positivo baixo pode, portanto, também ter um Clima Negativo baixo.

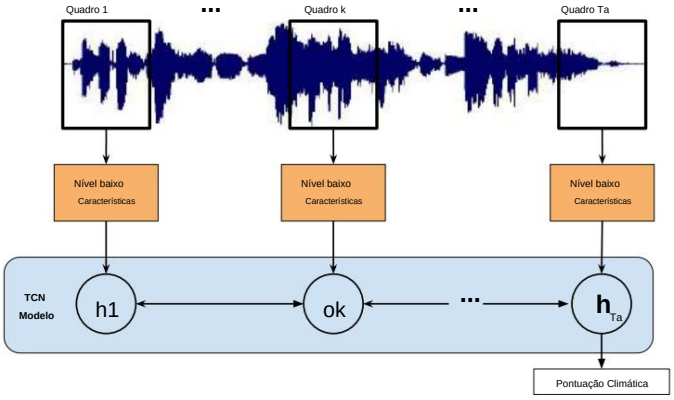
Para se tornarem proficientes na codificação CLASS, os observadores humanos normalmente se inscrevem em um seminário de treinamento de vários dias e depois continuam a praticar e a receber feedback ao longo de várias semanas ou meses. A proficiência é certificada por um exame online. Uma vez treinados, os codificadores CLASS podem assistir a sessões de sala de aula ao vivo ou gravadas em vídeo e fornecer um serviço valioso para professores, administradores e pesquisadores. Contudo, a quantidade de tempo envolvida na codificação CLASS é significativa e o trabalho é caro.

4 ABORDAGEM DE APRENDIZAGEM DE MÁQUINA MULTIMODAL

Nossa filosofia de projeto ao projetar a Rede Automática de Reconhecimento de Observação em Sala de Aula (ACORN) foi combinar recursos selecionados manualmente, conforme sugerido pelo Manual CLASS (por exemplo, estados afetivos dos participantes da sala de aula, estimados a partir de seus



(a) A via visual para prever as pontuações CLASS PC e NC.



(b) A via auditiva para prever os escores CLASS PC e NC.

1: Rede de Reconhecimento Automático de Observação em Sala de Aula (ACORN), compreendendo uma via visual e uma auditiva cujos resultados são calculados em média para estimar as pontuações CLASS.

expressões faciais) com características auditivas e visuais de baixo nível, como pixels brutos e coeficientes MFCC que são analisados por redes neurais convolucionais. Tratamos a estimativa da pontuação CLASS como uma multiclassificação em vez de um problema de regressão (ver discussão nos Materiais Suplementares) para que o sistema produza um elemento do conjunto {1, 2,. . . , 7}, conforme prescrito pelo Manual CLASSE.

Este artigo explora e estima o poder preditivo das representações de recursos visuais e auditivos para prever CLASSE PC e NC. Também avaliamos a precisão de diferentes abordagens para integrar informações ao longo do tempo.

4.1 Recursos visuais

Há uma variedade de marcadores comportamentais visuais que sugerem Clima Positivo. Por exemplo, o afeto positivo é sinalizado, até certo ponto, por expressões faciais como o sorriso, e os relacionamentos positivos estão associados a expressões faciais congruentes entre o professor e os seus alunos, ou seja, o professor demonstra emoção positiva quando os alunos demonstram emoção positiva. . Da mesma forma, demonstrações evidentes de raiva, frustração ou sarcasmo indicam um clima negativo.
Finalmente, também consideramos que importantes eventos e interações em sala de aula podem ser identificados por uma rede neural convolucional que analisa a imagem inteira de cada quadro de vídeo. Para evitar overfitting, principalmente na análise dos pixels dos frames do vídeo, utilizamos CNNs pré-treinados no ImageNet.

4.2 Características auditivas A

fala em sala de aula é claramente um fator crucial para todas as dimensões da CLASSE, incluindo PC e NC. De forma análoga à estimativa de emoções pela expressão facial de vídeo, treinamos detectores automáticos de emoções a partir de áudio e os usamos para estimar PC e NC. Além disso, de forma análoga à análise de todos os pixels de cada quadro de vídeo, também extraímos recursos de áudio de baixo nível (por exemplo, representação MFCC) do áudio da sala de aula que podem capturar recursos paralinguísticos e prosódicos, como sarcasmo, risadas, gritos, berros, chorando, etc

4.3 Integração Temporal Dada uma

série temporal de recursos (por exemplo, recursos baseados em CNN dos pixels de cada quadro de vídeo, expressão facial de professores e alunos em cada momento, enunciados de frases-chave, etc.), devemos analisar esta série temporal para chegar a uma estimativa final para as notas da CLASSE. Exploramos diversas abordagens: Mais simplesmente, podemos calcular a média de toda a série temporal (15 minutos em nossos conjuntos de dados). Podemos usar redes neurais recorrentes, como LSTMs e LSTMs bidirecionais. Mais recentemente, [46] mostraram que uma rede de convolução temporal (TCN) introduzida em [47] poderia superar os LSTMs em termos de velocidade, ao mesmo tempo que demonstrava uma memória efetiva mais longa.

4.4 Visão geral dos experimentos

Nas seções abaixo, descrevemos nossos experimentos para investigar os métodos mais eficazes de integração temporal (por exemplo, BiLSTM, TCN), representação de recursos (por exemplo, múltiplas expressões faciais, pixels de todo o quadro de vídeo) e arquiteturas de redes neurais (por exemplo, atenção , convolução gráfica). Para estimativa de pontuação CLASS PC e NC, as Seções 6 (áudio), 7 (vídeo) e 8 (conjunto) analisam o conjunto de dados UVA Toddler, enquanto a Seção 10 examina o conjunto de dados MET muito maior e explora como a precisão do modelo varia conforme o tamanho do conjunto de treinamento aumenta . Para encontrar os momentos-chave em um vídeo de sala de aula de 15 minutos com PC alto versus baixo, os experimentos na Seção 11 são conduzidos no conjunto de dados UVA Toddler.

Observe que não tentamos todas as combinações possíveis de todos os recursos, projetos de redes neurais e métodos de integração temporal, pois isso resultaria em um número muito grande de experimentos. Em vez disso, seguimos uma abordagem de desenvolvimento iterativo em que a arquitetura mais promissora que identificamos até agora foi ligeiramente modificada (por exemplo, inclusão de um modelo de atenção neural) para ver se o novo componente fazia diferença. O design do nosso sistema ACORN final é mostrado na Figura 1, que representa as vias visuais e auditivas cujos votos para as pontuações CLASS são calculados em média.

5 CONJUNTOS DE DADOS

Treinamos e testamos nossos modelos em dois conjuntos de dados codificados por CLASS: o conjunto de dados para crianças da Universidade da Virgínia (UVA), que contém salas de aula pré-escolares de crianças pequenas (2 a 3 anos de idade), e as Medidas de Ensino Eficaz (MET) [48] hospedado no Universidade de Michigan, que contém salas de aula do ensino médio

(a) Refeição em grupo (b) Atividades Individuais

(c) Brincadeiras ao ar livre (d) Atividades em grupo

Figura 2: Exemplos de configurações de sala de aula presentes no conjunto de dados UVA. Imagens mostradas com permissão.

(normalmente de 10 a 14 anos). Ambos os conjuntos de dados foram recolhidos em escolas reais em ambientes naturais; eles não são de estudos de laboratório.

5.1 UVA Criança

O conjunto de dados para crianças da Universidade da Virgínia (UVA) [49] consiste em 192 vídeos codificados em CLASS (ver Figura 2), com duração de 45 a 60 minutos, de 61 creches, onde os alunos são crianças de 2 a 3 anos de idade. (Observe que este conjunto de dados é uma versão expandida daquele que analisamos em nosso trabalho anterior [15].) UVA Toddler foi coletado como parte de um estudo financiado pelo Instituto de Ciências da Educação (IES) para explorar novos modelos de desenvolvimento profissional para professores. Os vídeos foram gravados por um observador treinado (professor ou cinegrafista) por meio de uma câmera digital montada em tripé com microfone integrado, com o objetivo de acompanhar visualmente os aspectos mais interessantes da dinâmica da sala de aula em cada momento. Cada vídeo mostra imagens de sala de aula de um dia típico de instrução pré-escolar, incluindo atividades individuais, atividades em grupo, brincadeiras ao ar livre e refeições compartilhadas (ver Figura 2). As salas de aula da pré-escola geralmente incluem canto, atividades de leitura conduzidas pelo professor, brincadeiras com blocos e outros brinquedos e café da manhã. Na maioria das salas de aula, pelo menos dois zeladores (professores e auxiliares) estão presentes: na média de todos os vídeos, há 1,70 professores (dp 0,787) e 7,59 alunos (dp 2,22) por sessão de sala de aula. Conforme mostrado na figura, os vídeos de observação em sala de aula são altamente desafiadores para os sistemas de visão computacional devido à iluminação descontrolada e à postura altamente não frontal da cabeça e do corpo dos participantes; fala sobreposta e fundos ruidosos também contribuem para a dificuldade da análise auditiva.

5.1.1 Demografia

Todos os professores eram mulheres. Raça dos professores: negra/afro-americana (48,2%), branca/caucasiana (39,3%), asiática (3,6%), multirracial (3,6%) e outras (3,6%). Etnia: 1,8% dos professores relataram ser hispânicos. Todos os vídeos foram gravados em salas de aula em um estado do Meio Atlântico dos EUA.

5.1.2 Codificação de CLASSE

De acordo com as diretrizes de codificação descritas pelo oficial CLASS Manual [1], cada vídeo é dividido em segmentos de 15 minutos, e cada segmento é rotulado para as 10 dimensões do protocolo CLASS-Toddler. No total, isso equivale a 300 segmentos de vídeo de 15 minutos distribuídos pelas 7 turmas, conforme mostrado na Tabela 2. Este tamanho é comparável a estudos recentes de computação afetiva e análise de sala de aula [20], [50]. A codificação CLASS foi realizada por 9 codificadores, que passaram por 2 dias de treinamento para o protocolo CLASS-Toddler e completaram uma avaliação de confiabilidade antes da codificação.

Uma amostra aleatória de cerca de 10% dos segmentos de vídeo foi rotulada por vários codificadores CLASS para avaliar a confiabilidade entre codificadores; veja a matriz de confusão na Tabela 5 (parte inferior). Além disso, entre as dimensões PC e NC para esses vídeos, a correlação de Pearson foi de ÿ0,446, ou seja, vídeos com maior PC tenderam a ter menor NC.

	Pontuação da CLASSE						
Dimensão	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Clima Positivo	0	7	28	74	78	92	21
Clima Negativo	243	43	11	3	0	0	0

TABELA 2: # segmentos de vídeo rotulados para cada pontuação CLASS no conjunto de dados UVA Toddler.

Para treinar nossos modelos, tratamos cada rótulo de cada codificador como um exemplo distinto; esta abordagem foi demonstrada em alguns estudos anteriores para aumentar a precisão em comparação com o treinamento no rótulo médio para cada exemplo [51], [52]. Para avaliação, usamos a pontuação média do CLASS, de todos os rotuladores, como a verdade básica.

5.2 MET Ensino Fundamental e Médio

O conjunto de dados Measures of Effective Teaching (MET) [8] é um dos maiores conjuntos de dados de vídeo codificados CLASS já coletados. Ele contém mais de 16.000 vídeos de 3.000 professores ensinando aulas de matemática, ciências ou artes da linguagem em escolas de ensino fundamental e médio em 6 distritos dos EUA. Em cada sala de aula, uma câmera esférica de 360 graus com microfone integrado foi colocada no centro da sala e usada para gravar o professor e os alunos simultaneamente. O MET foi coletado pela Fundação Bill & Melinda Gates e é hospedado pela Universidade de Michigan. Os vídeos são acessíveis apenas dentro do Virtual Data Enclave (VDE), que é um conjunto de máquinas virtuais que fornecem acesso restrito aos dados. Nenhuma transferência de dados é possível dentro ou fora do VDE sem autorização explícita da Universidade de Michigan.

Todas as análises devem ser realizadas em software aprovado.

5.2.1 Demografia

Calculada a média de todos os 6 distritos escolares do estudo (ponderada pelo número de participantes de cada distrito), a demografia [53] dos professores foi a seguinte: 77,8% mulheres, 24,7% afro-americanos, 9,1% latinos/latinos, 62,8% % de brancos não hispânicos e 3,4% de outra raça ou etnia. Demografia dos alunos: 48,7% mulheres, 30,4% afro-americanos, 33,9% latinos/latinos; para os alunos restantes, faltavam dados sobre raça e etnia.

5.2.2 Codificação de CLASSE

O histograma das pontuações PC e NC é mostrado na Tabela 3. Assim como o UVA Toddler, o conjunto de dados MET foi pontuado por vários (71) codificadores exclusivos para todas as dimensões CLASS. A matriz de confusão entre codificadores é mostrada na Tabela 6 (parte inferior). Entre as dimensões PC e NC, a correlação de Pearson foi de ÿ0,335.

	Pontuação da CLASSE						
Dimensão	**1**	**2**	**3 4 5**			**6**	**7**
Clima Positivo 271	23		883 1458 1632	1037	270	80 31 21	
Clima Negativo 3727	1385 323						7

TABELA 3: # segmentos de vídeo rotulados para cada pontuação CLASS no conjunto de dados MET.

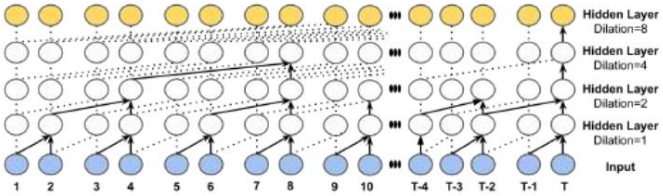


Figura 3: Rede Convolucional Temporal (TCN) [47] com avanços de dilatação de 1,2,4,8.

6 EXPERIMENTOS: VIA AUDITIVA

Primeiro consideramos arquiteturas de predição que usam apenas recursos auditivos. Esta abordagem tem uma possível vantagem em termos de privacidade: alguns alunos e professores podem sentir-se mais confortáveis com a gravação das suas vozes do que com vídeos com os seus rostos.
As características auditivas podem ser preditivas da CLASSE PC e NC de várias maneiras: No nível gestalt, elas podem dar uma noção de quanta excitação ou atividade está ocorrendo na sala de aula. Em um nível mais refinado, o áudio registra quem disse o quê, para quem, quando e com que emoção. Aqui usamos recursos espectrais como MFCC e Chroma.

Os experimentos nesta seção, todos conduzidos no conjunto de dados UVA Toddler, investigam qual mecanismo de integração temporal (média simples, 1-D CNN, BiLSTM, TCN) é mais eficaz para agregar informações de áudio para estimativa de CLASSE.

6.1 Procedimentos

Nossos modelos foram treinados com Adam usando uma taxa de aprendizado inicial de 0,001 com recozimento, por 500 épocas, com paciência de parada antecipada de 25 épocas. Treinamos e avaliamos nossos modelos no conjunto de dados UVA Toddler usando validação cruzada de 10 vezes em sala de aula, sujeito às seguintes restrições de estratificação: (1)
Sempre que possível, todos os níveis climáticos (1-7) foram representados em cada dobra; e (2) duas dobras não continham um videoclipe da mesma sala de aula. Subdividimos ainda cada dobra de treinamento (ou seja, validação cruzada dupla) em dois subconjuntos: um para otimização de parâmetros (treinamento) e outro para otimização de hiperparâmetros (validação). Pouco antes de enviar o artigo, reamostramos todas as dobras de validação cruzada e reexecutamos todos os experimentos para garantir que nossas descobertas fossem robustas em relação à escolha específica de dobras.
Quase todas as tendências que encontramos em relação a qual modelo funcionou melhor que outros permaneceram as mesmas; relatamos apenas as tendências que permaneceram consistentes após a remodelação das dobras.

Treinamos nossas redes neurais para estimar uma distribuição de probabilidade em 7 saídas discretas {1, 2,. . . , 7} usando perda de entropia cruzada. Em seguida, tratamos o rótulo de classe previsto como um número real e calculamos a correlação de Pearson com pontuações CLASS ordinais codificadas por humanos, de forma semelhante a como a estimativa da intensidade da expressão facial foi avaliada em estudos anteriores (54). Para testes de significância estatística, calculamos testes t bicaudais de que a média das correlações em todas as 10 dobras é diferente de 0. Relatamos os resultados da correlação no corpo do texto abaixo, e a maioria deles também é mostrada na Tabela 4. Ao lado cada título “Resultados” abaixo, relatamos o número do modelo correspondente na tabela.

6.2 Recursos auditivos de baixo nível Extraímos

todos os 34 recursos disponíveis do kit de ferramentas PyAudioAnal-ysis [55]. Cada arquivo de áudio é primeiro particionado em janelas não sobrepostas de 50 ms de comprimento (e, portanto, a taxa de amostragem das janelas é de 20 Hz); cada um deles é então particionado em duas subjanelas de 25 ms de comprimento. Os recursos são calculados dentro

cada subjanela e, em seguida, calculada a média das duas subjanelas; esta é a abordagem sugerida por [55]. Assim, para um áudio de 15 minutos (900 segundos), existem 18.000 vetores contendo 34 recursos, como coeficientes cepstrais de frequência Mel (MFCC), recursos cromáticos e vários outros recursos espectrais. Dadas essas características, comparamos vários métodos de integração temporal, descritos a seguir.

6.2.1 Média Simples No início

do desenvolvimento do ACORN, queríamos explorar se havia alguma correlação entre os recursos médios de áudio em cada vídeo e as pontuações CLASS PC e NC correspondentes. Assim, treinamos duas árvores de decisão (uma para PC e uma para NC) usando o algoritmo CART [56] que pegou o vetor médio de recursos de áudio de 34 dimensões como entrada e previu a pontuação CLASS. Observe que as árvores de decisão podem capturar relacionamentos não lineares.

Resultados (modelo #1): PC e NC foram previstos com correlações de Pearson de 0,27 e 0,26 em relação aos escores de verdade, respectivamente; ambos foram estatisticamente significativos. Estes fornecem evidências de que recursos de áudio de baixo nível, mesmo sem reconhecimento de fala downstream ou PNL, podem ser úteis para previsão de CLASSE.

6.2.2 Abordagem de convolução 1D Como

uma representação de recursos mais poderosa do que apenas o vetor médio de recursos de áudio, treinamos uma CNN 1-D que analisa os 18.000 vetores de áudio em todo o vídeo, aplicando kernels temporais de comprimento fixo. Em particular, a rede consistia em uma camada convolucional 1-D (128 canais de saída, largura do kernel de 10 passos de tempo), seguida por uma função de ativação ReLU, agrupamento de média global ao longo do eixo do tempo e, finalmente, uma camada densa combinada com softmax para estimar a pontuação CLASSE.

Resultados (modelo #2): PC e NC foram previstos com correlações de 0,28 e 0,26, respectivamente; ambos foram estatisticamente significativos.

6.2.3 BiLSTM

Para capturar não apenas o comportamento local médio, mas também a dinâmica das séries temporais de áudio, aplicamos redes neurais recorrentes. Em particular, treinamos um BiLSTM com 1 camada oculta contendo 100 neurônios que recebe cada recurso de áudio de 34 dimensões como entrada e produz uma estimativa de pontuação CLASS no intervalo de tempo final.

Resultados (modelo #3): PC e NC foram previstos com correlações de 0,23 e 0,22, respectivamente; ambos foram estatisticamente significativos. Na verdade, essas correlações são inferiores à Média Simples e à 1D-CNN. Isso pode ser devido ao desaparecimento dos gradientes do grande número de passos de tempo.

6.2.4 Rede Convolucional Temporal CNNs 1-D simples

podem ser vistas como um caso especial da Rede de Convolução Temporal (TCN) mais poderosa. A Figura 3 mostra um TCN com bloco residual único com passos de dilatação de

{1, 2, 4, 8}. Através do empilhamento de camadas de convolução dilatadas, o TCN pode ter um campo receptivo muito grande com relativamente poucas camadas e, assim, manter a eficiência computacional. Os TCN são uma alternativa às redes LSTM e GRU e podem manter alta precisão enquanto reduzem os custos de tempo de execução. Assim, tentamos usar um TCN para prever pontuações CLASS a partir do recurso de áudio. O TCN pegou um vetor de recursos de áudio de 34 dimensões em cada intervalo de tempo e produziu uma única estimativa de pontuação CLASS como saída no intervalo de tempo final.

Resultados (modelo #4): PC e NC foram previstos com correlações de 0,29 e 0,33, respectivamente; ambos foram estatisticamente significativos. Esta é uma melhoria modesta na precisão em relação aos modelos Mean+DT, 1D-CNN, BiLSTM. Uma explicação é que a dinâmica auditiva, ao invés de apenas o comportamento local ou o comportamento médio global, é preditiva de PC e NC. Uma explicação alternativa é que aumentar a profundidade computacional além de apenas uma única camada convolucional pode transformar os sinais locais em mais preditivos.

6.3 Experimentos adicionais Os materiais suplementares contêm resultados experimentais adicionais sobre o uso de classificação de frases-chave de uma rede neural treinada de forma personalizada [57], redes de detecção de eventos de áudio pré-treinadas, classificação de emoções auditivas e transcrição de fala para texto usando DeepSpeech [58], [59]. Esses resultados não são estatisticamente significativos. Nosso trabalho anterior [15] também inclui experimentos para explorar como lidar com o desequilíbrio de dados, que omitimos aqui por questões de brevidade.

7 EXPERIMENTOS: CAMINHO VISUAL

Aqui exploramos arquiteturas de predição para CLASS PC e NC que usam recursos puramente visuais de expressão facial e o número de rostos detectados em cada quadro. Variamos aspectos da arquitetura, como o backbone da rede neural convolucional (CNN) para reconhecimento (VGG-16 vs. Resnet-50), se alunos e professores são aglomerados ou tratados separadamente, e o método de integração temporal. Os experimentos são conduzidos no conjunto de dados UVA Toddler.

7.1 Expressão Facial Com base em nosso trabalho anterior [15], exploramos se as expressões faciais de alunos e professores, estimadas por classificadores faciais automáticos e integradas ao longo do tempo, poderiam prever CLASSE PC e NC. Para tanto, treinamos classificadores binários de sorriso/não-sorriso, raiva/não-raiva e tristeza/não-tristeza, bem como um detector criança versus adulto para distinguir entre alunos e professores na sala de aula. . Conforme relatado em [15], treinamos os detectores de sorriso e de criança/adulto no conjunto de dados de sala de aula do YouTube que coletados, e a raiva e a tristeza no conjunto de dados AffectNet [60].
Como primeira etapa de processamento, cada vídeo foi dividido em quadros a uma taxa de quadros fv de 3 Hz. Cada quadro foi então analisado pelo detector facial Faster R-CNN [61], que é robusto para faces não frontais.

Em termos de precisão da classificação binária de sorriso/não sorriso e criança/adulto, descobrimos que o Resnet-50 como backbone da CNN deu um pequeno, mas valioso, aumento na precisão em comparação com o VGG-16: No conjunto de dados do YouTube que coletamos contendo 70 vídeos das salas de aula pré-escolares [15], os classificadores baseados em Resnet alcançaram uma área sob a curva de características operacionais do receptor (AUC) de 0,967 (versus 0,942 para VGG) e 0,90 (versus 0,879 para VGG) para criança/adulto e sorriso/não -sorriso, respectivamente.
O modelo baseado em Resnet alcançou pontuações AUC de 0,872 e 0,884 para as tarefas de tristeza/não tristeza e raiva/não raiva.
Assim, investigamos se esse aumento de precisão para a percepção facial de baixo nível se traduziu em um aumento semelhante na precisão de previsão de pontuação CLASS posterior. Nota: esta seção examina apenas o sorriso, não as demais expressões faciais; na Seção 8 usamos todas as três expressões para predição de CLASSE. Comparamos várias abordagens de integração temporal, descritas abaixo.

7.1.1 Média Simples Como os sorrisos e as risadas são alguns dos indicadores comportamentais do Clima Positivo, parecia plausível que o sorriso médio, em todos os participantes e em todos os frames do vídeo, pudesse ser preditivo. Para explorar isso, treinamos uma árvore de decisão (usando o algoritmo CART [56]) que tomou como entrada a média de todos os quadros, da estimativa média do sorriso de cada rosto detectado em cada quadro, e previu a pontuação CLASS.

Resultados (modelo #5): PC e NC foram previstos com correlações de 0,11 e 0,08, respectivamente; nenhum dos dois foi estatisticamente significativo.

7.1.2 LSTMs

Para explorar se os valores de sorriso dinâmicos, em vez de apenas a média, podem ser preditivos, calculamos as pontuações médias de sorriso dentro de cada quadro de vídeo e, em seguida, passamos essas pontuações para um LSTM com 1 camada oculta contendo 100 unidades ocultas. O número de etapas recorrentes foi de 2.700 (900 segundos para um segmento de vídeo de 15 minutos a 3 quadros/segundo). Ao final da série temporal, foi previsto um único resultado que é a pontuação CLASS.

Ao calcular o valor do sorriso dentro de cada quadro, comparamos quatro estratégias: (1) O sorriso médio de todos os participantes (professores e alunos misturados); (2) o sorriso médio apenas dos alunos (ou seja, usamos as pontuações de sorriso apenas daqueles rostos que são considerados “crianças” pelo detector de crianças/adultos); (3) o sorriso médio apenas dos professores; (4) o sorriso médio de alunos e professores separadamente (ou seja, como duas características de entrada diferentes). Além disso, comparamos o VGG-16 ao Resnet-50 como backbone da CNN.

Resultados (modelos #6-#9): Nas quatro maneiras diferentes de calcular o valor médio do sorriso, conforme descrito acima, o método mais promissor foi calcular o sorriso médio do professor e o sorriso médio do aluno separadamente e, em seguida, integrar esses valores ao longo do tempo com um LSTM. Esse método obteve correlação de 0,13 para PC e 0,14 para NC; esses resultados foram estatisticamente significativos. Usar apenas o sorriso do professor ou o sorriso do aluno (mas não ambos) proporcionou menor precisão, assim como simplesmente mesclar todas as pessoas. Além disso, quase todos os resultados baseados em LSTM foram superiores aos do modelo 5, sugerindo que a dinâmica do sorriso era mais preditiva do que apenas o valor médio do sorriso em todos os quadros.

7.1.3 BiLSTMs Como geralmente não há restrições para estimar pontuações CLASS em tempo real, tentamos usar BiLSTMs, que podem aproveitar o conhecimento de eventos futuros para compreender o contexto de eventos atuais. Como antes, comparamos VGG com Resnet.

Resultados (modelos 10-11): A análise do vídeo em ambas as direções proporcionou um pequeno aumento de precisão em comparação com o modelo 9, gerando estatísticas melhoradas. ass. correlações de 0,19 e 0,21 com PC e NC. Além disso, descobrimos que o Resnet era um pouco mais preciso do que o VGG, fornecendo estatísticas. ass. correlações de 0,21 e 0,23; isso sugere que o aumento de precisão no reconhecimento de expressões faciais

(Seção 7.1) pode se traduzir em um aumento modesto na precisão da previsão CLASS downstream.

7.2 Número de rostos detectados

Na exploração piloto, levantamos a hipótese de que um recurso muito simples que consiste no número médio de rostos detectados em cada quadro de vídeo poderia prever as pontuações CLASS. A intuição é que os professores podem ser menos eficazes quando precisam atender muitas pessoas ao mesmo tempo. Assim, alimentamos uma série temporal, consistindo no número de faces detectadas em cada quadro, para um BiLSTM e previmos PC e NC com este único recurso.

Resultados (modelo #12): Houve correlação fraca de #faces com PC e NC: 0,07 e 0,09, respectivamente; nenhum dos dois foi estatisticamente significativo. Na verdade, esses números são maiores do que os do modelo nº 5 (com base no sorriso médio). Como devemos detectar rostos de qualquer maneira para calcular as características da expressão facial, decidimos manter o recurso #faces em nosso ACORN final.

8 EXPERIMENTOS: MODELOS DE CONJUNTO

Como próximo passo na construção do ACORN, combinamos as vias auditiva e visual. Em particular, cada caminho foi treinado de forma independente para produzir uma estimativa independente da pontuação CLASS, e o modelo conjunto calcula a média não ponderada das previsões desses modelos. (Na experimentação piloto, descobrimos que o aprendizado dos pesos nos dois caminhos não proporcionou nenhum benefício confiável.) Exploramos fatores como a inclusão de mais expressões faciais, todo o quadro da imagem e um modelo de atenção neural. As análises nesta seção são realizadas em UVA Toddler.

8.1 Recursos de sorriso e áudio espectral Avaliamos o quanto a precisão melhora se combinarmos (1) um modelo auditivo que prevê pontuações CLASS com um 1D-CNN a partir de recursos de áudio espectral (modelo # 2) e (2) um modelo visual que consiste em um BiLSTM em cima de um VGG que classifica separadamente os sorrisos de professores e alunos (modelo #10). Isto tem implicações práticas importantes: se o modelo auditivo for quase tão bom quanto o modelo de conjunto, então pode ser sensato, do ponto de vista da privacidade, eliminar completamente a via visual.

Resultados (modelo #13): A abordagem combinada produz correlações com PC e NC de 0,35 e 0,39; ambos foram estatisticamente significativos. Esses números representam uma melhoria substancial apenas nos modelos visual (0,19 e 0,21) e auditivo (0,28 e 0,26) , indicando que essas duas vias são altamente complementares. Em particular, a via visual contém informações valiosas não previstas pela nossa via auditiva.

8.2 Número de rostos detectados

Semelhante à Seção 7.2, tentamos adicionar o número de faces detectadas como entrada, para cada quadro de vídeo, ao BiLSTM.

Resultados (modelo #14): Incluir esse recurso aumentou o correlações muito ligeiramente (modelo wrt #13) para 0,35 e 0,40.

8.3 Mais expressões faciais Além do sorriso estimado (Sm) de cada aluno e professor, investigamos se o uso também de detectores de raiva (A) e tristeza (Sa) poderia aumentar a precisão da previsão. Estas duas emoções negativas podem ser particularmente úteis para o Clima Negativo.
Dentro de cada quadro de vídeo, calculamos a expressão média valor, usando o classificador de face binária apropriado, para professores e alunos separadamente.

Resultados (modelo #15): A inclusão de raiva e tristeza aumentou as correlações (modelo wrt #14) para 0,39 e 0,46.
Isto sugere que representações mais ricas de emoções faciais podem aumentar a precisão na análise de observação em sala de aula.

8.4 Análise de imagem completa Além de analisar o rosto de cada participante da sala de aula, outros recursos visuais que respondem a perguntas como “onde estão todos”, “o que eles estão fazendo” e “quais são as relações uns com os outros” também podem ser importantes para estimar o CLASS PC e NC. Portanto, investigamos se a inclusão dos pixels de todo o quadro da imagem poderia melhorar a precisão do reconhecimento. Em particular, usamos um VGG-16 (pré-treinado no ImageNet e depois ajustado no conjunto de dados UVA Toddler) para mapear cada imagem de entrada em um vetor de características com 7 × 7 × 512 = 25.088 dimensões. Este vetor foi então concatenado com os recursos de expressão facial e #faces e passado para o BiLSTM para estimativa da pontuação CLASS; portanto, os recursos de imagem inteira e expressão facial foram usados em conjunto durante o treinamento para estimar as pontuações CLASS.

Resultados (modelo #16): A inclusão dos pixels de cada quadro de vídeo aumentou substancialmente as correlações (modelo wrt #15) para 0,47 e 0,53. Isso sugere que há informações visuais substanciais além dos rostos que podem ser efetivamente aproveitadas pelas CNNs modernas para previsão de pontuação CLASS. É notável que uma CNN possa extrair dos pixels brutos uma construção semanticamente de alto nível como pontuações CLASS.

8.5 Modelos de Atenção

Nos últimos anos, os modelos de atenção neural melhoraram significativamente a precisão das redes neurais, não apenas para tarefas de análise sequencial, como tradução [62], mas também em tarefas de visão computacional [63], [64]. Mecanismos de autoatenção permitem que uma rede neural atenda às partes mais importantes de uma determinada entrada, de uma forma vagamente motivada pelo processamento visual humano [65]. Em nosso trabalho, implementamos uma variação da autoatenção apresentada em [66] que adicionamos aos modelos Resnet-50 e VGG-16 antes da camada final de nivelamento/pooling. Para calcular os pesos de atenção, realizamos o seguinte cálculo:

$$a = \ddot{y}(\mathbf{W}ah) \quad (1)$$

$$o = \text{softmax}(a)\, h \quad (2)$$

Dada a saída da camada de convolução h, primeiro calculamos a saída de autoatenção a usando pesos de atenção aprendidos **Wa** (Equação 1). Aplicamos um sigmóide para comprimir as saídas de **Wah** em (0, 1); isso ajuda a evitar que qualquer recurso domine demais outros recursos. Em seguida, aplicamos um softmax sobre as saídas de atenção a e então multiplicamos pelo próprio mapa de características original para obter a saída final atendida o.

Resultados (modelo #17): Incorporando o modelo de atenção em aumentou as correlações (modelo wrt #16) para 0,51 e 0,58.

8.6 Redes Convolucionais Temporais Semelhante à Seção 6.2.4, aqui investigamos se o uso de um TCN para as vias auditiva e visual melhoraria a precisão em comparação com um BiLSTM.

Resultados (modelo #18): Com o TCN, as correlações foram ligeiramente piores em comparação com a abordagem BiLSTM (modelo #17):

0,50 e 0,56. No entanto, o TCN é significativamente mais rápido em tempo de treinamento e teste do que o BiLSTM: o treinamento de um modelo BiLSTM leva cerca de 9 horas em uma GPU P100, enquanto um modelo TCN leva apenas 6 horas. No momento do teste, apenas para a integração temporal (sem contar a análise Resnet do quadro da imagem ou dos rostos), o BiLSTM leva cerca de 8 a 9 minutos por vídeo de 15 minutos, enquanto o TCN leva cerca de 3 minutos.

8.7 Resnet vs.

Devido à maior precisão relatada para o reconhecimento de expressões faciais na Seção 7.1, substituímos VGG por Resnet para ver se ele aumentava as correlações com as pontuações CLASS verdadeiras.

Resultados (modelo #21): Usando Resnet como backbone da CNN aumentou as correlações (modelo wrt #18) para 0,55 e 0,63.

8.8 Comparação com trabalhos anteriores [37]

O único trabalho anterior (além do nosso [15]) do qual temos conhecimento sobre a previsão automática de pontuação CLASS é o de James et al. [37].
Em vez de aderir à definição CLASSE de deteção de Clima Positivo e Clima Negativo como resultados independentes, eles tratam-nos como dois lados de um continuum e tentam distinguir entre clima positivo e clima negativo num vídeo de 15 minutos. Relatamos o desempenho de nosso modelo limitando nosso modelo para PC com uma pontuação de 4. Usando esse limite, a pontuação F1 de nosso modelo é 0,86 no conjunto de dados UVA Toddler, em comparação com 0,78 em [37] em seu próprio conjunto de dados. Observamos que esta comparação não é igual devido a uma formulação de problema e conjunto de dados de teste diferentes.

9BOLOTA: COMPARAÇÃO COM CODIFICADORES HUMANOS

Escolhemos o modelo #21 como nosso ACORN final. Quão precisa é esta rede em comparação com as pontuações CLASS verdadeiras no conjunto de dados UVA Toddler (definido como a pontuação média de todos os codificadores CLASS humanos que rotularam cada exemplo), não apenas em um nível agregado, mas dividido por pontuação CLASS (1- 7)? A máquina comete erros semelhantes aos dos codificadores humanos?

9.1 Agregado Usando

20% do conjunto de dados UVA Toddler que foi pontuado por vários codificadores CLASS, estimamos a confiabilidade entre codificadores tomando cada codificador c como o codificador verdadeiro, calculando a correlação de Pearson das pontuações dos outros codificadores em relação ao pontuações de c e, em seguida, calculando a média de todos os c. Isto resultou em uma correlação média de Pearson de 0,38 para PC e 0,44 para NC. (As correlações de Spearman correspondentes foram ligeiramente superiores em 0,44 e 0,49). A precisão dos códigos CLASSE humanos ACORN escritos neste conjunto de dados é, surpreendentemente, maior do que a confiabilidade entre codificadores.

9.2 Matrizes de Confusão

A Tabela 5 mostra a matriz de confusão das previsões (linhas) da ACORN em relação às pontuações PC e NC (colunas) verdadeiras, conforme anotado por codificadores CLASS especializados. As tabelas foram calculadas concatenando as previsões de 7 vias da máquina em todas as 10 dobras de validação cruzada (300 previsões no total) e depois normalizando dentro de cada pontuação de verdade. Eles representam as distribuições de probabilidade condicionais P(ˆy | y), onde yˆ é a estimativa da máquina e y é a verdade básica. Para comparação, também calculamos as matrizes de confusão entre codificadores de codificadores CLASS humanos no subconjunto de 20% que foi codificado multiplicadamente. Nós tratamos cada codificador c como a verdade básica e cada outro codificador c como um estimador; em seguida, calculamos a média de todos c e normalizamos dentro de cada coluna.

Resultados: Comparando as duas tabelas para PC e as duas tabelas para NC, vemos evidências de que a máquina às vezes comete grandes erros – ou seja, uma grande diferença absoluta entre y e yˆ – que os codificadores humanos não cometem. Por exemplo, para PC, a máquina às vezes confundia uma pontuação de 2 com 6. Por outro lado, também houve casos de grande discrepância entre codificadores humanos, por exemplo, a variância sobre as distribuições P(ˆy | y = 3) para tanto o PC quanto o NC eram grandes para codificadores humanos. Não há nenhum padrão óbvio de rotulagem incorreta na máquina que os codificadores humanos fizeram.
não.

9.3 Experimentos Adicionais Os

Materiais Suplementares contêm uma análise adicional comparando abordagens baseadas em classificação e regressão para estimativa de pontuação CLASS e como isso influencia a correlação empírica entre pontuações estimadas de PC e NC.

10 RESULTADOS NO MET DATASET

Todos os resultados até agora foram obtidos no conjunto de dados UVA Toddler.

A abordagem de alto nível generaliza-se para outras populações de estudantes mais velhos, onde os tipos de interações, pedagogias e estilos de sala de aula são muito diferentes daqueles das salas de aula pré-escolares? Para explorar esta questão, treinamos e testamos modelos de previsão CLASS no conjunto de dados Measures of Effective Teaching (MET).

Dado que o MET contém milhares de vídeos, investigamos duas questões principais: (1) Como a precisão da previsão (medida pela correlação de Pearson) aumenta com a quantidade de dados de treinamento e a complexidade do modelo? (2) Como a precisão do modelo treinado e testado em alunos do ensino fundamental e médio se compara a um modelo análogo treinado e testado em crianças pequenas?

10.1 Procedimentos

Do total de mais de 16.000 segmentos de vídeo no MET, 5.574 deles são codificados para a CLASSE. Dividimos esses segmentos de vídeo em 3.874 segmentos de treinamento e 2.000 segmentos de teste. Devido às restrições de RAM no VDE que nos impediram de treinar um único modelo em todos os 3.874 segmentos, dividimos novamente os 3.874 segmentos de treinamento em 10 dobras diferentes. Devido às restrições de software no VDE, não foi possível instalar as bibliotecas necessárias para realizar visão computacional neste conjunto de dados. Em vez disso, implementamos apenas uma via auditiva: usando o pacote de análise de áudio tuneR [67], extraímos os 200 principais recursos do MFCC com as maiores energias em cada janela de 1 segundo a uma frequência de 1 Hz de cada vídeo. Usando esses recursos, treinamos florestas aleatórias de n árvores de decisão (n é um hiperparâmetro) para prever pontuações CLASS.

Para cada k = 2,. . . , 10, treinamos florestas aleatórias em k - 1 dobras, testamos na dobra restante como um conjunto de resistência e calculamos a média dos resultados nas k dobras. (Observe que k não é o número de conjuntos nos quais particionamos o conjunto de treinamento como na validação cruzada normal; em vez disso, k - 1 é o número de dobras usadas para treinamento.) Realizamos esse processo para cada número de dobras k e cada número de árvores de decisão n ÿ {10, 15, 20, . . . , 50}. Em seguida, escolhemos a melhor combinação (n, k) e treinamos um modelo final, que avaliamos nos 2.000 segmentos de vídeo no conjunto de teste que nunca foram vistos durante o treinamento ou otimização de hiperparâmetros.

#	Caminho visual da abordagem de estimativa							Positivo Clima		Negativo Clima	
	Temp da Via Auditiva .	de pontuação CLASS									
	Int..	Expressões	#Rostos?	Quadro?	Atenção CNN?	Temperatura	Internacional	R	p	R	p
	Média+DT	–	–	–	–	–	–	0,27	0,002	0,26	0,003
1	1D-CNN	–	–	–	–	–	–	0,28	<0,001	0,26	<0,001
2	BiLSTM	–	–	–	–	–	–	0,23	0,004	0,22	0,009
3 4	TCN	–	–	–	–	–	–	0,29	<0,001	0,33	<0,001
5	–	{Sm} × {Todos}	Não	Não	VGG Sem Média+DT 0,11 VGG Não LSTM			0,162		0,08	0,192
6	–	{Sm} × {Todos}	Não	Não	0,10 VGG Não LSTM 0,09 VGG Não LSTM			0,113		0,13	0,091
7	–	{Sm} × {St}	Não	Não	0,03 VGG Não LSTM 0,13 VGG Não BiLSTM			0,121		0,10	0,119
8	–	{Sm} × {Te}	Não	Não	0,19			0,511		0,06	0,291
9	–	{Sm} × {St,Te}	Não	Não				0,023		0,14	0,033
10	–	{Sm} × {St,Te}	Não	Não				0,009		0,21	0,006
11	–	{Sm} × {St,Te}	Não	Não	Resnet sem BiLSTM 0,21			0,008		0,23	0,007
12	–	–	Sim	Não	–	Sem BiLSTM 0,07		0,311		0,09	0,285
13	1D-CNN	{Sm} × {St,Te}	Não	Não	VGG Sem BiLSTM 0,35 <0,001 0,39 <0,001						
14	1D-CNN	{Sm} × {St,Te}	Sim	Não	VGG Sem BiLSTM 0,35 <0,001 0,40 <0,001						
15	1D-CNN	{Sm,A,Sa} × {St,Te}	Sim	Não	VGG Sem BiLSTM 0,39 <0,001 0,46 <0,001						
16	1D-CNN	{Sm,A,Sa} × {St,Te}	Sim	Sim VGG Não BiLSTM 0,47 <0,001 0,53 <0,001							
17	1D-CNN	{Sm,A,Sa} × {St,Te}	Sim	Sim VGG Sim BiLSTM 0,51 <0,001 0,58 <0,001							
18	TCN	{Sm,A,Sa} × {St,Te}	Sim	Sim VGG Sim 0,50 <0,001 0,56 <0,001TCN							
19	1D-CNN	{Sm,A,Sa} × {St,Te}	Sim	Não	Resnet Sem BiLSTM 0,40 <0,001 0,49 <0,001						
20	1D-CNN	{Sm,A,Sa} × {St,Te}	Sim	Sim	Resnet Sem BiLSTM 0,51 <0,001 0,56 <0,001						
21	TCN	{Sm,A,Sa} × {St,Te}	Sim	Sim	Resnet 0,55 <0,001 0,63 <0,001TCN						

TABELA 4: Precisão da previsão (correlação de Pearson r, valor p bicaudal) no conjunto de dados UVA Toddler de 21 modelos diferentes para estimar CLASSE Clima Positivo e Clima Negativo. St=aluno, Te=professor, Sa=Tristeza, Sm=Sorriso, A=Raiva, DT=Árvore de Decisão.

Matrizes de confusão: máquina-humano

Clima Positivo

	1	2	3	4	5	6	7
1	0	0	0	0	0	0	0
2	0	0	0	,03	,03	0	,05
3	0	,14	,41	,07	,08	,03	0
4	0	,43	,25	,53	,18	,13	0
5	0	0	,16	,14	,47	,14	,15
6	0	,43	,14	,15	,16	,55	,1
7	0	0	,04	,08	,08	,15	,7

Clima Negativo

	1	2	3	4	5	6	7
1	,86	,25	,11	,11	0	0	0
2	0	,5	,05	,05	0	0	0
3	,14	,2	,84	,2	0	0	0
4	0	0,05	0	0,64	0	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0

Matrizes de confusão: humano-humano

Clima Positivo

	1	2	3	4	5	6	7
1	0	0	0	0	0	0	0
2	0	,5	,06	,03	0	0	0
3	0	,17	,5	,03	,02	,04	0
4	0	,33	,13	,68	,18	,09	0
5	0	0	,06	,15	,6	,13	0
6	0	0	,25	,11	,2	,73	,5
7	0	0	0	0	0	,01	,5

Clima Negativo

	1	2	3	4	5	6	7
1	,91	,31	,25	0	0	0	0
2,08,67	0,25	0	0	0			
3,01	0,5	,25	0	0	0		
4	0	0,02	0,25	0,5	0	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0

TABELA 5: **Parte superior:** Matrizes de confusão normalizadas da máquina (modelo #21) versus codificadores CLASS humanos no UVA Toddler conjunto de dados. As linhas são as previsões (arredondadas); colunas são aterradas verdade. **Parte inferior:** Matrizes de confusão entre codificadores (humanos).

10.2 Resultados

Os resultados são mostrados na Figura 4. As correlações de Pearson de as pontuações CLASS previstas e verdadeiras aumentaram quase monotonicamente à medida que k aumentou de 2 para 10 (correspondendo a um tamanho do conjunto de treinamento de cerca de 380 segmentos de vídeo até 3874) para PC e NC. Em k = 10, a precisão para PC e NC é ainda aumentando, embora a curva esteja se achatando ligeiramente para PC. Isto sugere que significativamente mais precisão pode ser obtida simplesmente adicionando mais dados de treinamento, mesmo usando este arquitetura com características puramente auditivas. Em termos de número de complexidade do modelo, a precisão da floresta aumentou com n para quase todos k. Houve, no entanto, retornos decrescentes acima n = 35 árvores de decisão.

Com base nesses resultados, treinamos uma floresta aleatória final (n = 35 , pois era mais simples e dava precisão equivalente a n = 50) em todas as 10 dobras de dados de treinamento e avaliou-os no teste de 2000 vídeos; isso alcançou correlações de Pearson de 0,36 e 0,41 no PC e NC, respectivamente. É provável que os modelos MET tenham sofrido devido aos modelos relativamente superficiais que treinamos, mas que eles também se beneficiou de ter muito mais dados de treinamento em comparação com Criança UVA.

10.3 Matrizes de Confusão

Semelhante à Seção 9.2, calculamos a confiabilidade entre codificadores de codificadores CLASS humanos no MET. O intercodificador Pearson correlações no conjunto de dados MET, conforme avaliado no vídeo 1044 segmentos que foram duplamente codificados, foram 0,42 e 0,51 para PC e NC respectivamente. (As correlações de Spearman foram ligeiramente mais altas em 0,48 e 0,53.) Estes são superiores à precisão da máquina, mas não dramaticamente. Também calculamos tanto a relação máquina-humano e matrizes de confusão humano-humano para PC e NC no MET conjunto de dados; veja a Tabela 6. Para ambos os casos, isso acontece ocasionalmente que um codificador pode atribuir uma pontuação que difere em 3 níveis de pontuação atribuída por outro codificador. Não há nenhuma tendência óbvia de que o máquina comete erros flagrantes com muito mais frequência do que humanos os codificadores fazem.

11 IDENTIFICANDO OS MOMENTOS CHAVE DA SALA DE AULA

Indiscutivelmente, as oportunidades mais impactantes de observação em sala de aula habilitada por IA são fornecer feedback específico sobre determinados assuntos. momentos em uma sessão de sala de aula. Passando da análise agregada ao específico é um grande desafio de pesquisa: baseado em aprendizado de máquina sistemas para detectar e rastrear rostos, reconhecer estados emocionais e outras tarefas perceptivas geralmente são executadas com alta precisão em média, mas ainda assim podem cometer erros embaraçosos em tarefas específicas.

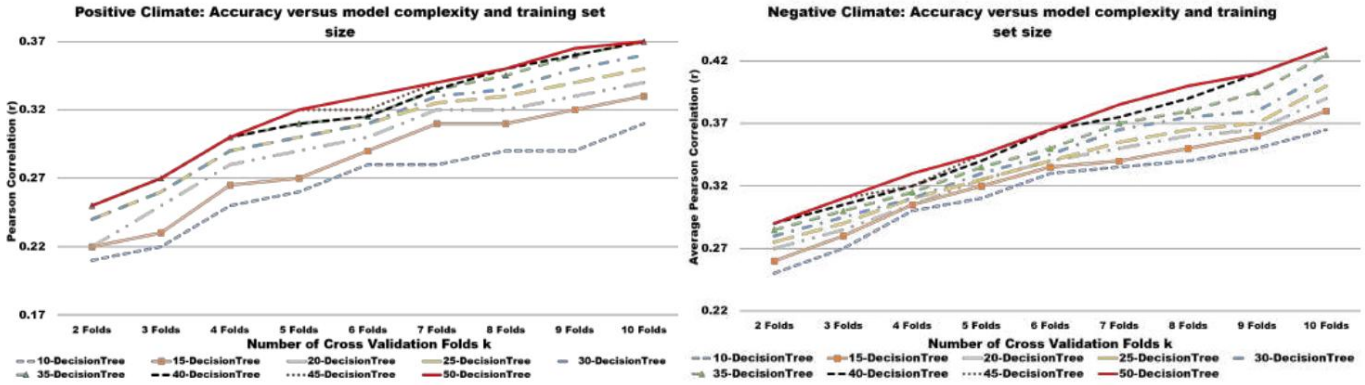


Figura 4: Correlação de Pearson entre pontuações CLASS previstas e codificadas por humanos no conjunto de dados MET. Cada modelo foi treinado como uma floresta aleatória de n árvores em (k ÿ 1) dobras de treinamento e testado na dobra restante. **Esquerda:** Clima Positivo. **À direita:** Clima negativo.

Clima Positivo

	1	**2**	**3**	**4**	**5**	**6**	**7**				
1	.4	.4	.15	0	0	0	0	**2**	.3	.4	.25
3	.3	.2	.5	.1	.1	0	0	**4**	0	0	.1
0	0	0	.1	.4	.25	0	**6**	0	0	0	.1
0	0	.1	.0	.0	.05	.6					

Clima Negativo 1 2 3

	4	**5**	**6**	**7**					
1	.8	.1	.1	0	0	0	0	**2**	.1
.05	.2	.4	.05	0	0	0	**4**	.05	.04
5	0	. 03	.1	.1	.8	1.	0	**6**	0
7	0	0	0	0	.1	0	1.		

Clima Positivo

	1	**2**	**3**	**4**	**5**	**6**	**7**					
1	.4	.1	.02	0	0	0	0	**2**	.5	.5	.1	0
.1	.4	.5	.2	.05	0	0	**4**	**0**	0	.28	.4	.1
5	0	0	.1	.2	.4	.2	.1	**6**	0	0	0	.2
0	0	0	0	.05	.1	.7						

Clima Negativo 1 2 3

	4	**5**	**6**	**7**					
1	.6	.2	.2	0	0	0	0	**2**	.3
3	.1	.2	.5	.2	0	0	0	**4**	0
5	0	0	0	. 2	.7	.1	0	**6**	0
7	0	0	0	0	0	0	1.		

TABELA 6: Parte **superior:** Matrizes de confusão normalizadas da máquina (modelo #21) versus codificadores CLASS humanos no conjunto de dados MET.
As linhas são as previsões (arredondadas); colunas são verdades básicas.
Parte inferior: Matrizes de confusão entre codificadores (humanos).

pessoas ou em momentos específicos. Nesta seção, exploramos algumas abordagens para encontrar automaticamente as interações mais importantes em sala de aula dentro de um segmento de vídeo de 15 minutos, semelhante a alguns trabalhos anteriores sobre resumo de vídeo [22], [68], [69]. Em particular, nos concentramos em distinguir momentos (45-90 segundos de duração) que apresentam PC muito baixo de momentos que exibem PC muito alto. (Como a rotulagem nesta curta escala de tempo é nova e está fora do alcance do Manual CLASS, decidimos não rotular PC “intermediário” para tornar a tarefa mais tratável para codificadores humanos.) Este é um problema de classificação binária, e medimos precisão como a Área Sob a Curva ROC (AUC), que é igual à probabilidade de a máquina, ao se deparar com um momento com PC alto e um momento com PC baixo, conseguir distinguir corretamente qual é qual. Se for bem-sucedida, essa ferramenta poderá ajudar os professores a identificar os momentos-chave que contêm CP muito alto ou muito baixo e a compreender quando e por que razão as suas interações com os alunos foram particularmente eficazes. Nosso problema de sumarização é supervisionado porque os momentos que queremos encontrar dependem de uma determinada dimensão CLASS (PC). Para resolver esse problema, coletamos mais rótulos para o conjunto de dados UVA Toddler e exploramos quatro abordagens algorítmicas diferentes.

11.1 Conjunto de dados

Recrutamos vários programadores da Curry School of Education da Universidade da Virgínia que foram treinados no CLASS para assistir aos vídeos UVA Toddler e encontrar vários clipes dentro de cada vídeo que exibem um clima positivo “alto” e vários clipes que exibem um clima positivo “baixo”. ; os clipes variaram de 45 a 90 segundos.

Também pedimos aos codificadores que fornecessem uma breve descrição explicando o raciocínio por trás do rótulo fornecido. Por exemplo, por um momento classificado como PC baixo, o codificador observou “nenhuma interação entre crianças e professores – os alunos também não estão interagindo uns com os outros”.
Por um momento classificado como PC alto, o codificador notou a presença de “tons entusiasmados e animados”. No total, foram obtidos 717 clipes rotulados. Dividimos esses clipes rotulados usando as mesmas dobras de validação cruzada de nossos experimentos anteriores no UVA Toddler.

11.2 Abordagem 1: Saída gradual do TCN A primeira abordagem

que tentamos foi treinar um classificador binário momentâneo de “PC alto/baixo” (uma saída por intervalo de tempo) em conjunto com o detector agregado que estima uma pontuação CLASS PC para todos os 15 minutos. segmento de vídeo (uma saída no final de toda a sequência), ou seja, adicionando uma tarefa secundária para prever momentos de PC baixo/alto. Nosso objetivo aqui foi determinar se o treinamento conjunto para prever a pontuação do PC para todo o vídeo e os momentos de PC baixo/alto levaram a um melhor desempenho do modelo, aprendendo recursos generalizados. Para tanto, expandimos o TCN no modelo #21 para produzir uma previsão em cada intervalo de tempo para indicar se aquele momento estava associado a PC “alto” (1) ou “baixo” (0). Em seguida, adicionamos termos binários de perda de entropia cruzada a todos os intervalos de tempo t que coincidiram com um clipe para o qual um rótulo codificado por humanos (PC alto ou baixo) foi fornecido. O modelo também incluiu uma perda de entropia cruzada de 7 vias para a pontuação agregada do PC. Como em todos os nossos experimentos no conjunto de dados UVA Toddler, treinamos e testamos os modelos em uma validação cruzada de 10 vezes.

Resultados: A AUC (tarefa de classificação binária) média (em todas as dobras) para determinar se cada momento exibia PC alto/baixo foi de apenas 0,39 – pior do que adivinhar. Além disso, a correlação de Pearson para prever a própria CLASSE PC (1-7) diminuiu consideravelmente para 0,47 (abaixo de 0,55). Isso sugere que a estimativa agregada da pontuação CLASS pode não se decompor trivialmente na média de muitas previsões momentâneas.

11.3 Abordagem 2: Classificação Binária Aqui treinamos um

classificador binário baseado em TCN que analisa videoclipes individuais de duração variável (45-90 segundos) e estima

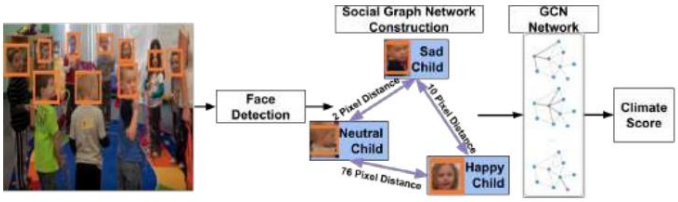


Figura 5: Analisando uma cena de sala de aula como uma rede social usando uma Rede de Convolução de Grafos.

através de uma unidade logística sigmóide se o clipe apresenta PC alto ou baixo. Ao contrário da Abordagem 1, este modelo também não tenta estimar o CP agregado. Comparamos duas arquiteturas: (1) modelo #21 mas sem via auditiva; (2) o modelo completo nº 21 com via auditiva.

Resultados: O TCN com apenas informação visual não teve um bom desempenho: a AUC média foi de 0,35 (pior que o acaso).
No entanto, utilizando a informação áudio, a AUC melhorou ligeiramente para 0,58 e foi estatisticamente significativa (teste de classificação de sinais de Wilcoxon, p = 0,009). Com base nas informações que coletamos dos codificadores humanos explicando por que eles classificaram cada clipe como PC alto/baixo, especulamos que os marcadores comportamentais auditivos do Clima Positivo – por exemplo, o grau de calor e positividade na voz de um professor – podem ser mais fáceis de identificar. detectar. Em contraste, alguns dos recursos visuais preditivos de PC alto/baixo envolviam interações mais complexas, como “revezar-se com as crianças jogando basquete e mostrar-lhes como arremessar” (como foi rotulado por um codificador para um vídeo).

11.4 Abordagem 3: GCNs para análise de redes sociais Motivados pelas recentes arquiteturas de aprendizagem profunda para gráficos [21], exploramos se há informações subjacentes que poderiam ser extraídas da sala de aula, visualizando-as como um gráfico social de interações entre alunos e professores (ver Figura 5). A convolução de grafos tornou-se popular nos últimos anos e impactou uma variedade de campos relacionados na computação afetiva, por exemplo, reconhecimento de emoções de grupo [70]. Em nosso trabalho, treinamos uma Rede Convolucional de Grafos (GCN) [21] construindo um gráfico a partir dos participantes da sala de aula detectados: Cada face detectada em cada quadro de vídeo é um nó, e a matriz de adjacência ponderada de nós em cada quadro é calculada como a distância inversa do pixel (no espaço 2-D) entre os centros das caixas faciais. Para diminuir o efeito de alarme falso detecções de rosto, limitamos o número de detecções em cada quadro para 22, que é o número máximo de participantes em qualquer sala de aula UVA Toddler. A cada nó associamos um vetor de características quadridimensional composto pelas três previsões probabilísticas de tristeza, raiva e sorriso, bem como a probabilidade criança/adulto de acordo com os classificadores automáticos de faces.

Para classificar um videoclipe curto como PC alto/baixo, construímos um gráfico social para cada vídeo; aplicou duas operações sequenciais de convolução de gráfico (100 filtros cada), cada uma seguida por uma camada de ativação e eliminação de ReLU; e então calculou uma soma de todos os recursos no gráfico ponderados por pontuações de atenção, semelhante à Seção 8.5. Embora métodos de agrupamento baseados em atenção tenham sido propostos antes [71], nosso mecanismo de atenção é aplicado no gráfico de saída após a convolução do gráfico. A ideia por trás da atenção é que podemos identificar os principais participantes presentes em cada quadro do vídeo usando os pesos de autoatenção. Também condensa o gráfico em uma representação de comprimento fixo. Descobrimos que tanto o mecanismo de atenção quanto o abandono foram essenciais para obter um bom desempenho com o GCN. Finalmente, agregamos o vetor de recursos para todos os quadros ao longo do tempo usando um BiLSTM (3 camadas de profundidade, 10 unidades ocultas). Nossos modelos foram treinados com Adam como otimizador, usando uma taxa de aprendizado inicial de 0,001 por 100 épocas, usando as mesmas dobras de validação cruzada dos outros experimentos UVA Toddler.

11.4.1 A topologia é importante?

Dentro de uma rede maior, uma camada GCN calcula uma agregação não linear de vetores de características de múltiplos nós, ponderados de acordo com a matriz Laplaciana do grafo induzida pela topologia do grafo. Para explorar se a topologia gráfica de quem está onde e quando era realmente importante, ou se o GCN simplesmente calcula a média das características individuais de todos os participantes em vez de examinar as interações entre eles, comparamos a abordagem GCN descrita acima com as duas alternativas a seguir: (1) Definimos a matriz Laplaciana normalizada que codifica a topologia do grafo como a matriz identidade I. Neste caso, cada nó do grafo é completamente isolado, ou seja, cada nó está conectado apenas a si mesmo. (2)

Definimos a matriz Laplaciana normalizada como uma matriz uniforme com todas as entradas iguais ao valor 1/d, onde d é o número de nós no gráfico. Neste caso, o grafo é uma clique (com autoconexões).

Resultados: Usando a topologia gráfica induzida pelas detecções faciais reais, a AUC média nas 10 dobras, para discriminar PC alto e baixo, foi de 0,70 (p = 0,005, teste de classificação de sinais de Wilcoxon). Embora ainda permita melhorias, este é o melhor resultado de todas as abordagens que tentamos para detectar PC alto versus baixo em videoclipes curtos. Em comparação, as AUCs obtidas para as matrizes de identidade ou de adjacência uniforme foram aleatórias (0,48 e 0,52, respectivamente). Isso sugere que a topologia de quem está onde e interagindo com quem e quando é importante para estimar o CP em sala de aula. Além disso, é digno de nota que o modelo GCN que analisa apenas os dados de emoção e idade tem um desempenho melhor do que a abordagem da Seção 11.3 baseada no modelo nº 21, que analisa a imagem inteira, os recursos de áudio, juntamente com os dados emocionais agregados para quadros individuais ao longo de toda a sequência.

11.5 Abordagem 4: Rede Siamesa Em contraste com as Abordagens 1, 2 e 3 acima, que tentam classificar um videoclipe como PC alto versus baixo em uma escala absoluta, aqui exploramos uma abordagem de destaque e resumo de vídeo de última geração [22] que usa uma rede siamesa para pegar dois videoclipes do mesmo vídeo e exibir qual deles apresenta PC mais alto. As duas entradas para a rede (uma de cada clipe) são produzidas pelo TCN no modelo #21 que foi modificado para produzir um escalar. Então, esses dois escalares são processados de forma não linear por uma rede neural densa de 2 camadas (2 neurônios ocultos cada) e uma unidade logística sigmóide para indicar se o primeiro clipe (0) ou o segundo clipe (1) possui PC mais alto. Como não tínhamos certeza se esta arquitetura siamesa funcionaria para a previsão de momentos-chave, implementamos um “controle positivo”, ou seja, treinamos outro modelo com exatamente a mesma arquitetura, mas com objetivo de treinamento diferente, ou seja, para distinguir entre dois vídeos inteiros de 15 minutos. segmentos – um com PC alto (ÿ 4) e outro com PC baixo (< 4).

Resultados: Apesar de uma pesquisa de hiperparâmetros sobre a passada de dilatação do TCN, número de blocos residuais, taxa de aprendizagem, etc., não fomos capazes de treinar a rede de previsão do momento-chave usando a arquitetura Siamesa - a perda de treinamento nunca diminuiu significativamente. Curiosamente, o controle positivo (ou seja, o mesmo

arquitetura treinada para uma tarefa diferente, conforme descrito acima), apesar de precisar analisar uma série temporal muito mais longa (15min vs. 45-90seg), foi capaz de treinar com sucesso e alcançou uma AUC média de 0,82 (média de todos os 10 cruzamentos -dobras de validação).
Isso sugere que identificar momentos-chave com PC alto/baixo pode ser uma tarefa mais difícil do que estimar a pontuação agregada do PC em um vídeo inteiro, ou pode exigir uma arquitetura e um conjunto de recursos muito diferentes do problema de estimativa agregada do PC.

12 CONCLUSÕES

Desenvolvemos uma arquitetura multimodal de aprendizado de máquina e um procedimento de treinamento para criar uma Rede Automática de Reconhecimento de Observação em Sala de Aula (ACORN). O ACORN é, até onde sabemos, o primeiro sistema totalmente automatizado que pode analisar vídeos de salas de aula escolares e estimar as dimensões do Clima Positivo (PC) e do Clima Negativo (NC) do protocolo CLASS. O melhor sistema (modelo # 21) apresentado neste artigo pode prever PC e NC com uma precisão (correlação de Pearson) de 0,55 e 0,63 rótulos wrt fornecidos por codificadores CLASS especializados, o que é uma melhoria substancial em nosso trabalho anterior [15] (com Correlações de Pearson de 0,40 e 0,51 com base na verdade).

Esses níveis de precisão são semelhantes à confiabilidade entre codificadores de codificadores humanos. Também apresentamos resultados estatisticamente significativos (AUC=0,70) na detecção automática dos momentos-chave em um vídeo de sala de aula quando o PC está mais alto ou mais baixo.

12.1 Principais Resultados

Abaixo resumimos os principais resultados empíricos do nosso artigo: **Integração temporal:** (1) A integração temporal usando uma Rede Convolucional Temporal oferece precisão semelhante, mas é substancialmente mais rápida que um BiLSTM tanto para treinamento quanto para teste. (2)
A inclusão de módulos upstream adicionais (aqui os módulos upstream são redes neurais que extraem recursos dos diferentes
caminhos de informação diferentes) e caminhos de informação (expressões faciais, estimativas de emoções separadas de professores e alunos, quadro de imagem completo, etc.) são mais críticos do que a escolha dos módulos temporais a jusante (aqui os módulos a jusante são redes temporais que integram as várias entradas de os módulos upstream, como BiLSTM, TCN, etc.). (3) A dinâmica emocional de alunos e professores, e não apenas os seus valores emocionais médios, é importante para estimar CP e NC.

Escolha e qualidade dos recursos: (4) Maior precisão nos módulos upstream (por exemplo, mudança de VGG-16 para Resnet-50 para análise de rosto e imagem) se traduz em maior precisão nas previsões do modelo downstream (pontuações CLASS). Em outras palavras, as melhorias de precisão nas primeiras camadas perceptivas podem persistir ao longo de todo o gráfico computacional. (5) Embora características interpretáveis por humanos, como expressões faciais, sejam úteis para prever pontuações CLASS, informações complementares substanciais podem ser obtidas, embora ao preço da interpretabilidade, a partir da análise de entradas de baixo nível, como características de áudio espectral e os pixels de cada quadro de vídeo inteiro.

Via auditiva: (6) As características auditivas já fornecem poder preditivo não trivial (alcançando correlações em torno de 0,30), mas a adição de características visuais fornece informações complementares que aumentam ainda mais a precisão (para cerca de 0,60). É importante considerar isso ao avaliar privacidade versus precisão. (7) Ao utilizar apenas a via auditiva, foi alcançada uma precisão semelhante para os conjuntos de dados UVA Toddler e MET, apesar das diferentes faixas etárias e das diferentes definições do protocolo CLASS. (Observe que não A conclusão está disponível para a via visual, uma vez que não pudemos testá-la no conjunto de dados MET.)

Tamanho do conjunto de treinamento: (8) A precisão da previsão CLASS PC e NC da via auditiva aumenta constantemente até 3.500 exemplos de treinamento (segmentos de vídeo de 15 minutos); a trajetória sugere que continuará a aumentar.

PC vs NC: (9) Obtivemos consistentemente maior precisão na previsão de NC em comparação com a previsão de PC para os conjuntos de dados UVA e MET.

Prever os momentos-chave: (10) Prever os momentos-chave em que o Clima Positivo é alto/baixo parece ser uma tarefa mais difícil do que estimar a pontuação agregada do CLASS. Isto ocorre possivelmente porque muitos erros perceptuais momentâneos podem “atingir a média” ao longo de muitos intervalos de tempo. Em várias abordagens diferentes, incluindo CNN+TCN, rede siamesa e GCN+BiLSTM, descobrimos que a abordagem baseada em convolução de grafos funcionou melhor porque pode aproveitar interações entre diferentes participantes ponderadas pela proximidade entre si. Até onde sabemos, este é um dos primeiros resultados na literatura sobre a aplicação da convolução profunda de grafos para classificar a interação social entre humanos.

12.2 Pesquisa Futura

Existem diversas direções e questões de pesquisa que estamos considerando e/ou explorando ativamente. Para melhorar a precisão do modelo, estamos explorando: (1) Como podemos incluir informações linguísticas mais poderosas para a previsão da pontuação CLASS que vão além do que os recursos de áudio espectral de baixo nível podem capturar? Uma abordagem possível é treinar um classificador que possa estimar a complexidade linguística de um clipe de áudio como um recurso adicional. (2) Pode ser útil acompanhar as trajetórias de expressão de pessoas individuais na sala de aula ao longo do tempo, em vez de apenas tratar cada quadro como um “saco” de expressões. (3) Há amplo espaço para explorar o aproveitamento da abordagem GCN que usamos para previsão de momentos-chave para estimativa geral da pontuação CLASS. (4) A arquitetura para PC e NC generaliza para outras dimensões de CLASSE? Quais recursos adicionais seriam necessários?

Em última análise, a questão de investigação mais importante é sobre como tornar a observação automatizada da sala de aula mais útil para os professores: (5) A precisão do nosso actual ACORN (modelo #21) é suficientemente elevada para proporcionar experiências úteis de formação de professores e de desenvolvimento profissional? Estamos nos estágios iniciais da condução de um experimento para ver como os professores podem usar os resultados do nosso sistema para se tornarem mais perceptivos das interações em sala de aula e, eventualmente, implementarem interações mais eficazes em suas próprias salas de aula.

REFERÊNCIAS

[1] Robert C Pianta, Karen M La Paro e Bridget K Hamre. Classroom Assessment Scoring System™: Manual K-3. Publicação Paul H Brookes, 2008.

[2] Susan Kontos e Amanda Wilcox-Herzog. Interações dos professores com as crianças: por que são tão importantes? pesquisa em revisão. Crianças pequenas, 52(2):4–12, 1997.

[3] Andrew J Mashburn, Robert C Pianta, Bridget K Hamre, Jason T Downer, Oscar A Barbarin, Donna Bryant, Margaret Burchinal, Diane M Early e Carollee Howes. Medidas da qualidade da sala de aula na pré-escola e do desenvolvimento de habilidades acadêmicas, linguísticas e sociais das crianças. Desenvolvimento infantil, 79(3):732–749, 2008.

[4] Claire Cameron Ponitz, Megan M McClelland, JS Matthews e Frederick J Morrison. Uma observação estruturada da autorregulação comportamental e sua contribuição para os resultados do jardim de infância. Psicologia do desenvolvimento, 45(3):605, 2009.

[5] Deborah Lowe Vandell, Jay Belsky, Margaret Burchinal, Laurence Steinberg e Nathan Vandergrift. Os efeitos dos cuidados na primeira infância estendem-se até aos 15 anos de idade? resultados do estudo específico sobre cuidados na primeira infância e desenvolvimento de jovens. Desenvolvimento infantil, 81(3):737–756, 2010.

[6] Ellen S Peisner-Feinberg, Margaret R Burchinal, Richard M Clifford, Mary L Culkin, Carollee Howes, Sharon Lynn Kagan e Noreen Yazejian. A relação da qualidade do cuidado infantil pré-escolar com as trajetórias de desenvolvimento cognitivo e social das crianças até a segunda série.
Desenvolvimento infantil, 72(5):1534–1553, 2001.

[7] Barbara Nye, Spyros Konstantopoulos e Larry V Hedges. Quão grandes são os efeitos do professor? Avaliação educacional e análise de políticas, 26(3):237–257, 2004.

[8] Thomas J Kane, Daniel F McCaffrey, Trey Miller e Douglas O Staiger.
Identificamos professores eficazes? validar medidas de ensino eficaz usando atribuição aleatória. Em artigo de pesquisa. Projeto MET. Fundação Bill e Melinda Gates. Citeseer, 2013.

[9] Kenneth Holstein, Bruce M McLaren e Vincent Aleven. O aprendizado dos alunos se beneficia de uma ferramenta de conscientização de professores de realidade mista em salas de aula aprimoradas com IA. Na Conferência Internacional sobre Inteligência Artificial na Educação, páginas 154–168. Springer, 2018.

[10] Amanjot Kaur, Aamir Mustafa, Love Mehta e Abhinav Dhall. Previsão e localização do envolvimento dos alunos na natureza. Em 2018 Computação Digital de Imagens: Técnicas e Aplicações (DICTA), páginas 1–8. IEEE, 2018.

[11] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster e Javier R Movellan. As faces do envolvimento: Reconhecimento automático do envolvimento dos alunos a partir de expressões faciais. Transações IEEE em Computação Afetiva, 5(1):86–98, 2014.

[12] Beverly Park Woolf, Ivon Arroyo, David Cooper, Winslow Burleson e Kasia Muldner. Tutores afetivos: Detecção automática e resposta às emoções dos alunos. Em Avanços em sistemas de tutoria inteligentes, páginas 207–227. Springer, 2010.

[13] Sean Kelly, Andrew M Olney, Patrick Donnelly, Martin Nystrand e Sidney K D'Mello. Medir automaticamente a autenticidade das perguntas em salas de aula do mundo real. Pesquisador Educacional, 47(7):451–464, 2018.

[14] Karan Ahuja, Dohyun Kim, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan e Yuvraj Agarwal. Edusense: detecção prática em sala de aula em escala. Procedimentos da ACM sobre tecnologias interativas, móveis, vestíveis e ubíquas, 3(3):1–26, 2019.

[15] Anand Ramakrishnan, Erin Ottmar, Jennifer LoCasale-Crouch e Jacob Whitehill. Rumo à observação automatizada da sala de aula: Prevendo clima positivo e negativo. Em 2019, 14ª Conferência Internacional IEEE sobre Reconhecimento Automático de Rosto e Gestos (FG 2019), páginas 1–8. IEEE, 2019.

[16] Melinda T Owens, Shannon B Seidel, Mike Wong, Travis E Bejines, Susanne Lietz, Joseph R Perez, Shangheng Sit, Zahur-Saleh Subedar, Gigi N Acker, Susan F Akana, et al. O som da sala de aula pode ser usado para classificar práticas de ensino em cursos universitários de ciências. Anais da Academia Nacional de Ciências, 114(12):3085–3090, 2017.

[17] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka e Bernt Schiele. Um banco de dados para detecção detalhada de atividades culinárias. Em 2012, Conferência IEEE sobre Visão Computacional e Reconhecimento de Padrões, páginas 1194–1201. IEEE, 2012.

[18] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem e Juan Carlos Niebles. Activitynet: um benchmark de vídeo em grande escala para compreensão da atividade humana. Em Anais da conferência ieee sobre visão computacional e reconhecimento de padrões, páginas 961–970, 2015.

[19] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. O conjunto de dados de vídeo de ação humana cinética. Pré-impressão do arXiv arXiv:1705.06950, 2017.

[20] Nigel Bosch e Sidney D'Mello. Detecção automática de divagações mentais a partir de vídeos no laboratório e na sala de aula. Transações IEEE em Computação Afetiva, 2019.

[21] Thomas N Kipf e Max Welling. Classificação semissupervisionada com redes convolucionais de grafos. Pré-impressão do arXiv arXiv:1609.02907, 2016.

[22] Ting Yao, Tao Mei e Yong Rui. Detecção de destaque com classificação profunda em pares para resumo de vídeo em primeira pessoa. Em Anais da conferência IEEE sobre visão computacional e reconhecimento de padrões, páginas 982–990, 2016.

[23] Ashish Kapoor, Selene Mota, Rosalind W Picard, et al. Rumo a um companheiro de aprendizagem que reconhece o afeto. No simpósio AAAI Fall, número 543, páginas 2–4, 2001.

[24] Goren Gordon, Samuel Spaulding, Jacqueline Kory Westlund, Jin Joo Lee, Luke Plummer, Marayna Martinez, Madhurima Das e Cynthia Breazeal. Personalização afetiva de um robô social tutor para crianças
habilidades de segunda língua. Na Trigésima Conferência AAAI sobre Inteligência Artificial, 2016.

[25] Joseph Grafsgaard, Joseph B Wiggins, Kristy Elizabeth Boyer, Eric N Wiebe e James Lester. Reconhecendo automaticamente a expressão facial: Prevendo envolvimento e frustração. Em Mineração de Dados Educacionais, 2013.

[26] Nigel Bosch, Sidney D'Mello, Ryan Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang e Weinan Zhao. Detecção automática de estados afetivos centrados na aprendizagem em estado selvagem. Em Internacional conferência sobre interfaces de usuário inteligentes, 2015.

[27] Tsung-Yen Yang, Ryan S Baker, Christoph Studer, Neil Heffernan e Andrew S Lan. Aprendizagem ativa para detecção de afeto do aluno. Nos Anais da 12ª Conferência Internacional sobre Mineração de Dados Educacionais (EDM 2019), páginas 208–217. ÉRIC, 2019.

[28] Zachary A Pardos, Ryan SJD Baker, Maria OCZ San Pedro, Sujith M Gowda e Supreeth M Gowda. Estados afetivos e testes de estado: Investigar como o afeto ao longo do ano letivo prevê os resultados de aprendizagem no final do ano. Na Conferência Internacional sobre Análise de Aprendizagem e Conhecimento, páginas 117–124. ACM, 2013.

[29] Anthony F Botelho, Ryan S Baker, Jaclyn Ocumpaugh e Neil T Heffernan. Estudando dinâmica de afeto e cronometria usando detectores sem sensor. Sociedade Internacional de Mineração de Dados Educacionais, 2018.

[30] Ashish Kapoor, Winslow Burleson e Rosalind W Picard. Previsão automática de frustração. Jornal internacional de estudos de computador humano, 65(8):724–736, 2007.

[31] Wanli Xing, Hengtao Tang e Bo Pei. Além das emoções positivas e negativas: analisando o papel das emoções de realização nos fóruns de discussão dos moocs. A Internet e o Ensino Superior, 43:100690, 2019.

[32] Sidney D'Mello, Rosalind W Picard e Arthur Graesser. Em direção a um autotutor sensível ao afeto. Sistemas Inteligentes IEEE, 22(4):53–61, 2007.

[33] Ivon Arroyo, David G Cooper, Winslow Burleson, Beverly Park Woolf, Kasia Muldner e Robert Christopherson. Sensores de emoção vão para a escola. Em AIED, volume 200, páginas 17–24, 2009.

[34] Robin Cosbey, Allison Wusterbarth e Brian Hutchinson. Aprendizado profundo para detecção de atividades em sala de aula a partir de áudio. Na ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), páginas 3727–3731. IEEE, 2019.

[35] Patrick J Donnelly, Nathaniel Blanchard, Borhan Samei, Andrew M Olney, Xiaoyi Sun, Brooke Ward, Sean Kelly, Martin Nystrand e Sidney K D'Mello. Modelagem multissensor de segmentos instrucionais de professores em salas de aula ao vivo. Na conferência internacional ACM sobre interação multimodal, páginas 177–184. ACM, 2016.

[36] Qifeng Qiao e Peter A Beling. Avaliação e recuperação de vídeo em sala de aula por meio de aprendizagem em múltiplas instâncias. Na Conferência Internacional sobre Inteligência Artificial na Educação, páginas 272–279. Springer, 2011.

[37] Anusha James, Mohan Kashyap, Yi Han Victoria Chua, Tomasz Maszczyk, Ana Moreno Núñez, Rebecca Bull e Justin Dauwels.
Inferindo o clima nas salas de aula a partir de gravações de áudio e vídeo: uma abordagem de aprendizado de máquina. Em 2018, Conferência Internacional IEEE sobre Ensino, Avaliação e Aprendizagem para Engenharia (TALE), páginas 983–988. IEEE, 2018.

[38] Anthony F Botelho, Ryan S Baker e Neil T Heffernan. Melhorando a detecção de efeitos sem sensores usando aprendizado profundo. Na Conferência Internacional sobre Inteligência Artificial na Educação, 2017.

[39] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei e Yaser Sheikh.
Openpose: estimativa de pose 2D de várias pessoas em tempo real usando campos de afinidade de peças. Pré-impressão do arXiv arXiv:1812.08008, 2018.

[40] Tadas Baltrušaitis, Peter Robinson e Louis-Philippe Morency. Open-face: um kit de ferramentas de análise de comportamento facial de código aberto. Em Aplicações de Visão Computacional (WACV), 2016.

[41] Sidney K D'Mello, Andrew M Olney, Nathan Blanchard, Borhan Samei, Xiaoyi Sun, Brooke Ward e Sean Kelly. Captura multimodal de interações professor-aluno para análise dialógica automatizada em salas de aula ao vivo. Em Internacional conferência sobre interação multimodal, 2015.

[42] Zuowei Wang, Xingyu Pan, Kevin F Miller e Kai S Cortina. Classificação automática de atividades no discurso da sala de aula. Computadores e Educação, 78:115–123, 2014.

[43] Anusha James, Yi Han Victoria Chua, Tomasz Maszczyk, Ana Moreno Núñez, Rebecca Bull, Kerry Lee e Justin Dauwels. Classificação automatizada do clima da sala de aula por análise de áudio. No 9º Workshop Internacional sobre Tecnologia de Sistemas de Diálogo Falado, páginas 41–49.
Springer, 2019.

[44] Thomas J Kane e Douglas O Staiger. Coletar feedback para o ensino: Combinar observações de alta qualidade com pesquisas de alunos e ganhos de desempenho. artigo de pesquisa. projeto conhecido. Fundação Bill e Melinda Gates, 2012.

[45] Lia E Sandilos e James C DiPerna. Confiabilidade entre avaliadores do sistema de pontuação de avaliação em sala de aula-pré-k (classe pré-k). Jornal da Primeira Infância e Psicologia Infantil, (7), 2011.

[46] Shaojie Bai, J Zico Kolter e Vladlen Koltun. Uma avaliação empírica de redes convolucionais e recorrentes genéricas para modelagem de sequências. Pré-impressão do arXiv arXiv:1803.01271, 2018.

[47] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior e Koray Kavukcuoglu. Wavenet: Um modelo generativo para áudio bruto. Pré-impressão do arXiv arXiv:1609.03499, 2016.

[48] Fundação Bill e Melinda Gates. Medidas de ensino eficaz: 2 - arquivos principais, 2009-2011, 2014.

[49] B Hamre, J LoCasale-Crouch, F Romo e J Whittaker. O curso eficaz de interações em sala de aula para professores de cuidados infantis: resultados preliminares de um ensaio de controle randomizado. Na Conferência Nacional de Pesquisa sobre a Primeira Infância, 2018.

[50] Hamed Monkaresi, Nigel Bosch, Rafael A Calvo e Sidney K D'Mello. Detecção automatizada de envolvimento usando estimativa baseada em vídeo de expressões faciais e frequência cardíaca. Transações IEEE em Computação Afetiva, 8(1):15–28, 2016.

[51] Yelin Kim e Jeesun Kim. Reconhecimento de emoções semelhantes às humanas: aprendizagem multi-rótulo a partir de fala expressiva audiovisual rotulada com ruído. Em 2018, Conferência Internacional IEEE sobre Acústica, Fala e Processamento de Sinais (ICASSP), páginas 5104–5108. IEEE, 2018.

[52] Arkar Min Aung e Jacob Whitehill. Aproveitando a incerteza do rótulo para melhorar a modelagem: uma aplicação para o reconhecimento do envolvimento dos alunos. Em FG, páginas 166–170, 2018.

[53] Fundação Gates. Aprendendo sobre o ensino: descobertas iniciais das medidas de projetos de ensino eficazes. https://docs.gatesfoundation.org/documents/preliminary-findings-research-paper.pdf , 2020. Acesso em: 17/08/2020.

[54] László A Jeni, Jeffrey M Girard, Jeffrey F Cohn e Fernando De La Torre. Estimativa contínua de intensidade de au usando espaço de características faciais esparsas e localizadas. Em 2013, 10ª conferência internacional IEEE e workshops sobre reconhecimento automático de rosto e gestos (FG), páginas 1–7. IEEE, 2013.

[55] Theodoros Giannakopoulos. pyaudioanálise: Uma biblioteca python de código aberto para análise de sinal de áudio. PloS um, 10(12), 2015.

[56]Leo Breiman. Árvores de classificação e regressão. Routledge, 2017.

[57] Brian Zylich e Jacob Whitehill. Detectores de frases-chave robustos a ruído para feedback automatizado em sala de aula. Na ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), páginas 9215–9219. IEEE, 2020.

[58] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Fala profunda: ampliando o reconhecimento de fala de ponta a ponta. Pré-impressão do arXiv arXiv:1412.5567, 2014.

[59] Rúben Morais e Kelly Davis. Documentação do Deepspeech!

[60] Ali Mollahosseini, Behzad Hasani e Mohammad H Mahoor. Affect-net: Um banco de dados para expressão facial, valência e computação de excitação em estado selvagem. Pré-impressão do arXiv arXiv:1708.03985, 2017.

[61] Huaizu Jiang e Erik Learned-Miller. Detecção de rosto com o r-cnn mais rápido. Em Reconhecimento Automático de Rosto e Gestos do IEEE, 2017.

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ŸukaszKaiser e Illia Polosukhin. Atenção é tudo que você precisa. Dentro Avanços em sistemas de processamento de informação neural, páginas 5998–6008, 2017.

[63] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens e Quoc V Le. Atenção redes convolucionais aumentadas. Em Anais da Conferência Internacional IEEE sobre Visão Computacional, páginas 3286–3295, 2019.

[64] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya e Jonathon Shlens. Autoatenção autônoma em modelos de visão. Pré-impressão do arXiv arXiv:1906.05909, 2019.

[65] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel e Yoshua Bengio. Mostre, assista e conte: geração de legendas de imagens neurais com atenção visual. Na conferência internacional sobre aprendizado de máquina, páginas 2048–2057, 2015.

[66] Minh-Thang Luong, Hieu Pham e Christopher D Manning. Abordagens eficazes para tradução automática neural baseada na atenção. Pré-impressão do arXiv arXiv:1508.04025, 2015.

[67] Uwe Ligges, Sebastian Krey, Olaf Mersmann e Sarah Schnackenberg. tuneR: Análise de Música e Fala, 2018.

[68] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu e Xian-Sheng Hua. Autoencoder espaço-temporal para detecção de anomalias de vídeo. Dentro Anais da 25ª conferência internacional ACM sobre Multimídia, páginas 1933–1941, 2017.

[69] Huijuan Xu, Abir Das e Kate Saenko. R-c3d: Rede 3D convolucional regional para detecção de atividade temporal. Em Anais da conferência internacional IEEE sobre visão computacional, páginas 5783–5792, 2017.

[70] Minghui Zhang, Yumeng Liang e Huadong Ma. Raciocínio gráfico afetivo sensível ao contexto para reconhecimento de emoções. Na Conferência Internacional sobre Multimídia e Expo, páginas 151–156. IEEE, 2019.

[71] Junhyun Lee, Inyeop Lee e Jaewoo Kang. Agrupamento de gráficos de autoatenção. Pré-impressão do arXiv arXiv:1904.08082, 2019.

Anand Ramakrishnan Anand Ramakrishnan é doutorando no Departamento de Ciência da Computação do Worcester Polytechnic Institute (WPI) e trabalha com o Prof. Jacob Whitehill. Sua pesquisa envolve a utilização de métodos de aprendizado de máquina multimodais para capturar interações professor-aluno.

Ele também tem interesse na área de robótica e aprendizagem única. Ele possui mestrado pela WPI em Engenharia Robótica e BE pela SRM University em Engenharia Mecânica.

Brian Zylich Brian Zylich é estudante de doutorado na Faculdade de Ciências da Informação e da Computação da Universidade de Massachusetts, Amherst. Seus interesses de pesquisa estão em aprendizado de máquina multimodal, processamento de linguagem natural e suas aplicações na educação. Ele possui bacharelado e mestrado pelo Worcester Polytechnic Institute.

Erin Ottmar Erin Ottmar é professora assistente de Ciências da Aprendizagem e Psicologia no Worcester Polytechnic Institute. Ela recebeu seu doutorado. em Psicologia Educacional pela Universidade da Virgínia. Ela conduz pesquisas sobre mecanismos cognitivos, perceptivos e sociais de aprendizagem e examina como as tecnologias e intervenções podem facilitar os processos de instrução e aprendizagem. Ela tem experiência na concepção, desenvolvimento e avaliação em larga escala de tecnologias de aprendizagem para alunos e professores.

Jennifer LoCasale-Crouch Jennifer LoCasale-Crouch é professora pesquisadora associada no Centro de Estudos Avançados de Ensino e Aprendizagem da Universidade da Virgínia. Ela recebeu seu bacharelado e mestrado pela Florida State University e doutorado. em Risco e Prevenção em Ciências da Educação pela Universidade da Virgínia. Ela conduz pesquisas sobre os apoios e sistemas que influenciam as crianças e como podemos melhorar essas experiências para apoiar o desenvolvimento das crianças.

Jacob Whitehill Jacob Whitehill trabalha no Departamento de Ciência da Computação em Worcester Instituto Politécnico e membro do programa Learning Science & Technologies. Sua pesquisa está em aprendizado de máquina multimodal, computação afetiva e aprendizado único, bem como suas aplicações à medição educacional. Ele possui doutorado pela Universidade da Califórnia, San Diego, mestre pela Universidade do Western Cape e bacharelado em Stanford.