Processamento de sinal, imagem e vídeo (2024) 18:3553–3563
https://doi.org/10.1007/s11760-024-03020-8

PAPEL ORIGINAL

Um modelo leve de reconhecimento de expressão facial para detecção automatizada de engajamento

Zibin Zhao1 · Yinbei Li2 · Jiaqiang Yang2 · Yuliang Ma1

Recebido: 14 de agosto de 2023 / Revisado: 25 de dezembro de 2023 / Aceito: 14 de janeiro de 2024 / Publicado online: 25 de fevereiro de 2024

Abstrato

O monitoramento em tempo real do nível de envolvimento dos alunos na sala de aula é de suma importância na educação moderna. O reconhecimento de expressões faciais tem sido extensivamente explorado em vários estudos para atingir esse objetivo. No entanto, os modelos convencionais muitas vezes lidam com um elevado número de parâmetros e custos computacionais substanciais, limitando a sua praticidade em aplicações em tempo real e em cenários do mundo real. Para resolver esta limitação, este artigo propõe “Light_Fer”, um modelo leve projetado para atingir alta precisão enquanto reduz parâmetros. A novidade do Light_Fer reside na integração de convolução separável em profundidade, convolução de grupo e estrutura de gargalo invertido. Essas técnicas otimizam a arquitetura dos modelos, resultando em precisão superior com menos parâmetros. Resultados experimentais demonstram que Light_Fer, com apenas 0,23M de parâmetros, atinge precisões notáveis de 87,81% e 88,20% em conjuntos de dados FERPLUS e RAF-DB, respectivamente. Além disso, ao estabelecer uma correlação entre as expressões faciais e os níveis de envolvimento dos alunos, estendemos a aplicação do Light_Fer para detecção e monitoramento em tempo real do envolvimento dos alunos durante as atividades em sala de aula. Concluindo, o modelo Light_Fer proposto, com seu design leve e precisão aprimorada, oferece uma solução promissora para monitoramento do envolvimento dos alunos em tempo real por meio do reconhecimento de expressões faciais.

Detecção de engajamento **de palavras-chave** · Reconhecimento de expressão facial · Modelo leve

1. Introdução

A educação é um caminho importante para herdar a civilização e o conhecimento, nutrir a geração mais jovem e criar uma melhor qualidade de vida. Um aspecto essencial da educação eficaz reside em orientar os alunos para melhorar a sua eficácia auditiva na sala de aula. A premissa para atingir esse objetivo é compreender prontamente a situação dos alunos em sala de aula, tendo o envolvimento dos alunos como medida fundamental. De acordo com para o Conselho Australiano de Pesquisa em Educação [1], o envolvimento dos alunos abrange atividades e condições que promovem experiências de aprendizagem de alta qualidade. O impacto do envolvimento no sucesso dos alunos [2, 3], conforme apoiado pela pesquisa de Trowler [4], destaca uma forte correlação entre o nível de envolvimento e os resultados positivos da aprendizagem no desenvolvimento dos alun No entanto, os resultados de pesquisas mostram que apenas 46% a 67% dos alunos participam ativamente da sala de aula, enquanto uma parcela significativa requer atenção e orientação dos professores [5].

A detecção em tempo real do envolvimento dos alunos representa um desafio considerável. Em resposta às exigências da educação moderna, os investigadores têm tentado alcançar a detecção automática sem contacto físico. O rápido desenvolvimento da aprendizagem profunda nos últimos anos desbloqueou inúmeras aplicações, incluindo detecção de objetos, detecção de rostos, reconhecimento de expressões, etc., fornecendo uma base técnica para o uso da aprendizagem profunda para detectar engajamento.

Entre os estudos, muitos pesquisadores ressaltaram a importância de monitorar as expressões faciais dos alunos [6]. A expressão facial serve como uma forma primária de expressão emocional para os indivíduos, com a teoria de Mehrabian

B Jiaqiang Yang
yangjiaq@zju.edu.cn

B Yuliang Ma
mayuliang@hdu.edu.cn

Zibin Zhao
zhaozb1999@163.com

Yinbei Li
liyinbei@outlook.com

1 Escola de Automação, Universidade Hangzhou Dianzi, Hangzhou 310018, China

2 Faculdade de Engenharia Elétrica, Universidade de Zhejiang, Hangzhou 310027, China

[7] sugerindo que a expressão facial é responsável por 55% informações transmitidas. Categorização de humano de Ekman expressão em seis emoções básicas (ou seja, “triste”, “feliz”, “raiva”, “nojo”, “surpresa” e “medo”) [8] reforça ainda mais a importância dos sinais faciais. Na educação contexto, os pesquisadores associaram expressões faciais a envolvimento dos alunos, pois oferecem informações valiosas sobre seus estados emocionais e níveis de compreensão na sala de aula [9]. Trabalho pioneiro de Khawlah Altuwairqi [10] inclui experimentos sociais que mapeiam a expressão facial para o níveis de engajamento correspondentes, estabelecendo base para detectar envolvimento com base na expressão facial.

O núcleo do aprendizado profundo está na construção de modelos de redes neurais, com o tamanho do modelo representado pelo número de parâmetros. Os modelos de aprendizagem profunda normalmente consistem em múltiplas camadas, cada uma contendo vários parâmetros. Complexo modelos com um grande número de parâmetros possuem capacidades de ajuste. No entanto, também implicam maiores custos operacionais e de armazenamento, o que pode ser problemático para plataformas de implantação com recursos limitados. Portanto, encontrar um equilíbrio entre precisão e tamanho do modelo é crucial. Neste artigo, nos concentramos no design de um modelo leve para atender aos requisitos de implantação em cenários do mundo real.

Em resumo, este artigo apresenta as três contribuições a seguir: mas:

- Um modelo leve de reconhecimento de expressão facial “Light_Fer” é introduzido. É especialmente projetado para detecção automática em tempo real do envolvimento dos alunos nível em ambientes de sala de aula. Light_Fer estabelece um novo padrão ao alcançar precisão de última geração enquanto minimiza os parâmetros do modelo para eficiência.
- O bloco residual no gargalo do grupo é projetado como um bloco de construção fundamental para Light_Fer. Este módulo fundamental integra convolução separável em profundidade, convolução de grupo, estruturas de gargalo invertido e incorpora conexões residuais, coletivamente melhorando o desempenho geral e a flexibilidade do modelo.
- Experimentos extensos validam a eficácia do propôs Light_Fer. Ele atinge a maior precisão em conjunto de dados FERPLUS quando comparado com grandes e modelos leves. Além disso, demonstra resultados comparáveis no conjunto de dados RAF-DB, reafirmando ainda mais a sua eficácia geral.

As seções restantes deste artigo estão organizadas da seguinte forma: Seção. 2 fornece uma visão geral das pesquisas relacionadas; Seita. 3 apresenta uma descrição detalhada do modelo proposto; Seita. 4 cobre treinamento de modelo e apresenta experimental junto com uma comparação com outros modelos existentes; Seita. 5 conclui o trabalho e traça perspectivas futuras.

2. Trabalho relacionado

2.1 Pesquisa sobre detecção de engajamento

Nos últimos anos, esforços substanciais de investigação concentraram-se em a detecção do envolvimento dos alunos [11, 12]. Entre estes abordagens, os métodos baseados em expressões faciais ganharam destaque devido à sua acessibilidade e precisão [13, 14]. Notavelmente, Grafsgaard et al. [15] utilizou a ação facial sistema de codificação para monitorar os movimentos faciais dos alunos, como como movimentos das sobrancelhas, aperto das pálpebras e canto da boca levantamento, durante sessões de aprendizagem on-line. Seu trabalho estabeleceu uma ligação direta entre a intensidade da expressão facial e eficácia do ensino, enfatizando o potencial do tratamento facial reconhecimento de expressão (FER) na educação.

Dubbaka et al. [16] empregaram câmeras para registrar as expressões faciais dos alunos enquanto assistiam às aulas. vídeos, possibilitando uma análise de seus níveis de engajamento. Eles aplicou uma rede neural convolucional para detectar ações faciais unidades e usou regressão vetorial de suporte para mapear essas unidades aos estados emocionais e à atenção, alcançando um notável 95% de precisão entre os assuntos. Shen et al. [17] proposto um modelo de avaliação do envolvimento na aprendizagem que introduziu reconhecimento de expressão facial adaptável ao domínio para capturar mudanças emocionais dos alunos. Os resultados experimentais demonstraram a eficácia do modelo na avaliação dos alunos envolvimento durante o processo de aprendizagem.

Liao et al. [18] introduziram um modelo inovador para previsão de emoções, incorporando um SE-ResNet-50 pré-treinado para extrair características espaciais faciais e um longo curto prazo rede de memória (LSTM) com atenção global (GALN) para gerando estados de atenção. Este modelo capturou tanto o rosto informações espaciais e temporais, permitindo percepção do estado de engajamento e melhoria do desempenho de previsão de engajamento.

Swadha Gupta et al. [19] conduziram uma análise aprofundada de expressões faciais e classificação de emoções dos alunos, levando ao cálculo de um índice de engajamento (IE) usado para prever dois estados de engajamento: “engajado” e “desengajado”. O desempenho do sistema foi avaliado usando conjuntos de dados como FER2013, CK+ e RAF-DB, com os selecionados Modelo ResNet-50 alcançando os melhores resultados.

2.2 Pesquisa sobre reconhecimento de expressões faciais

No domínio da detecção de engajamento por meio da análise de expressões faciais, os modelos de reconhecimento de expressões faciais servem como um componente fundamental. Embora o reconhecimento de expressões faciais tenha recebido atenção substancial, houve limitação foco no desenvolvimento e implementação de soluções leves modelos.

Hewitt e Gunes et al. [20] propuseram três variações de modelos CNN bem estabelecidos, nomeadamente a variante AlexNet,

Variante VGGNet e variante MobileNet. Esses modelos foram treinados e validados usando o conjunto de dados Affectnet, alcançando uma precisão de até 58%. Posteriormente, eles foram implantados em dispositivos móveis.

Barros e cols. [21] introduziram um modelo leve de reconhecimento de expressão facial chamado FaceChannel. Este modelo apresenta dez camadas convolucionais e quatro camadas de pooling em seu processo de extração de características, seguidas por uma camada totalmente conectada. Além disso, campos inibitórios de desvio foram aplicados à camada final, produzindo uma precisão de 80,54% no conjunto de dados FABO.

Ferro-Perez e Mitre-Hernandez et al. [22] desenvolveram outro modelo leve de reconhecimento de expressão conhecido como ResMoNet. Este modelo é baseado nos princípios de convolução separável profunda e conexão residual. Consiste em cinco módulos principais: bloco de haste, bloco móvel, bloco residual, bloco de transição e bloco denso. Notavelmente, alcançou uma precisão de 90% no conjunto de dados RAFD.

Zengqun Zhao et al. [23] introduziram o modelo Efficientface, que aumenta a robustez da rede leve através de um extrator de características locais e um modulador espacial de canal, incorporando convolução em profundidade. Reconhecendo que as emoções são muitas vezes uma mistura de emoções básicas, eles introduziram um método simples, mas eficaz, de aprendizagem de distribuição de rótulos (LDL) como uma nova estratégia de treinamento. Seu modelo apresenta uma precisão impressionante de 88,36% quando avaliado usando o conjunto de dados RAF-DB.

Além disso, os desenvolvimentos recentes no domínio dos produtos leves modelos levaram a abordagens notáveis, incluindo Xception [24], SqueezeNet [25], MobileNet [26–28] e ShuffleNet [29, 30]. Esses modelos se distinguem por seus recursos arquitetônicos, como convoluções separáveis em profundidade, convoluções 1x1, blocos residuais invertidos e convoluções de grupo. Eles demonstraram melhorias notáveis de desempenho e, ao mesmo tempo, minimizaram a sobrecarga computacional.

3 Método

3.1 Fluxo de trabalho geral

O método de detecção de engate proposto está descrito na Fig. 1 e consiste nas seguintes etapas sequenciais:

1. Processamento de entrada:

- O vídeo de aprendizagem do aluno é inserido no sistema, onde o OpenCV é usado para extrair imagens de quadros individuais do vídeo.
- O algoritmo Haarcascade no OpenCV é usado para detectar o rosto na imagem e a região facial é enquadrada de acordo.
- As imagens da região facial são então cortadas e ajustado para um tamanho de 48×48.

2. Reconhecimento de expressão:

- As imagens faciais recortadas são encaminhadas para módulo de reconhecimento de expressão.
- O módulo realiza a classificação e atribui cada imagem a uma das oito categorias: raiva, nojo, medo, alegria, tristeza, surpresa, neutro ou desprezo.

3. Mapeamento do nível de engajamento:

- Os resultados das emoções detectadas são mapeados em três níveis de engajamento: alto, médio e baixo, com base na relação estabelecida entre expressão facial e níveis de engajamento.

3.2 Arquitetura geral do modelo

A pedra angular da detecção de engajamento neste estudo é o desenvolvimento de um modelo leve de reconhecimento de expressão facial denominado Light_Fer. A arquitetura geral do modelo proposto é mostrada na Fig. 2. Inspirando-se na arquitetura do mini_Xception, Light_Fer começa com duas camadas de convolução padrão, cada uma seguida por uma camada de normalização em lote e função de ativação Mish. Posteriormente, quatro blocos residuais de gargalo de grupo e (módulo de atenção do bloco de convolução) módulos CBAM são incorporados.

Em seguida, outra camada de convolução padrão e um módulo de módulo de atenção de canal (CAM) são introduzidos.
Finalmente, a camada de pooling de média global e a função Softmax são aplicadas para classificação de emoções.

3.3 Bloco residual no gargalo do grupo

Para o nosso modelo proposto, o bloco residual do grupo no gargalo desempenha um papel fundamental na obtenção de alta precisão, conforme mostrado na Fig. 3. Usamos quatro desses blocos para extrair completamente os recursos da imagem.

O bloco residual em gargalo do grupo consiste em duas vias paralelas. O caminho esquerdo é composto por duas camadas de grupo no gargalo (G no gargalo) seguidas por uma camada de pooling máximo. Em contraste, o caminho certo é uma conexão residual envolvendo uma convolução regular com o tamanho do kernel de 1x1. A integração de uma conexão residual permite que o modelo aprenda a diferença entre o mapa de características original e as características desejadas, mitigando assim o risco de desaparecimento do gradiente. Além disso, a adoção do gargalo do grupo de duas camadas permite uma extração de recursos mais eficiente com

menos parâmetros. O bloco residual do gargalo do grupo serve como um componente crítico na arquitetura do modelo, otimizando o compromisso entre eficiência computacional e representação de recursos, contribuindo em última análise para um desempenho superior de detecção de engajamento.

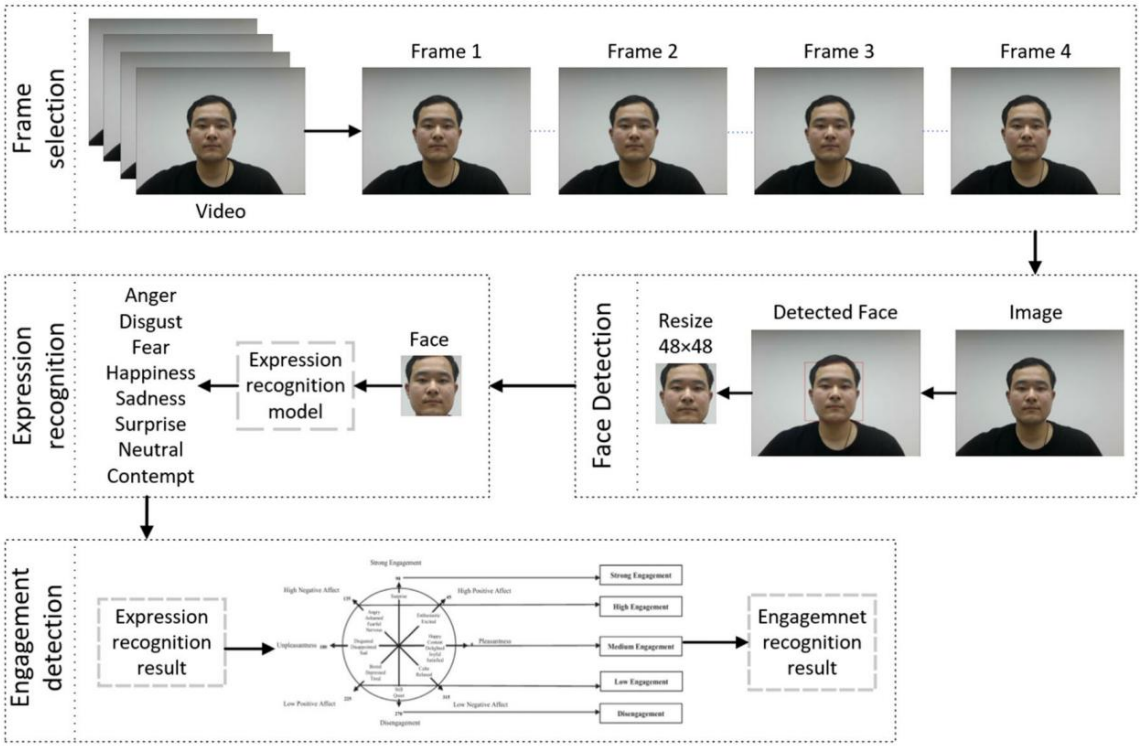

Fig. 1 Fluxo de trabalho geral de detecção de envolvimento dos alunos

3.4 Gargalo do grupo

O gargalo do Grupo serve como a essência do bloco residual mencionado acima, incorporando três operações principais:
convolução separável em profundidade, gargalo invertido e
convolução do grupo. Estas operações contribuem sinergicamente para um número reduzido de parâmetros, preservando
excelentes recursos de extração de recursos.

Convolução separável em profundidade, ilustrada na Fig. 4, constitui a operação fundamental em nosso modelo. Isto
decompõe uma convolução regular em duas etapas separadas:
convolução profunda e convolução pontual.

No processo de convolução em profundidade, cada canal do o mapa de recursos é convolvido independentemente por um kernel de convolução, com o kernel tendo uma contagem de canais de 1. O resultado mapas de características são então concatenados para capturar correlações espaciais na dimensão do canal, preservando ao mesmo tempo a dimensão do canal. número do mapa de recursos de entrada.

Ao mesmo tempo, no processo de convolução pontual, um Kernel 1×1×M é empregado, onde M é o número de canais na camada anterior. Esta operação utiliza o correlações entre canais para realizar combinações ponderadas do mapa de características anterior na dimensão de profundidade. Como um resultado, o número de canais no mapa de recursos de saída é determinado pelo número de núcleos convolucionais. Por totalmente aproveitando informações espaciais e de profundidade, separação em profundidade convolução confiável garante nenhum sacrifício na extração de recursos capacidade.

Em relação ao número de parâmetros, assumindo M canais no mapa de características de entrada e N canais na saída
mapa de recursos, com um tamanho de kernel de convolução de D×D, o número de parâmetros na convolução padrão é D×D×M×N.
Em contraste, o número total de parâmetros para profundidade convolução separável é D×D×M+M×N, que é 1/N + 1/(D×D) vezes a convolução padrão. Esta redução significativa nos parâmetros aumenta a eficiência do modelo.

Para aprimorar ainda mais os recursos de extração de recursos, introduzimos uma estrutura de gargalo invertida. Consiste em três camadas de convolução separável dpethwise, esta estrutura incorpora uma camada de expansão para aumentar o número de canais, uma camada de manutenção para manter o número expandido de canais e uma camada de redução para alinhar com operações subsequentes. A taxa de expansão empregada neste artigo garante que o número de canais na camada keep é o dobro do a camada de redução.

Esta configuração aumenta a profundidade do modelo enquanto ampliando o número de canais, permitindo extrair mais informações e melhorar a capacidade de adaptação.

Para contrariar o aumento do número de parâmetros introduzido pelo gargalo invertido, integramos o grupo convolução para obter maior redução de parâmetros. Esse a convolução de grupo divide o mapa de recursos de entrada em grupos distintos com base em canais, e cada grupo sofre

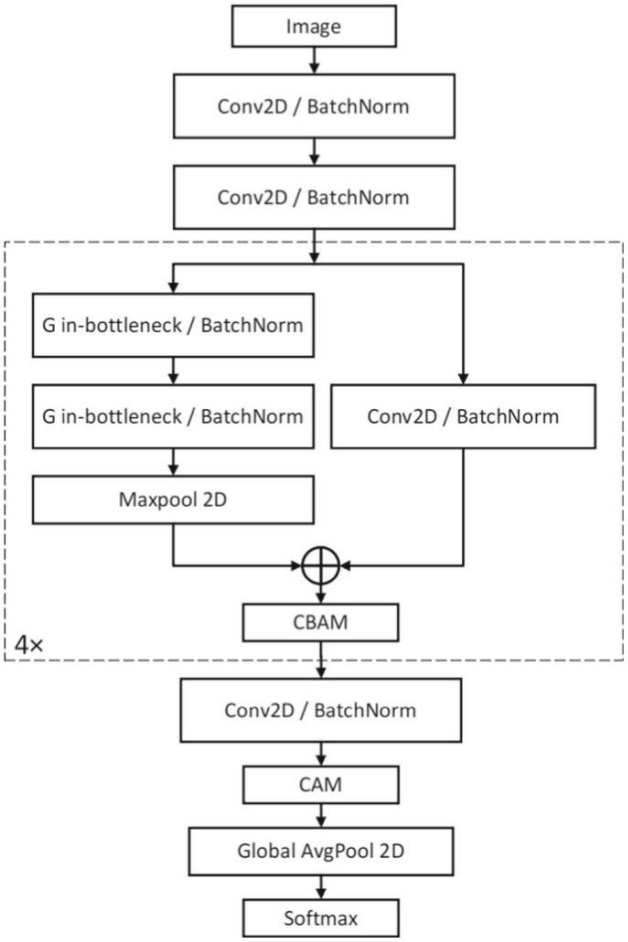

Fig. 2 Estrutura geral do nosso modelo

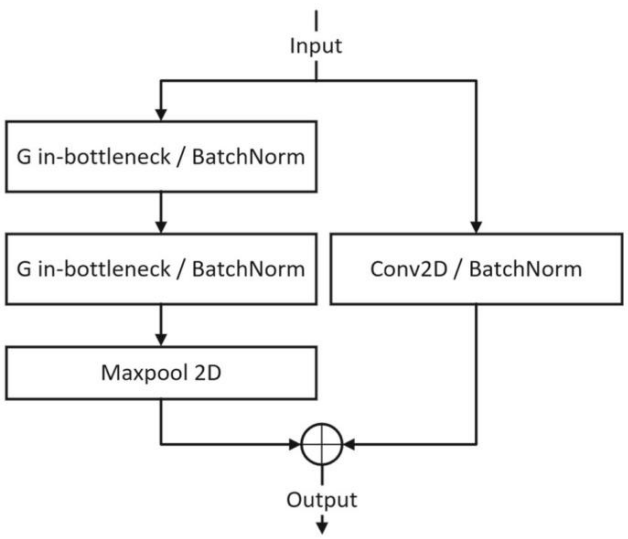

Fig. 3 Estrutura do bloco residual no gargalo do grupo

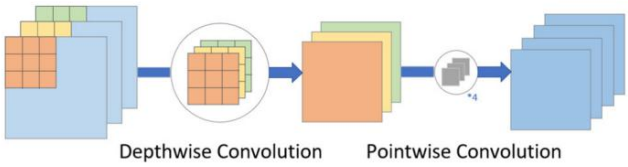

Fig. 4 Estrutura da convolução separável em profundidade

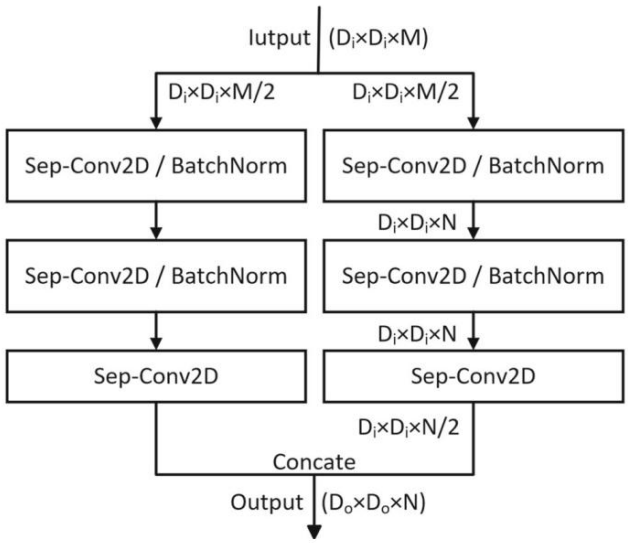

Fig. 5 Estrutura do gargalo agrupado

convolução independente com sua convolução correspondente
núcleo. As saídas de todos os grupos são então concatenadas para
obter o mapa de recursos de saída final. Com grupos C, o número de
parâmetros de convolução do grupo é apenas 1/C daquele
na convolução padrão. Além disso, devido aos núcleos de convolução
distintos em cada grupo, a convolução do grupo aumenta
a não linearidade do modelo.

Para aproveitar os benefícios da convolução do grupo e
gargalo invertido, nós os combinamos com dois grupos para
cada caminho, passando individualmente por um gargalo invertido. Por fim,
eles são concatenados ao longo da dimensão do canal para formar o
gargalo agrupado (G in-gargalo),
servindo como elemento central do bloco residual do gargalo do grupo,
conforme ilustrado na Figura 5.

3.5 Módulo CBAM

O mecanismo de atenção é um método poderoso para aprender
a importância dos dados de entrada e ponderá-los eficazmente. Isto
pode melhorar significativamente o desempenho do modelo com o mínimo
parâmetros adicionais. Neste estudo, utilizamos o módulo de atenção de
bloco convolucional (CBAM) [31], que consiste
de módulo de atenção de canal conectado (CAM) e espacial
módulo de atenção (SAM) sequencialmente, conforme mostrado na Fig .

CAM é responsável por identificar recursos significativos
submetendo o mapa de recursos ao pooling máximo paralelo e

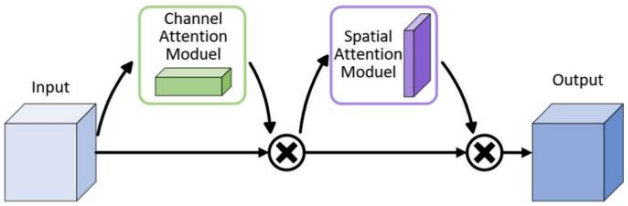

Fig. 6 Estrutura do mecanismo de atenção do CBAM

camadas médias de pooling, mantendo a dimensão do canal enquanto comprime a dimensão espacial.

Por outro lado, o SAM realiza pooling máximo serial e pooling médio na entrada, preservando a dimensão espacial enquanto comprime a dimensão do canal. O mapa de características enviado ao módulo CBAM é multiplicado elemento a elemento com o canal gerado e o mapa de atenção espacial.

Em nosso modelo, um módulo CBAM é colocado após cada bloco residual do gargalo do grupo. Embora um módulo CBAM típico empregue um tamanho de kernel 7×7 no SAM para melhor compreender as informações globais, descobrimos que usando um tamanho de kernel 7×7 para o SAM do módulo final, com um mapa de recursos de 3×3, requer preenchimento. Isso pode introduzir informações menos relevantes aos recursos e interferir no processo de extração de recursos, diminuindo potencialmente o desempenho do modelo. Como resultado, utilizamos um tamanho de kernel 3×3 para SAM no último módulo CBAM, que foi validado para melhorar o desempenho do modelo. Além disso, um módulo CAM é adicionado antes da camada final de pooling médio global para aumentar ainda mais o foco em canais mais importantes.

3.6 Relação entre expressões faciais e nível de engajamento

Ao aplicar o modelo de reconhecimento de expressões faciais, a imagem de entrada é classificada em diferentes expressões faciais. Com base em pesquisas anteriores sobre a relação entre expressão facial e envolvimento dos alunos [10], os resultados da discriminação produzem o nível de envolvimento dos alunos. Conforme ilustrado na Figura 7, este estudo adaptou o modelo de relacionamento original, mesclando a escala de engajamento de 5 níveis em três níveis. Especificamente, os níveis de engajamento “forte” e “alto” foram combinados em um único nível de engajamento “alto”, mantendo o nível de engajamento “médio” como estava.

Além disso, o nível de engajamento “baixo” e o “sem engajamento” foram mesclados em um único nível de engajamento “baixo”.

4 Resultado

4.1 Plataforma experimental e configuração de treinamento

Nosso processamento de treinamento de modelo foi realizado em uma placa gráfica NVIDIA RTX 3090 equipada com. Esta placa gráfica possui 10.496 núcleos CUDA, 24 GB de memória GDDR6X, uma frequência de núcleo de 1.695 MHz e uma capacidade de operação de ponto flutuante de precisão única de 35,7 TFLOPs. Projetamos nosso algoritmo e o implementamos usando Python 3.8 e PyTorch 1.13.1.

Durante o treinamento do modelo, utilizamos um tamanho de lote de 64, empregando descida gradiente estocástica (SGD) como otimizador com um momento de 0,9 e queda de peso definida como 1e-4. A estratégia de ajuste da taxa de aprendizagem OneCycleLR foi aplicada, começando com uma taxa inicial de 0,0008, ajustando-se progressivamente até uma taxa máxima de 0,02 e, posteriormente, ajustando-se da taxa máxima para uma taxa mínima de 2e-6. O processo de treinamento durou um total de 700 épocas e uma semente aleatória de 1.029 foi empregada.

4.2 Conjunto de dados

O principal conjunto de dados utilizado para reconhecimento de expressões faciais e detecção de engajamento neste estudo é o FERPLUS, uma versão melhorada do FER2013. FERPLUS consiste em imagens 48 × 48 e abrange 3.111 casos de “raiva”, 248 casos de “nojo”, 819 casos de “medo”, 9.355 casos de “feliz”, 4.371 casos de “triste”, 4.462 casos de “Surpresa, ” 12.906 casos de “neutro”, 216 casos de “desprezo”, 222 casos de “desconhecido” e 177 casos de “não é um rosto”. Após observação cuidadosa, descobriu-se que “desconhecido” e “não é um rosto” são relativamente raros em comparação com outras classes, potencialmente introduzindo ruído na rede neural durante o treinamento. Portanto, modificamos o conjunto de dados para remover essas duas classes.

Para validação e expansão adicional do modelo, o conjunto de dados RAF-DB também foi utilizado. É um conjunto de dados de expressões faciais em grande escala que contém 15.339 imagens faciais. O conjunto de dados exibe variações significativas de idade, sexo, etnia, pose da cabeça, condições de iluminação, oclusões e operações de pós-processamento.

4.3 Aumento de dados

Antes do treinamento do modelo, operações de aumento de dados online foram aplicadas ao conjunto de dados. Essas operações incluem equalização do histograma, normalização, corte aleatório, rotação aleatória e apagamento aleatório de certas regiões.
Esses aumentos enriqueceram significativamente o conjunto de dados. Durante a validação foi utilizado o método TenCrop, que resultou em um aumento de dez vezes no número de casos.

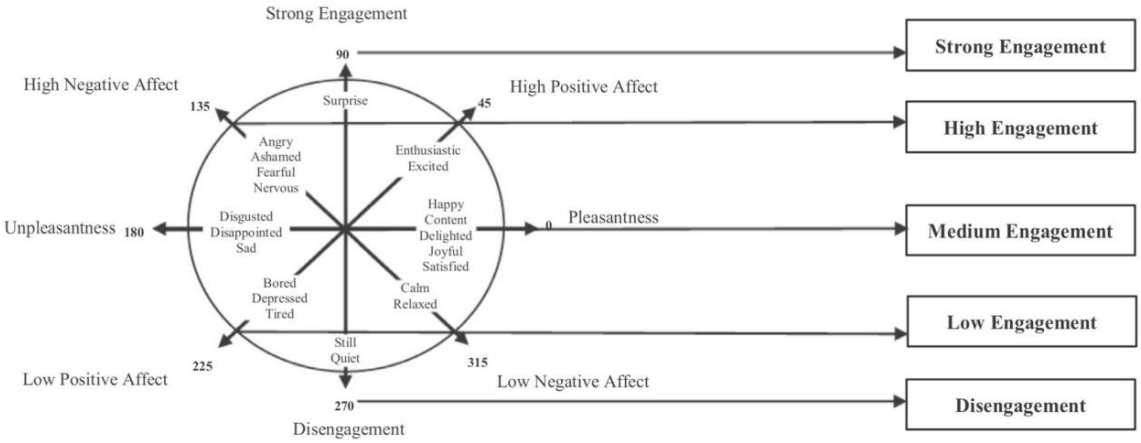

Fig. 7 A relação entre expressão facial e envolvimento

Para resolver o problema de desequilíbrio de dados no FERPLUS, o aumento de dados offline foi realizado complementando classes sub-representadas com dados parciais de expressão facial do conjunto de dados Affectnet. Após o aumento dos dados, o conjunto de dados FERPLUS aumentado abrange 4.495 casos de “raiva”, 1.809 casos de “nojo”, 2.082 casos de “medo”, 9.045 casos de “feliz”, 6.209 casos de “triste”, 5.579 casos de “surpresa, ” 11.012 casos de “neutro” e 673 casos de “desprezo”. A distribuição específica das expressões faciais é mostrada na Tabela 1.

4.4 Resultados

Avaliamos o modelo sob três aspectos: precisão, parâmetros e FLOPs. A precisão representa a proporção de previsões corretas no conjunto de teste. Uma maior precisão indica melhor desempenho. Os parâmetros referem-se ao número de parâmetros que precisam ser treinados no modelo. Mais parâmetros do modelo requerem maior espaço de memória e mais recursos para treinamento. FLOPs é uma medida da complexidade computacional do modelo, representando o número de operações de ponto flutuante executadas durante uma passagem direta. FLOPs mais altos levam a maior tempo e consumo de recursos computacionais.

No conjunto de dados FERPLUS, nosso modelo alcançou uma precisão de 87,81%, conforme mostrado na Fig. 8. Apesar do ligeiro overfitting durante o treinamento, o modelo alcançou a maior precisão no conjunto de teste. Esta observação pode ser atribuída ao design leve do modelo, que possui inerentemente algum nível de capacidade de generalização. O ligeiro overfitting observado durante o treinamento auxilia na extração eficaz dos recursos, enquanto a consistência dos dados de treinamento, validação e teste no conjunto de dados contribui para um melhor desempenho.

Depois de analisar a matriz de confusão mostrada na parte esquerda da Figura 9, fica evidente que entre as oito categorias de expressão, as expressões “raiva”, “feliz”, “surpresa” e “neutro” alcançaram alta precisão. Isso pode ser atribuído ao fato de essas categorias possuírem um maior número de casos

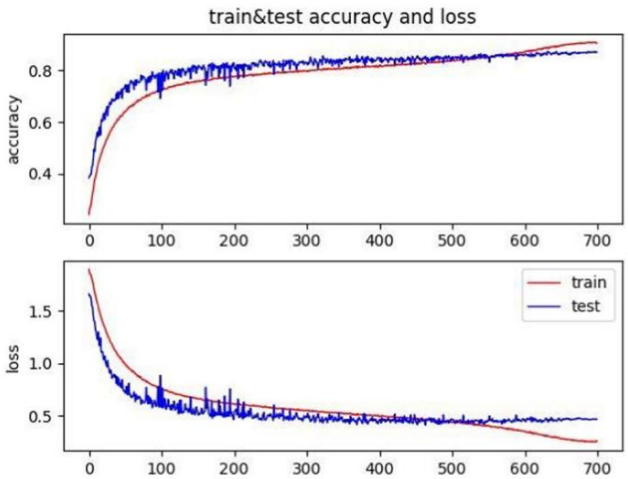

Fig. 8 A curva de treinamento do nosso modelo no conjunto de dados FERPLUS

no conjunto de dados, permitindo um treinamento mais abrangente. Em relação às emoções “feliz”, “surpresa” e “neutro”, elas compartilham algumas semelhanças com outras emoções. Em particular, a “raiva” tem características distintas, resultando em menos erros de classificação com outras categorias de emoções. Alcançou a maior precisão entre todas as categorias.

A precisão relativamente menor da categoria emoção “triste” pode ser devida à presença de microexpressões. Isto leva a erros de classificação, sendo muitas vezes categorizados como “neutros” devido à sua natureza subestimada.

Quanto às emoções de “nojo”, “medo” e “desprezo”, a sua menor precisão deve-se principalmente ao número limitado de casos de treino. Apesar de ter um número mínimo de casos, o “desprezo” demonstra características faciais distintas que lhe permitiram obter maior precisão em comparação com o “nojo”. Notavelmente, tanto as emoções de “raiva” quanto de “tristeza” envolvem o franzir da testa, que também está presente no “nojo”, tornando difícil diferenciar essas emoções.

Tabela 1 O específico distribuição de expressões faciais

Conjunto de dados	Raiva	Nojo	Medo	Feliz	triste	Surpresa	Desprezo	Neutro
Trem (FER)	3939	1767	1937	7287	5474	4768	8740	644
Teste Privado (FER)	273	18	83	893	384	396	1090	16
Teste Público (FER)	287	24	62	865	351	415	1182	13
Trem (RAF)	705	717	281	4772	1982	1290	2524	0
Teste (RAF)	162	160	74	1185	478	329	680	0

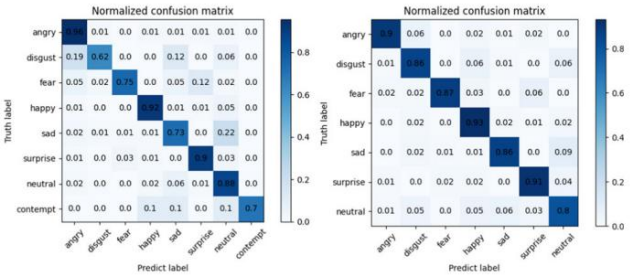

Fig. 9 A matriz de confusão do nosso modelo em FERPLUS e RAF-BD conjunto de dados

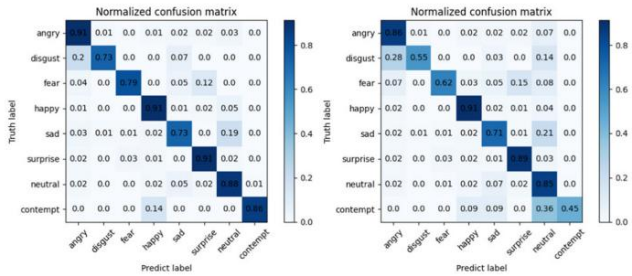

Fig. 10 A matriz de confusão de MobileNetV1 e ShufflenetV1 em Conjunto de dados FERPLUS

No conjunto de dados RAF-DB, nosso modelo atingiu uma precisão de 88,20% por meio de aprendizagem por transferência, aproveitando o treinamento resultados do conjunto de dados FERPLUS. Este desempenho é consistente com as tendências observadas no conjunto de dados FERPLUS, conforme mostrado na parte direita da Fig. 9, demonstrando a estabilidade e eficácia do modelo em diferentes conjuntos de dados.

O número de parâmetros em nosso modelo é 0,23 milhões, e a complexidade computacional no conjunto de dados FERPLUS é de 24,87 milhões de FLOPs. Isso indica que o modelo é um modelo leve.

Implantamos nosso modelo treinado para avaliar seu desempenho em uma plataforma de computação com recursos limitados. O experimento foi conduzido em Nvidia Jetson Nano, que está equipado com uma GPU baseada na arquitetura NVIDIA Maxwell de 128 núcleos, uma CPU ARM Cortex A-57 de quatro núcleos e um 4 GB de RAM dinâmica de baixo consumo de 64 bits. Durante a inferência, o o modelo proposto alcançou uma velocidade de 16,13 quadros por segundo (fps). Este resultado confirma a praticidade de implantação nosso modelo em sistemas embarcados modernos para a aplicação de detecção em tempo real dos níveis de envolvimento dos alunos em ambientes de sala de aula.

4.5 Discussão

Nesta seção, fornecemos uma análise comparativa de nossos modelo proposto, Light_Fer, juntamente com modelos leves e de grande escala comumente usados no reconhecimento de expressões faciais. As métricas de desempenho são apresentadas na Tabela 2.

Desempenho no conjunto de dados FERPLUS: Light_Fer alcançado uma precisão impressionante de 87,81% no conjunto de dados FERPLUS. Quando comparado com três modelos grandes, nomeadamente VGG11, Densenet-121 e Resnet-18, a precisão do Light_Fer os supera ligeiramente, com VGG11 sendo o mais próximo em 87,59%. Notavelmente, Light Fer exibe um valor significativamente menor número de parâmetros, com apenas 0,226 milhão, em contraste com 6,594 milhões para Densenet-121 e 49 vezes menos parâmetros que Resnet-18. Além disso, o cálculo computacional de Light_Fer complexidade, medida por FLOPs, é substancialmente menor do que a desses modelos de grande porte, enfatizando sua eficiência.

Para os cinco modelos leves — ShufflenetV1, Shuf-flenetV2, MobileNetV1, MobileNetV2 e EfficientFace — eles alcançaram precisões de 85,12%, 85,53%, 87,62%, 86,34% e 84,15%, respectivamente. Light_Fer supera estes modelos leves com precisão notavelmente maior. Até quando comparado ao modelo leve de melhor desempenho, MobilenetV1, a precisão do Light_Fer é 0,19% maior. É importante ressaltar que Light Fer mantém menos parâmetros do que qualquer um desses modelos leves, demonstrando sua eficiência.

Equilíbrio de desempenho de reconhecimento: Light_Fer exibe um desempenho de reconhecimento mais equilibrado em vários recursos faciais expressões quando comparado com MobileNetV1 e Shuf-flenetV1, como evidente nas matrizes de confusão (ver Fig. 10).
Este reconhecimento equilibrado reflete o desempenho geral da Light_Fer precisão e sua capacidade de extrair informações essenciais e de alto nível características faciais, contribuindo para um desempenho superior de reconhecimento mance.

Desempenho no conjunto de dados RAF-DB: Para substanciar ainda mais a eficácia do nosso modelo proposto, conduzimos uma avaliação abrangente, comparando-a com vários métodos de última geração usando o conjunto de dados RAF-DB, conforme resumido na Tabela 3.

Tabela 2 O desempenho do nosso modelo e de outros modelos

Modelo	Parâmetros	FLOPs (FERPLUS)	Precisão (FERPLUS) (%)	Precisão (RAF-DB) (%)
Nosso	0,23 milhões	24,87 milhões	87,81	88,20
VGG11 [32]	9,27 milhões	342,67 milhões	87,59	84,84
Densenet-121 [33]	6,95 milhões	127,63 milhões	86,25	84,29
Resnet-18 [34]	11,17 milhões	94,82 milhões	86,46	84.06
MobilenetV1 [26]	2,14 milhões	67,67 milhões	87,62	83,25
MobilenetV2 [27]	2,23 milhões	17,53 milhões	86,34	82,20
ShufflenetV1 [29]	1,91 milhões	14,88 milhões	85,12	79,66
ShufflenetV2 [30]	1,26 milhão	8,37 milhões	85,53	76,89
Cara Eficiente [23]	1,28 milhão	153,00 milhões	84,15	88,36

Tabela 3 Comparação com métodos de última geração no RAF-DB

Modelo	Parâmetros FLOP		Precisão (%)
Nosso	0,23 milhões	24,87 milhões	88,20
IPA2LT [35]	23,527 milhões	4109,48M	86,77
Perda separada [36]	11,18 milhões	1818,56M	86,38
gACNN [37]	134,29 milhões	15479,79 milhões	85.07
RAN [38]	11,19M	14548,45 milhões	86,90
LDL-ALSG [39]	23,52 milhões	4109,48M	85,53
DDA-Perder [40]	11,18 milhões	1818,56M	86,90
SCN [41]	11,18 milhões	1818,56M	87.03
SCN* [41]	11,18 milhões	1818,56M	88,14
Cara Eficiente [23]	1,28 milhão	154,18 milhões	88,36

Nos casos em que o código-fonte de modelos específicos não estava disponível, nós os reconstruímos diligentemente com base em nossa compreensão dos algoritmos subjacentes e arquiteturas. Para os demais modelos, realizamos nosso experimentos usando seu código-fonte aberto.

Os modelos incluídos na nossa análise comparativa abrangem uma ampla gama de abordagens de ponta, nomeadamente IPA2LT [35], Perda Separada [36], gACNN [37], RAN [38], LDL-ALSG [39], DDA-Lose [40], SCN [41], SCNÿ [41], e Eficiente face [23].

Desempenho de Light_Fer nos espelhos do conjunto de dados RAF-DB o seu desempenho no FERPLUS, reafirmando a sua eficácia global.

Fatores que contribuem: O sucesso do Light_Fer pode ser atribuído a múltiplos fatores. A convolução separável em profundidade e a convolução de grupo reduzem significativamente o número de parâmetros sem comprometer o desempenho. Esses as operações extraem recursos com eficiência e minimizam a sobrecarga computacional. A inclusão do gargalo invertido estrutura contribui para o aumento da complexidade computacional mas melhora a capacidade de ajuste de recursos e o desempenho do modelo. Além disso, o mecanismo de atenção permite que o modelo para se concentrar em recursos importantes, melhorando ainda mais sua precisão.

Detecção de engajamento em tempo real: Light_Fer alcança detecção automática e em tempo real do envolvimento dos alunos, integrando a relação entre expressões reconhecidas e níveis de engajamento. A eficácia do modelo foi avaliada através de extensos experimentos de testes conduzidos em diversas condições dentro de um ambiente de sala de aula real. Esses condições abrangeram vários cenários, incluindo: (i) condições normais de iluminação com imagens faciais nítidas, (ii) condições faciais imagens sob condições de pouca luz, (iii) imagens faciais não frontais imagens, (iv) pequenas imagens faciais e (v) parcialmente ocluídas imagens faciais, conforme ilustrado na Fig. 11. Os resultados obtidos demonstrar excelente desempenho de detecção em cenários caracterizado por iluminação normal, pouca iluminação e imagens faciais de tamanho pequeno. No entanto, surgem desafios no caso de imagens faciais não frontais, impactando a precisão da detecção. Em cenas envolvendo oclusão facial, o modelo ainda pode realizar certas detecções sob oclusão em pequena escala. Ainda, não consegue detectar oclusão facial em grande escala. Abordando o problemas associados a imagens e cenas faciais não frontais caracterizando a oclusão facial se destaca como um caminho crucial para exploração e melhoria futuras.

Em resumo, o sucesso do nosso modelo Light_Fer pode ser atribuído à sua arquitetura leve, recurso eficiente recursos de extração e detecção de engajamento em tempo real desempenho. Esses recursos tornam Light_Fer um valioso ferramenta para diversas aplicações, especialmente no contexto de monitoramento de sala de aula em tempo real e envolvimento dos alunos avaliação.

5 Conclusão e trabalhos futuros

Seja em salas de aula tradicionais ou on-line, monitorar o envolvimento dos alunos é essencial. Neste artigo, propomos Light_Fer, um modelo leve de reconhecimento de expressão facial, para detecção automatizada do envolvimento dos alunos na sala de aula.

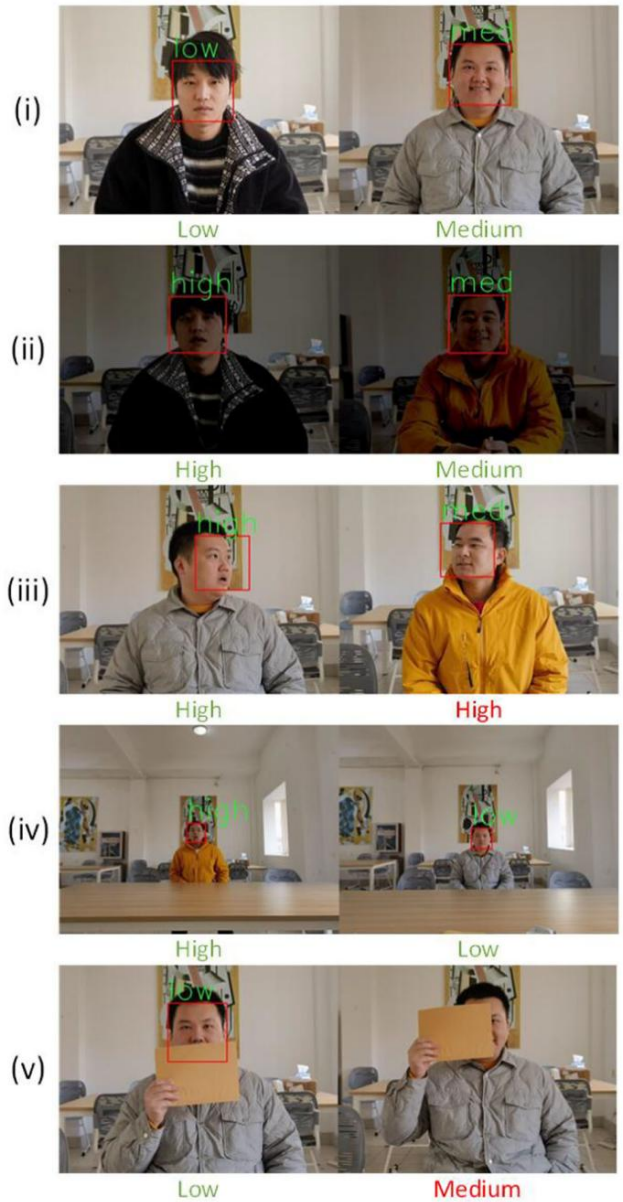

Fig. 11 Resultados de previsão dos modelos propostos em diversas condições dentro de uma sala de aula real. O texto abaixo da imagem mostra o nível real de atenção. O texto verde corresponde a previsões corretas e o texto vermelho corresponde a previsões incorretas (figura colorida online)

Light_Fer demonstra precisão excepcional nos conjuntos de dados FERPLUS e RAF-DB, superando outros modelos em termos de desempenho balanceado, incluindo precisão, contagem de parâmetros e FLOPs. Ao integrar as expressões faciais com o envolvimento dos alunos, nosso método oferece uma abordagem eficiente e adaptável para monitorar a participação dos alunos em sala de aula em diversos ambientes de aprendizagem.

Em nosso trabalho de pesquisa futuro, planejamos integrar as expressões faciais e a análise da postura corporal para detectar de forma abrangente o envolvimento dos alunos na sala de aula a partir de múltiplas perspectivas. Além disso, continuaremos nossa exploração de algoritmos leves de detecção e classificação de objetos para avançar ainda mais os objetivos de soluções leves e implantáveis.

Contribuições dos autores Zhao conduziu a proposta do modelo, a verificação experimental e a redação do rascunho inicial. Li forneceu aconselhamento técnico significativo e apoio na conceituação e implementação, e contribuiu para a redação do artigo. Yang e Ma forneceram orientações de pesquisa e gerenciamento de projetos.

Financiamento Este trabalho foi apoiado em parte pela Fundação Nacional de Ciências Naturais da China sob concessão (nº 62071161).

Disponibilidade de dados e materiais Os conjuntos de dados gerados e/ou analisados durante o estudo atual estão disponíveis no autor correspondente mediante solicitação razoável.

Declarações

Conflito de interesses Os autores não têm interesses conflitantes que possam influenciar os resultados e/ou discussão relatados neste artigo.

Aprovação ética Não aplicável.

Referências

1. Coates, H.: Envolvendo Alunos para o Sucesso - Pesquisa Australásia de Envolvimento Estudantil de 2008. Conselho Australiano de Pesquisa Educacional, Victoria, Austrália (2009)
2. Gunuc, S.: As relações entre o envolvimento dos alunos e o seu desempenho académico. Internacional J. Novas Tendências Educ. Seu Implícito. 5(4), 216–231 (2014)
3. Casuso-Holgado, MJ, Cuesta-Vargas, AI, Moreno-Morales, N., Labajos-Manzanares, MT, Barón-López, FJ, Vega-Cuesta, M.: A associação entre engajamento acadêmico e desempenho em ciências da saúde estudantes. BMC Med. Educ. 13(1), 1–7 (2013)
4. Trowler, V.: Revisão da literatura sobre envolvimento estudantil. Alto. Educ. Acad. 11(1), 1–15 (2010)
5. Bower, GH: Humor e memória. Sou. Psicol. 36(2), 129 (1981)
6. Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., Movellan, JR: As faces do envolvimento: reconhecimento automático do envolvimento do aluno a partir de expressões faciais. IEEE Trans. Afeto. Computação. 5(1), 86–98 (2014)
7. Mehrabian, A.: Comunicação sem palavras. Univ. Leste de Londres. 24(4), 1084–5 (1968)
8. Ekman, P., Friesen,W.: Sistema de codificação de ação facial: Uma técnica para medir o movimento facial. Em Psicologia Ambiental e Comportamento Não-verbal. Consulting Psychologists Press: Palo Alto, CA, EUA (1978)
9. Sathik, M., Jonathan, SG: Efeito das expressões faciais no reconhecimento da compreensão do aluno em ambientes educacionais virtuais. Springerplus 2, 1–9 (2013)
10. Altuwairqi, K., Jarraya, SK, Allinjawi, A., Hammami, M.: Um novo modelo afetivo baseado na emoção para detectar o envolvimento do aluno. J. Universidade Rei Saud. Computação. Inf. Ciência. 33(1), 99–109 (2021)
11. Fish, J., Brimson, J., Lynch, S.:Intervenções de mindfulness fornecidas por tecnologia sem envolvimento de facilitador: que pesquisas existem e quais são os resultados clínicos? Atenção Plena 7, 1011–1023 (2016)

12. Hew, KF: Promovendo o envolvimento em cursos on-line: quais estratégias podemos aprender com três MOOCS altamente avaliados. Ir. J.Educ. Tecnologia. 47(2), 320–341 (2016)
13. Aneja, D., Colburn, A., Faigin, G., Shapiro, L., Mones, B .: Modelagem de expressões de personagens estilizados por meio de aprendizagem profunda. In: ACCV (2016)
14. Mollahosseini, A., Chan, D., Mahoor, MH: Aprofundando o reconhecimento de expressões faciais usando redes neurais profundas. IEEE (2016)
15. Grafsgaard, J., Wiggins, JB, Boyer, KE, Wiebe, EN, Lester, J.: Reconhecendo automaticamente a expressão facial: prevendo envolvimento e frustração. In: Mineração de Dados Educacionais 2013 (2013)
16. Dubbaka, A., Gopalan, A.: Detectando o envolvimento do aluno em MOOCS usando reconhecimento automático de expressão facial. In: Conferência Global de Educação em Engenharia IEEE 2020 (EDUCON) (2020)
17. Shen, J., Yang, H., Li, J., Cheng, Z.: Avaliando o envolvimento na aprendizagem com base no reconhecimento de expressões faciais no cenário MOOC. Sistema Multimídia **28,** 469–478 (2022)
18. Liao, J., Liang, Y., Pan, J.: Rede espaçotemporal facial profunda para previsão de engajamento na aprendizagem online. Apl. Intel. **51,** 6609–6621 (2021)
19. Gupta, S., Kumar, P., Tekchandani, RK: Sistema de detecção de envolvimento do aluno em tempo real baseado em reconhecimento de emoções faciais em contexto de aprendizagem online usando modelos de aprendizagem profunda. Multimed. Ferramentas Appl. 82(8), 11365–11394 (2023)
20. Hewitt, C., Gunes, H.: Análise de afeto facial baseada em CNN em dispositivos móveis. Pré-impressão do arXiv arXiv:1807.08775 (2018)
21. Barros, P., Churamani, N., Sciutti, A .: O facechannel: uma rede neural profunda leve para reconhecimento de expressões faciais. In: 15ª Conferência Internacional IEEE de 2020 sobre reconhecimento automático de rosto e gestos (FG 2020), pp. IEEE 22. Ferro-Pérez, R., Mitre-Hernandez, H.: Resmonet: uma rede residual baseada em dispositivos móveis para reconhecimento de emoções faciais em sistemas com recursos limitados (2020)
23. Zhao, Z., Liu, Q., Zhou, F.: Rede robusta e leve de reconhecimento de expressão facial com treinamento em distribuição de rótulos. In: Anais da Conferência AAAI sobre Inteligência Artificial, vol. 35, pp.
24. Chollet, F .: Xception: aprendizagem profunda com convoluções separáveis em profundidade. In: Anais da Conferência IEEE sobre Visão Computacional e Reconhecimento de Padrões, pp.
25. Iandola, FN, Han, S., Moskewicz, MW, Ashraf, K., Dally, WJ, Keutzer, K .: Squeezenet: precisão de nível Alexnet com 50x menos parâmetros e tamanho de modelo <0,5 mb. pré-impressão arXiv arXiv:1602.07360 (2016)
26. Howard, AG, Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: redes neurais convolucionais eficientes para aplicações de visão móvel. pré-impressão arXiv arXiv:1704.04861 (2017)
27. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C .: Mobilenetv2: resíduos invertidos e gargalos lineares. In: Procedimentos da Conferência IEEE sobre Visão Computacional e Reconhecimento de Padrões, pp.
28. Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V.: Procurando por mobilenetv3. In: Anais da Conferência Internacional IEEE/CVF sobre Visão Computacional, pp.
29. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: uma rede neural convolucional extremamente eficiente para dispositivos móveis. In: Procedimentos da Conferência IEEE sobre Visão Computacional e Reconhecimento de Padrões, pp. 6848–6856 (2018)
30. Ma, N., Zhang, X., Zheng, H.-T., Sun, J.: Shufflenet v2: diretrizes práticas para design eficiente de arquitetura CNN. In: Anais da Conferência Europeia sobre Visão Computacional (ECCV), pp.
31. Woo, S., Park, J., Lee, J.-Y., Kweon, IS: CBAM: módulo de atenção de bloco convolucional. In: Anais da Conferência Europeia sobre Visão Computacional (ECCV), pp. 3–19 (2018)
32. Simonyan, K., Zisserman, A .: Redes convolucionais muito profundas para reconhecimento de imagens em grande escala. pré-impressão arXiv arXiv:1409.1556 (2014)
33. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, KQ: Redes convolucionais densamente conectadas. In: Anais da Conferência IEEE sobre Visão Computacional e Reconhecimento de Padrões, pp.
34. He, K., Zhang, X., Ren, S., Sun, J .: Aprendizagem residual profunda para reconhecimento de imagem. In: Anais da Conferência IEEE sobre Visão Computacional e Reconhecimento de Padrões, pp.
35. Zeng, J., Shan, S., Chen, X .: Reconhecimento de expressão facial com conjuntos de dados anotados inconsistentemente. In: Anais da Conferência Europeia sobre Visão Computacional (ECCV), pp.
36. Li, Y., Lu, Y., Li, J., Lu, G .: Perda separada para reconhecimento de expressões faciais básicas e compostas na natureza. In: Conferência Asiática sobre Aprendizado de Máquina, pp. 897–911 (2019). PMLR 37.
Li, Y., Zeng, J., Shan, S., Chen, X .: Reconhecimento de expressão facial com reconhecimento de oclusão usando CNN com mecanismo de atenção. IEEE Trans. Processo de imagem. 28(5), 2439–2450 (2018)
38. Wang, K., Peng, X., Yang, J., Meng, D., Qiao, Y .: Redes de atenção regional para reconhecimento robusto de expressão facial de pose e oclusão. IEEE Trans. Processo de imagem. **29,** 4057–4069 (2020)
39. Chen, S., Wang, J., Chen, Y., Shi, Z., Geng, X., Rui, Y .: Aprendizagem de distribuição de rótulos em gráficos de espaço de rótulos auxiliares para reconhecimento de expressões faciais. In: Anais da Conferência IEEE/CVF sobre Visão Computacional e Reconhecimento de Padrões, pp.
40. Farzaneh, AH, Qi, X .: Perda agnóstica de distribuição discriminante para reconhecimento de expressão facial na natureza. In: Anais da Conferência IEEE/CVF sobre Workshops de Visão Computacional e Reconhecimento de Padrões, pp.
41. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y .: Suprimindo incertezas para reconhecimento de expressões faciais em larga escala. In: Procedimentos da Conferência IEEE/CVF sobre Visão Computacional e Reconhecimento de Padrões, pp.