Redes Neurais Recorrentes para Reconhecimento de Emoções em Vídeo

Samira Ebrahimi Kahou École Polytechnique de Montréal, Canadá samira.ebrahimi kahou@polymtl.ca

Vincent Michalski Kishore Konda Université de Montréal, Goethe-Universität Frankfurt, Montreal, Canadá vincent.michalski@umontreal.ca konda.kishorereddy@gmail.com

Roland Memisevic Université de Montréal, Montreal, Canadá roland.memisevic@umontreal.ca

Christopher Pal École Polytechnique de Montréal, Canadá christopher.pal@polymtl.ca

RESUMO

Abordagens baseadas em aprendizado profundo para análise facial e análise de vídeo demonstraram recentemente alto desempenho em uma variedade de tarefas importantes, como reconhecimento facial, reconhecimento de emoções e reconhecimento de atividades. No caso do vídeo, as informações geralmente devem ser agregadas em uma sequência de quadros de tamanho variável para produzir um resultado de classificação. Trabalhos anteriores usando redes neurais convolucionais (CNNs) para reconhecimento de emoções em vídeo basearam-se em médias temporais e operações de agrupamento reminiscentes de abordagens amplamente usadas para a agregação espacial de informações. Redes neurais recorrentes (RNNs) têm visto uma explosão de interesse recente, pois fornecem desempenho de ponta em uma variedade de tarefas de análise de sequência. As RNNs fornecem uma estrutura atraente para a propagação de informações em uma sequência usando uma representação de camada oculta de valor contínuo. Neste trabalho apresentamos um sistema completo para o Desafio Emo tion Recognition in the Wild (EmotiW) 2015. Focamos nossa apresentação e análise experimental em uma arquitetura híbrida CNN-RNN para análise de expressão facial que pode superar uma abordagem CNN aplicada anteriormente usando média temporal para agregação.

Categorias e Descritores de Assunto I.5 [Reconhecimento de Padrões]: Modelos, Aplicações

Palavras -chave

reconhecimento de emoções; aprendizagem profunda; aprendizagem multimodal; combinação de modelos; redes neurais recorrentes

ICMI 2015, 9 a 13 de novembro de 2015, Seattle, WA, EUA.

ACM 978-1-4503-3912-4/15/11 ...$15,00.

DOI: http://dx.doi.org/10.1145/2818346.2830596 .

1. INTRODUÇÃO

A análise de emoções humanas é uma tarefa desafiadora de aprendizado de máquina com uma ampla gama de aplicações em interação humano-computador, e-learning, assistência médica, publicidade e jogos. A análise da emoção é particularmente desafiadora, pois múltiplas modalidades, tanto visuais quanto auditivas, desempenham um papel importante em sua compreensão. Dada uma sequência de vídeo com um sujeito humano, algumas das dicas importantes que ajudam a entender a emoção do usuário são expressões faciais, movimentos e atividades. Em alguns casos, a fala ou o contexto da cena de alto nível também podem ser úteis para inferir emoções. Na maioria das vezes, há uma sobreposição considerável entre as classes de emoção, tornando-se uma tarefa de classificação desafiadora. Neste artigo, apresentamos uma abordagem baseada em aprendizado profundo para modelar diferentes modalidades de entrada e combiná-las para inferir rótulos de emoção de uma determinada sequência de vídeo.

O desafio Reconhecimento de emoções na natureza (EmotiW 2015) [9] é uma extensão de um desafio semelhante realizado em 2014 [8]. A tarefa é prever um dos sete rótulos de emoção: raiva, nojo, medo, alegria, tristeza, surpresa e neutro. O conjunto de dados usado no desafio é o conjunto de dados Acted Facial Expressions in the Wild (AFEW) 5.0, que contém pequenos videoclipes extraídos de filmes de Hollywood. Os videoclipes apresentam emoções com alto grau de variação, por exemplo, identidade do ator, idade, pose e condições de iluminação. O conjunto de dados contém 723 vídeos para treinamento, 383 para validação e 539 clipes de teste.

As abordagens tradicionais para o reconhecimento de emoções foram baseadas em recursos de engenharia manual [17, 28]. Com a disponibilidade de grandes conjuntos de dados, o aprendizado profundo surgiu como uma abordagem geral para o aprendizado de máquina, produzindo resultados de ponta em muitas tarefas de visão computacional e processamento de linguagem natural [22, 19]. O princípio básico do aprendizado profundo é aprender representações hierárquicas de dados de entrada, de modo que as representações aprendidas melhorem o desempenho da classif

A principal contribuição deste trabalho é modelar a evolução espaço-temporal das expressões faciais de uma pessoa em um vídeo usando uma Rede Neural Recorrente (RNN) combinada com uma Rede Neural Convolucional (CNN) em uma arquitetura subjacente CNN-RNN. Além disso, também empregamos um pipeline de reconhecimento de atividade baseado em Autoencoder para modelar a atividade do usuário e uma abordagem simples baseada em Support Vector Machine (SVM) sobre energia e recursos espectrais para áudio. Também apresentamos uma rede neural baseada em

técnica de fusão de nível de recurso para combinar diferentes modalidades para a previsão de emoção final para um videoclipe curto.

Trabalhos anteriores [18, 25] alcançaram resultados de ponta no desafio de reconhecimento de emoções usando técnicas de aprendizado profundo, que inclui nosso trabalho que venceu o desafio EmotiW de 2013. Em contraste com [18, 16], que usa um método de agregação baseado em média para recursos visuais em vídeo, aqui empregamos um RNN para modelar a evolução temporal de recursos faciais em vídeo. Também exploramos a fusão de nível de recurso de nossos modelos específicos da modalidade e mostramos que isso aumenta o desempenho.

O restante deste artigo está organizado da seguinte forma. Nas Seções 2, 3 e 4 descrevemos cada um dos modelos utilizados para diferentes modalidades seguidas pela Seção 5, que fornece detalhes sobre os métodos de fusão que aplicamos. A Seção 6 apresenta nossos resultados experimentais e fornece uma lista de nossas submissões ao desafio. Finalmente, na Seção 7 tiramos algumas conclusões de nossos experimentos.

2. EVOLUÇÃO ESPAÇO-TEMPORAL DAS EXPRESSÕES FACIAIS

Modelar a evolução espaço-temporal da informação visual desempenha um papel importante na compreensão do comportamento de objetos e usuários em vídeo. O reconhecimento de emoções é uma das tarefas que envolvem a modelagem do comportamento de um usuário.
Neste trabalho, usamos uma abordagem de duas etapas para modelar a emoção como a evolução espaço-temporal da estrutura da imagem. Na primeira etapa, uma CNN é treinada para classificar imagens estáticas contendo emoções. Na segunda etapa, treinamos uma RNN na representação da camada superior da CNN inferida de quadros individuais para prever uma única emoção para todo o vídeo. RNNs passaram por um ressurgimento de interesse devido em parte ao seu desempenho impressionante em reconhecimento de escrita e fala [14, 13]. Grande parte desse interesse foi impulsionado pela estabilidade do aprendizado alcançado pelo uso das chamadas unidades de memória de longo prazo (LSTM) [15]. RNNs também provaram ser métodos poderosos para outros tipos de dados sequenciais, incluindo vídeo [1, 10] e processamento de linguagem natural [2, 32]. Como tal, usamos uma estrutura RNN para aprender um modelo para representação e classificação em nível de vídeo. A representação da camada superior da CNN fornece informações estruturais de um determinado quadro e a RNN modela a evolução espaço-temporal da estrutura ao longo do tempo.

Ao contrário de outros trabalhos envolvendo técnicas de vídeo e RNN, como [1, 10], não usamos LSTMs. Aqui usamos IRNNs [24] que são compostos de unidades lineares retificadas (ReLUs) e empregam uma estratégia de inicialização especial baseada em variações escalonadas da matriz identidade. Esses elementos de IRNNs visam fornecer um mecanismo muito mais simples para lidar com o problema de gradiente de desaparecimento e explosão em comparação com a estrutura LSTM mais complexa. Trabalhos recentes compararam IRNNs com LSTMs e descobriram que IRNNs são capazes de produzir resultados comparáveis em algumas tarefas, incluindo problemas que envolvem dependências de longo prazo [24].

Fornecemos uma explicação detalhada da estrutura da CNN na Seção 2.1 e da RNN na Seção 2.2. Para comparar com a abordagem não sequencial apresentada em [16], também agregamos características CNN a um vetor de características de comprimento fixo e treinamos um SVM. Isso é descrito na Seção 2.3.

2.1 Extração de recursos de quadro usando uma CNN

O conjunto de dados da competição possui um rótulo de emoção por vídeo que não corresponde a todos os quadros. Isso introduz muito ruído se os rótulos de vídeo forem usados como alvos para treinar uma CNN em quadros individuais. Nossos recursos visuais são, portanto, fornecidos por uma CNN treinada em uma combinação de dois conjuntos de dados de emoção adicionais de imagens estáticas. Além disso, o uso de dados adicionais abrange uma variedade maior de idade e identidade, em contraste com os dados de desafio em que o mesmo ator/atriz pode aparecer em vários clipes.

2.1.1 Conjuntos de dados

Os conjuntos de dados adicionais usados no treinamento da CNN consistem em dois grandes conjuntos de dados de emoção, ou seja, o Toronto Face Database (TFD) [31] com 4.178 imagens e o conjunto de dados Fa cial Expression Recognition (FER2013) [6] contendo 35.887 imagens, ambos com sete expressões básicas: raiva, nojo, medo, alegria, tristeza, surpresa e neutralidade.

2.1.2 Pré-processamento

Para levar em consideração as condições de iluminação variadas (em particular, entre conjuntos de dados), aplicamos a equalização de histograma. Usamos os rostos alinhados fornecidos pelos organizadores para extrair recursos da CNN. O alinhamento envolve uma abordagem combinada de detecção e rastreamento de pontos-chave faciais, explicada em [7]. Vamos nos referir a este conjunto de dados como AFEW-faces. Diferentes técnicas de detecção e/ou alinhamento de face foram usadas para FER2013, TFD e AFEW-faces. Para poder aproveitar os conjuntos de dados adicionais, realinhamos todos os conjuntos de dados para FER2013 usando o seguinte procedimento:

1. Detectamos cinco pontos-chave faciais para todas as imagens no conjunto de treinamento FER2013, TFD e AFEW-faces usando o método de cascata de rede neural convolucional em [30].

2. Para cada conjunto de dados, calculamos a forma média por av riando as coordenadas dos pontos-chave.

3. Os conjuntos de dados foram mapeados para FER2013 usando uma transformação de similaridade entre formas médias. Ao calcular uma transformação por conjunto de dados, deixamos os olhos, nariz e boca aproximadamente no mesmo local, mantendo uma pequena quantidade de variação. Adicionamos uma borda com ruído para TFD e AFEW-faces, pois as faces foram cortadas com mais precisão em comparação com FER2013.

4. A validação de ALGUMAS faces e os conjuntos de teste foram mapeados usando a transformação inferida no conjunto de treinamento.

Também realizamos a normalização do conjunto de dados com a imagem de média e desvio padrão do FER2013 e TFD combinados (FER+TFD).

2.1.3 Arquitetura CNN

Treinamos várias arquiteturas CNN em FER+TFD sem usar nenhum dado de desafio para cálculos de gradiente. Para a parada antecipada, tentamos deixar de fora 1.000 amostras de FER+TFD e os dados de desafio. Observamos que a RNN produz um desempenho ligeiramente melhor quando a parada antecipada da CNN foi feita nos dados de desafio, pois isso evita o ajuste excessivo para FER+TFD. Portanto, para nossa melhor estrutura de CNN, treinamos em todos os FER+TFD e executamos um ping de parada antecipada em AFEW-faces train+validation.

Exploramos três estruturas principais da CNN:

- uma estrutura muito profunda com tamanho pequeno de filtro 3x3 [26, 29],
- uma CNN de três camadas com filtros 5x5 [21, 22] e
- uma CNN semelhante de três camadas com tamanho de filtro 9x9.

A CNN é treinada principalmente para extração de características e usamos apenas o conjunto de dados adicional para a fase de treinamento. Portanto, buscamos uma estrutura que melhor generalize para outros conjuntos de dados. Estruturas profundas são conhecidas por aprender representações que melhor generalizam para outros conjuntos de dados [29]. No entanto, observamos que a estrutura muito profunda rapidamente se sobreajustou para FER+TFD e generalizou mal para o conjunto de dados de desafio. Isso pode ser devido à quantidade relativamente pequena de dados rotulados disponíveis para a tarefa de reconhecimento de emoções aqui. Por esta razão, tentamos uma rede mais rasa com três camadas que parece ter resolvido moderadamente o problema de sobreajuste. Finalmente, aumentamos o tamanho do filtro de 5 para 9 e reduzimos o número de filtros de 64-64-128 para 32-32-64. Para todos os experimentos, usamos aumento de dados (inversão horizontal com probabilidade de 0,5 e corte aleatório), bem como abandono (com taxa de 0,25).

2.2 Sequências de aprendizado usando uma RNN

Usamos um RNN para agregar recursos de quadro pelos seguintes motivos:

- A ordem temporal dos quadros é respeitada em contraste para abordagens bag-of-features.
- Uma RNN tem a capacidade de aprender a detectar um evento, como a presença de uma determinada expressão, independentemente do momento em que ocorre em uma sequência.
- As RNNs lidam naturalmente com um número variável de quadros.

RNNs são um tipo de rede neural que transforma uma sequência de entradas em uma sequência de saídas. A cada intervalo de tempo t, um estado oculto ht é calculado com base no estado oculto no tempo t ÿ 1 e na entrada xt no tempo t

$$ht = ÿ(Winxt + Wrechtÿ1), \quad (1)$$

onde Win é a matriz de peso de entrada, Wrec é a matriz recorrente e ÿ é a função de ativação oculta. Cada vez step também calcula as saídas, com base no estado oculto atual:

$$yt = f(Woutht), \quad (2)$$

onde Wout é a matriz de peso de saída e f é a função de ativação de saída. Um exemplo de uma RNN em que apenas o último passo de tempo produz uma saída é mostrado na Figura 1.

Usamos o IRNN, que como discutido acima é um RNN simples com unidades ocultas lineares retificadas (ReLUs) e com uma matriz recorrente, que é inicializada com variações escalonadas da matriz identidade [24]. O truque de inicialização de identidade garante um bom fluxo de gradiente no início do treinamento e nos permite treiná-lo em sequências relativamente longas.

Treinamos o IRNN para classificar um vídeo, alimentando os recursos para cada quadro da CNN sequencialmente para o trabalho de rede e usando a saída softmax do último intervalo de tempo como previsão de classe. Usamos Stochastic Gradient Descent (SGD) com uma taxa de aprendizado de 0,005, recorte de gradiente em 1,0 e um batchsize de 64 sequências. Experimentamos usar diferentes camadas da CNN como recursos de entrada e escolhemos a saída da segunda camada convolucional após o pooling máximo, pois isso teve melhor desempenho em dados de validação.

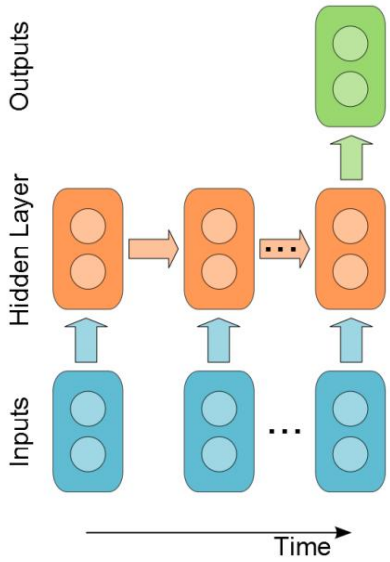


Figura 1: Estrutura da nossa rede neural recorrente.

2.3 Recursos agregados da CNN Como uma forma alternativa de agregar as representações estruturais em nível de quadro da CNN, empregamos agrupamento k-average juntamente com um SVM para classificação como em [16]. Nesta abordagem, os recursos CNN por quadro são calculados em média em compartimentos para gerar um vetor de comprimento fixo de tamanho k como representação de vídeo. Heuristicamente, selecionamos k = 15 e usamos as saídas pré-softmax da CNN como recursos por quadro.

Para vídeos com um número de quadros menor que k, os quadros são repetidos localmente até que o comprimento da sequência seja k.

As representações vetoriais de vídeos junto com os rótulos de emoção correspondentes são usados para treinar um SVM de kernel RBF. Os hiperparâmetros do SVM são definidos por meio da pesquisa de grade. Conforme mostrado na Tabela 1, a RNN atinge uma precisão de validação de 39,6%, significativamente maior do que a CNN agregada. A média simples das probabilidades por quadro produziu uma precisão de validação de apenas 23,7%.

3. ÁUDIO

Dado que o foco principal deste trabalho é o reconhecimento de emoções baseado na visão, simplesmente usamos os recursos de áudio empregados em [7] para o canal de áudio dos videoclipes.
Estes são baseados na abordagem de [27]. Ele usa 1582 recursos extraídos com o kit de ferramentas de código aberto Emotion and Affect Recognition (openEAR) [12] que usa openSMILE [11] como back-end.

O kit de ferramentas encapsula vários descritores de recursos de áudio de baixo nível (LLDs) e diferentes funções para aplicar neles. O conjunto de recursos consiste em 34 LLDs relacionados a energia e espectrais e 21 funcionais, 4 funcionais LLD × 19 relacionados a voz, 34 coeficientes delta de energia e LLD espectral × 21 funcionais, 4 coeficientes delta dos funcionais LLD × 19 relacionados a voz e 2 funcionais /características duracionais surdas.

Neste trabalho, usamos a redução de dimensionalidade baseada em Análise de Componentes Principais (PCA) como pré-processamento nos recursos de entrada de 1582 dimensões e um SVM de kernel RBF para classificação. Os hiperparâmetros para o SVM são definidos por meio de pesquisa de grade.

4. ATIVIDADE

Transformações espaço-temporais de características de imagens locais, ou atividades, podem ser uma pista importante para o reconhecimento de emoções. Um subconjunto de emoções pode ser representado como mudanças nas expressões faciais e, em alguns casos, na atividade de todo o corpo da pessoa. Outras abordagens, baseadas na visão, descritas neste trabalho tratam principalmente da análise da emoção em uma determinada sequência de vídeo com base em características estáticas da imagem e diferentes formas de agregá-las ao longo do tempo. O pipeline de análise de atividades é a única abordagem que se baseia no aprendizado de transformações espaço-temporais locais a partir do vídeo.

Nossa abordagem para análise de atividade é baseada no pipeline de reconhecimento de ação de [20, 23], que também foi usado para reconhecimento de emoções anteriormente em [16]. O pipeline consiste principalmente em três módulos diferentes, a saber, extração de recursos de movimento local, quantização de k-means e classificação baseada em SVM. Um Synchrony Autoencoder (SAE) [20] treinado em blocos de vídeo 3D recortados de tamanho 16 × 16 × 10 (espaço × espaço × tempo) é usado para extração de recursos de movimento local. A Figura 2 mostra os filtros aprendidos pelo modelo no conjunto de treinamento AFEW 5.0.

5. FUSÃO

Em muitas tarefas discriminativas, a fusão de previsões ou representações de modelos treinados usando diferentes modalidades de entrada produz uma melhoria significativa. Usamos dois tipos de abordagens de fusão para combinar os modelos específicos da modalidade descritos nas seções anteriores, fusão de nível de recurso e fusão de nível de decisão.

5.1 Nível do recurso

Nesta abordagem, uma combinação de representações de nível intermediário dos modelos treinados é usada como entrada para treinar um modelo adicional na tarefa de classificação. Para a fusão em nível de recursos, aplicamos uma variante da rede de fusão de recursos regularizada de [33]. A rede de fusão de recursos é um Multilayer Perceptron (MLP) com camadas ocultas separadas para cada modalidade, conforme mostrado na Figura 3. As saídas dessas camadas são concatenadas e alimentadas para outra camada oculta que é seguida por uma camada softmax cujo número de unidades é igual ao número de classes de emoção. A primeira camada da rede de fusão, consistindo em camadas específicas da modalidade, é regularizada para encorajar uma representação comum, compartilhando subconjuntos similares de unidades ocultas entre as modalidades, mantendo ainda as características discriminativas presentes em algumas modalidades.

A rede é treinada com SGD usando uma taxa de aprendizado de 0,1 e corte de gradiente usando limite de corte 10.
A função objetivo é a entropia cruzada categórica entre o rótulo de destino e a previsão. Como entrada para a rede de fusão, usamos os recursos agregados por quadro da CNN, os recursos de áudio clareados pelo PCA e as ativações da camada oculta do último passo de tempo da RNN. Excluímos o modelo de reconhecimento de atividade do mix, pois ele tende a se ajustar demais ao seu conjunto de treinamento. Também exploramos a adição de dropout às camadas ocultas para evitar o ajuste excessivo no pequeno desafio

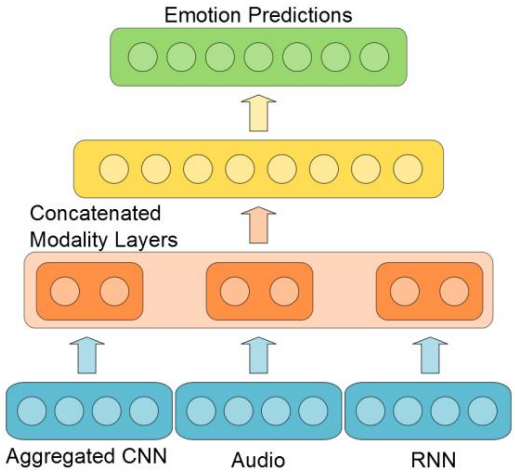


Figura 3: Estrutura da rede de fusão de recursos.

conjunto de dados. O número de camadas ocultas e seus tamanhos são selecionados usando o conjunto de validação. Nossa melhor arquitetura tem 100, 10 e 50 unidades nas camadas ocultas agregadas de CNN, áudio e específicas de RNN, respectivamente. A camada oculta comum tem 70 unidades. O espaço de busca para determinar o tamanho ideal das camadas específicas da modalidade foi selecionado considerando os tamanhos dos recursos de entrada e os desempenhos dos modelos individuais no conjunto de validação AFEW 5.0 durante o treinamento no conjunto de trens. Mais detalhes são fornecidos na Seção 6.

5.2 Nível de Decisão

Para fusão em nível de decisão, ou seja, a combinação de classificadores, usamos uma soma ponderada das probabilidades de classe estimadas pelos classificadores específicos da modalidade e pela rede de fusão. O classificador combinado tem um peso por modalidade por classe e a pontuação resultante para cada classe é a soma ponderada de todas as probabilidades para a respectiva classe.
Os pesos da combinação são determinados por busca aleatória [4], que também foi usada para combinação de modelos na abordagem vencedora do desafio EmotiW de 2013 [18].

Os pesos são amostrados uniformemente de [0,0, 1,0] seguidos por redimensionamento por classe, de modo que somam 1. Em seguida, os melhores pesos amostrados são escolhidos com base no desempenho da validação. Observe que, salvo indicação em contrário, sempre usamos a partição do conjunto de dados para a pesquisa aleatória que não foi usada para treinamento do modelo, ou seja, para modelos treinados no conjunto de treinamento, realizamos a pesquisa aleatória no conjunto de validação e vice-versa. Após uma busca aleatória inicial com 100.000 iterações, realizamos uma busca aleatória local em torno do melhor conjunto de pesos encontrado até o momento. Essa busca aleatória local consiste em ponderar amostras de um gaussiano com média definida para o melhor conjunto atual de ponderações e desvio padrão ÿ de 0,5. O melhor ˜w atual é atualizado assim que um novo melhor é encontrado. Após cada 100.000 iterações, o ÿ é diminuído por um fator de 0,9 e a busca local é interrompida quando ÿ é menor que 0,0001. Também realizamos busca local uniforme a partir de [˜w ÿ r, w˜ + r], onde ˜w é o melhor conjunto atual de pesos e r é o intervalo no qual pesquisar, porém obteve aproximadamente o mesmo desempenho. Tentamos explicitamente todas as combinações de subconjuntos de modalidades e fusão.

Consistentemente, descobrimos que a fusão do nível de decisão se beneficiou da inclusão de todos os modelos.

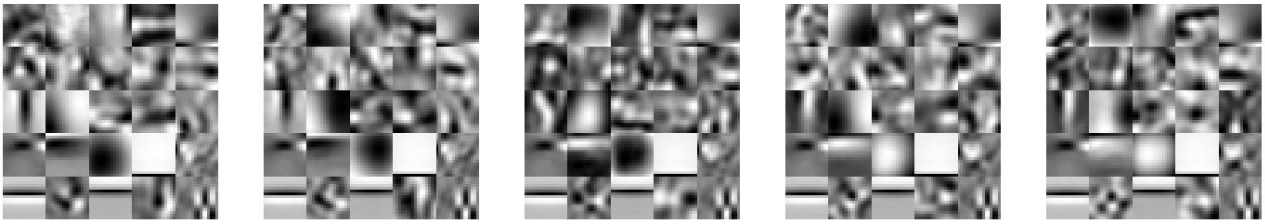


Figura 2: Subconjunto de filtros aprendidos pelo modelo SAE no conjunto de treinamento AFEW5. Da esquerda para a direita: Quadros 1,3,5,7 e 9.

Tabela 1: Precisões de Treinamento e Validação para Todas as Modalidades (Treinamento na Partição Train)

Modelo	Validação do Treinamento	
Atividade	0,983	0,266
Áudio	0,418	0,332
Agregado CNN 0,505 RNN	0,848	0,350
		0,396

6. RESULTADOS

Nesta seção, descrevemos nossas inscrições para o desafio EmotiW 2015. Fornecemos detalhes sobre estratégias de treinamento por modelo e variações de nossos métodos de fusão. Também apresentamos resultados e discutimos as escolhas que fizemos em cada etapa.

6.1 Desempenho por modelo

Este trabalho se concentra principalmente em uma abordagem RNN para recursos visuais. No entanto, dado o contexto do desafio, incluímos mais três modelos para alcançar desempenho competitivo.
A Tabela 1 mostra a precisão de cada modelo no conjunto de validação de desafio após o treinamento no conjunto de treinamento. As matrizes de confusão correspondentes são apresentadas na Figura 4. As matrizes mostram diferentes perfis e pontos fortes para classes de emoções específicas, o que é benéfico para a combinação.

6.2 Fusão de nível de recurso Como

mencionado anteriormente, excluímos o modelo de atividade da fusão de nível de recurso, pois ele tende a se ajustar demais em sua partição de treinamento. Isso pode ser visto na Tabela 1, onde a atividade tem uma discrepância extremamente alta entre as precisões de treinamento e validação. Os recursos de entrada para a rede de fusão são os seguintes:

- Os primeiros dez componentes dos recursos de áudio brancos do PCA (consulte a Seção 3).
- Os recursos CNN agregados, que são vetores de 105 dimensões (7 × 15 bins), conforme descrito na Seção 2.3.
- Os recursos RNN, que são as ativações ocultas da última etapa de tempo. Esses são os únicos recursos que foram aprendidos discriminativamente no nível do vídeo e que, portanto, contribuem fortemente para a rede de fusão. O número de unidades ocultas no RNN é 200 (ver Seção 2.2).

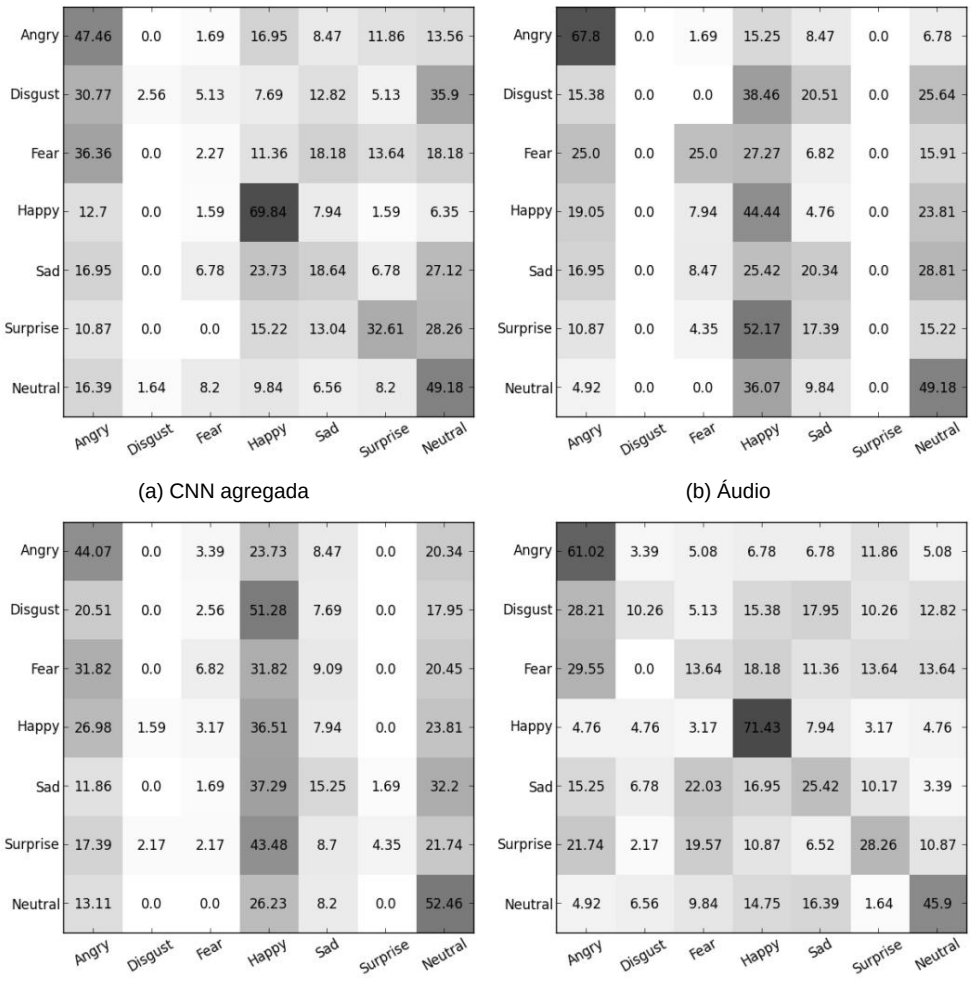
Para treinar a rede de fusão, tentamos substituir a função de ativação sig moid das camadas ocultas por unidades lineares retificadas ReLU(x) = max(0, x) e unidades tanh retificadas RectT anh(x) = max(0, tanh(x)). Enquanto isso melhorou o desempenho de validação em cerca de 2%, não rendeu uma melhoria no desempenho do teste. Uma observação durante o treinamento foi que as curvas de aprendizado estavam oscilando, o que tornava a parada precoce não confiável. Para estabilizar o aprendizado, reduzimos a taxa de aprendizado de 0,1 para 0,001 e adicionamos impulso de 0,9. A Figura 5 compara duas curvas de aprendizado antes e depois da estabilização. O número de épocas em cada subfigura corresponde à taxa de aprendizado selecionada.
Nossa rede de fusão atinge uma precisão de validação de 43,7%, que é maior do que qualquer classificador específico da modalidade.

6.3 Submissões

Nossos envios podem ser divididos em duas categorias: aqueles que usam o conjunto de treinamento para treinamento e o conjunto de validação para parada antecipada e busca aleatória e aqueles para os quais os conjuntos de treinamento e validação foram trocados. Para ambas as categorias, também enviamos uma versão em que os modelos foram retreinados no conjunto completo de treinamento e validação, retendo todos os hiperparâmetros, incluindo número de período de parada inicial e pesos de combinação. Observe que os modelos baseados em CNN também foram retreinados, mas não a CNN subjacente, pois usamos dados adicionais de emoção estática para treinamento. Para todos os envios, a pesquisa aleatória foi feita na partição de dados que não foi usada para treinar os modelos subjacentes. Por exemplo, se os modelos foram treinados na partição de treinamento, a pesquisa aleatória foi realizada no conjunto de validação. Pesquisar na mesma partição em que os modelos foram treinados não era uma opção, pois a pesquisa aleatória atribuiria pesos altos aos over-fitters, o que resultaria em um desempenho de generalização ruim.

A Tabela 2 lista nossos envios com suas precisões de treinamento, validação e teste. Na primeira categoria, treinamos modelos específicos da modalidade e a rede de fusão nos dados de treinamento de desafio e os dados de validação foram usados para parada antecipada. Então, para as previsões finais, realizamos uma pesquisa aleatória no conjunto de validação. Isso alcançou uma precisão de conjunto de teste de 44,341%. Com a rede de fusão estabilizada, a precisão melhorou para 48,979%. Retreinar os modelos com o conjunto combinado de treinamento e validação, mantendo os hiperparâmetros do experimento 2, rendeu uma precisão de teste de 50,463%.

Na segunda categoria, com conjuntos de treinamento e validação trocados, nossa apresentação inicial alcançou uma precisão de teste de 50,092%. Aqui, a fusão estabilizada não melhorou o desempenho, resultando em uma precisão de teste de 47,680%. A versão retreinada alcançou nosso melhor resultado de 52,875%. Observe que, para cada categoria, escolhemos a melhor inscrição para retreinamento. A pesquisa aleatória como a última etapa em nosso pipeline tem uma grande influência no potencial de generalização do todo



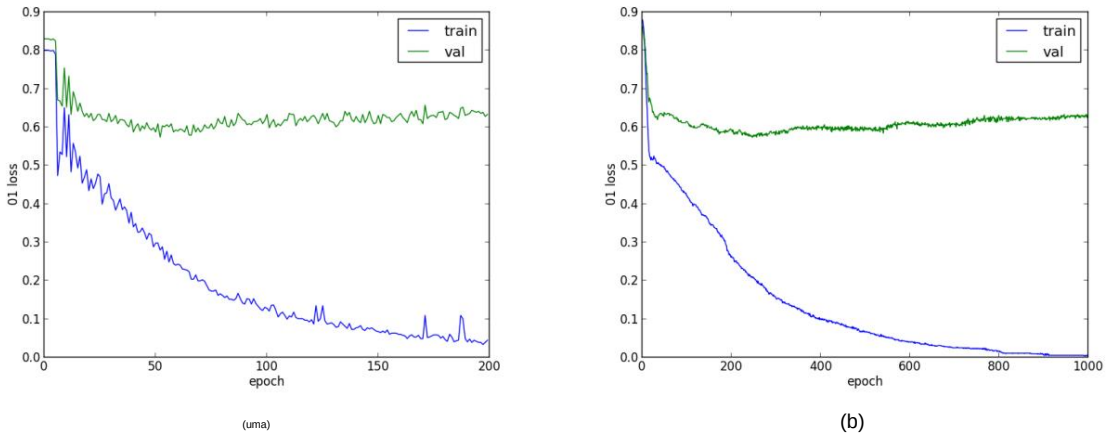
(a) CNN agregada (b) Áudio

(c) Atividade (d) RNN

Figura 4: Matrizes de confusão no conjunto de validação de desafio.

Tabela 2: Nossos envios com precisão de treinamento, validação e teste (em porcentagem) para a competição EmotiW 2015 (a fonte em negrito mostra a melhor precisão)

Teste	válido de subcombolo			Método
1 2	86.216 54.716 44.341			Treinamento em Train, Validação em Valid 81.997
	54.447 48.979			Treinamento em Train, Validação em Valid, fusão estável
3	-	-	50.463	Treinamento em Train+Val, hiperparâmetros da submissão 2, fusão estável
4	52.320 71.967 50.092			Treinamento em Val, Validação em Train 52.742
5	68.463 47.680			Treinamento em Val, Validação em Train, fusão estável 52.875 Treinamento
6	-	-		em Train+Val, hiperparâmetros de envio 4 49.907 Pesquisa aleatória sobre
7	-	-		combinações de envio 3 e 6 em Train+Val



(uma) (b)

Figura 5: Comparação das curvas de aprendizado (a) antes e (b) após a estabilização.

modelo e prováveis benefícios do conjunto de treinamento maior. Isso explica o maior desempenho das partições trocadas.

Nosso último envio foi uma tentativa de combinar nossos dois melhores envios que foram retreinados no conjunto de treinamento e validação. Combinamos os dois usando a mesma estratégia de fusão de nível de decisão de antes. As entradas para a busca aleatória foram as probabilidades previstas pelos dois modelos.
Uma pesquisa aleatória nesses dois modelos foi realizada no conjunto completo de treinamento e validação. O desempenho do teste resultante foi de apenas 49,907%. Isso pode ser explicado pelo fato de que todo o conjunto de dados foi visto, o que pode ter levado a um ajuste excessivo.

7. CONCLUSÕES

Descobrimos que a evolução espaço-temporal das características faciais é uma das pistas mais fortes para o reconhecimento de emoções. Apresentamos a aplicação de uma RNN para modelar essa evolução espaço-temporal por meio da agregação de características faciais para realizar o reconhecimento de emoções em vídeo. Nossos experimentos na Seção 2.3 mostraram que essa abordagem supera todas as outras modalidades, a média de classificações baseadas em visão por quadro e também o método de agregação mais sofisticado empregado pelos vencedores do desafio de 2013 [18].

Além disso, exploramos dois métodos de fusão, operando no recurso e no nível de decisão. Nossa rede de fusão de nível de recurso combina recursos de diferentes modalidades e atinge uma precisão de validação mais alta do que qualquer um dos classificadores de modalidade única. Nossos experimentos mostram que a fusão de nível de recurso e nível de decisão são complementares e, quando combinadas, atingem uma maior precisão de classificação. No entanto, é preciso ter cuidado para evitar o overfitting, seja excluindo overfitters fortes, como fizemos com o modelo de reconhecimento de atividade na rede de fusão, ou usando diferentes partições de conjunto de dados para combinação do que para treinamento do modelo, como feito na busca aleatória.

Achamos difícil tirar conclusões de alguns dos resultados de nossa submissão. Isso pode ser causado pelo grande número de casos ambíguos que existem nesse domínio. Descobrimos que um número razoavelmente grande de vídeos de treinamento poderia mostrar uma mistura de duas ou mais emoções básicas (como uma mistura de surpresa com medo ou felicidade). Esta sugere que explorar o uso de mais de um único rótulo para reconhecimento de emoções pode ser uma direção útil para pesquisas futuras.

8. AGRADECIMENTOS

Os autores gostariam de agradecer aos desenvolvedores do Theano [3, 5]. Este trabalho foi apoiado por um NSERC Discovery Award e pelo BMBF alemão, projeto 01GQ0841. Agradecemos também à Canadian Foundation for Innovation (CFI) pelo apoio ao programa Leaders.

9. REFERÊNCIAS

[1] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia e A. Baskurt. Aprendizagem profunda sequencial para reconhecimento de ação humana. Em A. Salah e B. Lepri, editores, Human Behavior Understanding, volume 7065 de Lecture Notes in Computer Science, páginas 29–39. Springer Berlin Heidelberg, 2011.

[2] D. Bahdanau, K. Cho e Y. Bengio. Tradução automática neural aprendendo em conjunto a alinhar e traduzir. CoRR, abs/1409.0473, 2014.

[3] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley e Y. Bengio. Theano: novos recursos e melhorias de velocidade. arXiv preprint arXiv:1211.5590, 2012.

[4] J. Bergstra e Y. Bengio. Pesquisa aleatória para otimização de hiperparâmetros. The Journal of Machine Learning Research, 13(1):281–305, 2012.

[5] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley e Y. Bengio. Theano: um compilador de expressões matemáticas de CPU e GPU. Em Proceedings of the Python for Scientific Computing Conference (SciPy), volume 4, página 3. Austin, TX, 2010.

[6] P.-L. Carrier, A. Courville, IJ Goodfellow, M. Mirza e Y. Bengio. Banco de Dados de Faces FER-2013. Relatório técnico, 1365, Universit´e de Montr´eal, 2013.

[7] A. Dhall, R. Goecke, J. Joshi, K. Sikka e T. Gedeão. Reconhecimento de emoções no desafio selvagem 2014: linha de base, dados e protocolo. Em Processos de

a 16ª Conferência Internacional sobre Interação Multimodal, ICMI '14, páginas 461–466, Nova York, NY, EUA, 2014. ACM.

[8] A. Dhall, R. Goecke, S. Lucey e T. Gedeon. Coleta de bancos de dados de expressões faciais grandes e ricamente anotados de filmes. MultiMedia, IEEE, 19(3):34–41, julho de 2012.

[9] A. Dhall, OVR Murthy, R. Goecke, J. Joshi e T. Gedeon. Desafios de reconhecimento de emoções baseados em vídeo e imagem na natureza: Emotiw 2015. In Proceedings of the 17th ACM on International Conference on Multimodal Interaction, ICMI '15, 2015.

[10] J. Donahue, LA Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko e T. Darrell. Redes convolucionais recorrentes de longo prazo para reconhecimento e descrição visual. 2014.

[11] F. Eyben, M. Wöllmer e B. Schuller. Opensmile: o versátil e rápido extrator de recursos de áudio de código aberto de Munique. Em Proceedings of the international conference on Multimedia, páginas 1459–1462. AC, 2010.

[12] F. Eyben, M. Wöllmer e B. Schuller. openear - apresentando o kit de ferramentas de reconhecimento de emoções e afetos de código aberto de Munique. Em ACII, páginas 576–581, 2009.

[13] A. Graves, A.-r. Mohamed e G. Hinton. Reconhecimento de fala com redes neurais recorrentes profundas. Em Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, páginas 6645–6649. IEEE, 2013.

[14] A. Graves e J. Schmidhuber. Reconhecimento de escrita offline com redes neurais recorrentes multidimensionais. Em Advances in Neural Information Processing Systems, páginas 545–552, 2009.

[15] S. Hochreiter e J. Schmidhuber. Memória de longo prazo. Computação neural, 9(8):1735–1780, 1997.

[16] SE Kahou, X. Bouthhillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, R. Chandias Ferrari, M Mirza, D. Warde-Farley, A. Courville, P. Vincent, R. Memisevic, C. Pal e Y. Bengio. Emonets: Abordagens de aprendizado profundo multimodal para reconhecimento de emoções em vídeo. Journal on Multimodal User Interfaces, páginas 1–13, 2015.

[17] SE Kahou, P. Froumenty e C. Pal. Facial análise de expressão baseada em recursos binários de alta dimensão. No ECCV Workshop on Computer Vision with Local Binary Patterns Variants, Zurique, Suíça, 2014.

[18] SE Kahou, C. Pal, X. Bouthhillier, P. Froumenty, C. Gulcehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, RC Ferrari, M. Mirza, S. Jean, P.-L. Carrier, Y. Dauphin, N. Boulanger-Lewandowski, A. Aggarwal, J. Zumer, P. Lamblin, J.-P. Raymond, G. Desjardins, R. Pascanu, D. Warde-Farley, A. Torabi, A. Sharma, E. Bengio, M. Cˆot´e, KR Konda e Z. Wu. Combinando redes neurais profundas específicas da modalidade para reconhecimento de emoções em vídeo. In Proceedings of the 15th ACM on International Conference on Multimodal Interaction, ICMI '13, 2013.

[19] N. Kalchbrenner, E. Grefenstette e P. Blunsom. Uma rede neural convolucional para modelagem de sentenças. arXiv:1404.2188, 2014.

[20] KR Konda, R. Memisevic e V. Michalski. Aprender a codificar o movimento usando a sincronia espaço-temporal. In Proceedings of ICLR, abril de 2014.

[21] A. Krizhevsky. Página inicial do código do Google Cuda-convnet. https://code.google.com/p/cuda-convnet/, agosto de 2012.

[22] A. Krizhevsky, I. Sutskever e GE Hinton. Classificação de Imagenet com redes neurais convolucionais profundas. Em Avanços em sistemas de processamento de informações neurais, páginas 1097–1105, 2012.

[23] Q. Le, W. Zou, S. Yeung e A. Ng. Aprendendo recursos espaço-temporais invariantes hierárquicos para reconhecimento de ação com análise de subespaço independente. CVPR, 2011.

[24] QV Le, N. Jaitly e GE Hinton. Uma maneira simples para inicializar redes recorrentes de unidades lineares retificadas. arXiv preprint arXiv:1504.00941, 2015.

[25] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang e X. Chen. Combinando vários métodos de kernel no coletor riemanniano para reconhecimento de emoções na natureza. Em Proceedings of the 16th International Conference on Multimodal Interaction, ICMI '14, páginas 494–501, Nova York, NY, EUA, 2014. ACM.

[26] Nagadomi. Github: kaggle-cifar10-torch7. https://github.com/nagadomi/kaggle-cifar10-torch7/, 2014.

[27] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie e M. Pantic. Avec 2011 – o primeiro desafio internacional de emoção audiovisual. Em Computação Afetiva e Interação Inteligente, páginas 415–424. Primavera, 2011.

[28] C. Shan, S. Gong e PW McOwan. Facial reconhecimento de expressão baseado em padrões binários locais: um estudo abrangente. Image Vision Comput., 27(6):803–816, maio de 2009.

[29] K. Simonyan e A. Zisserman. Redes convolucionais muito profundas para reconhecimento de imagens em larga escala. CoRR, abs/1409.1556, 2014.

[30] Y. Sun, X. Wang e X. Tang. Cascata de rede convolucional profunda para detecção de pontos faciais. Em Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '13, páginas 3476–3483, Washington, DC, EUA, 2013. IEEE Computer Society.

[31] J. Susskind, A. Anderson e G. Hinton. O banco de dados de rostos de Toronto. Relatório técnico, UTML TR 2010-001, Universidade de Toronto, 2010.

[32] I. Sutskever, O. Vinyals e QV Le. Aprendizado de sequência a sequência com redes neurais. Em Avanços em sistemas de processamento de informações neurais, páginas 3104–3112, 2014.

[33] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye e X. Xue. Modelando pistas espaço-temporais em uma estrutura de aprendizagem profunda híbrida para classificação de vídeo. arXiv preprint arXiv:1504.01561, 2015.