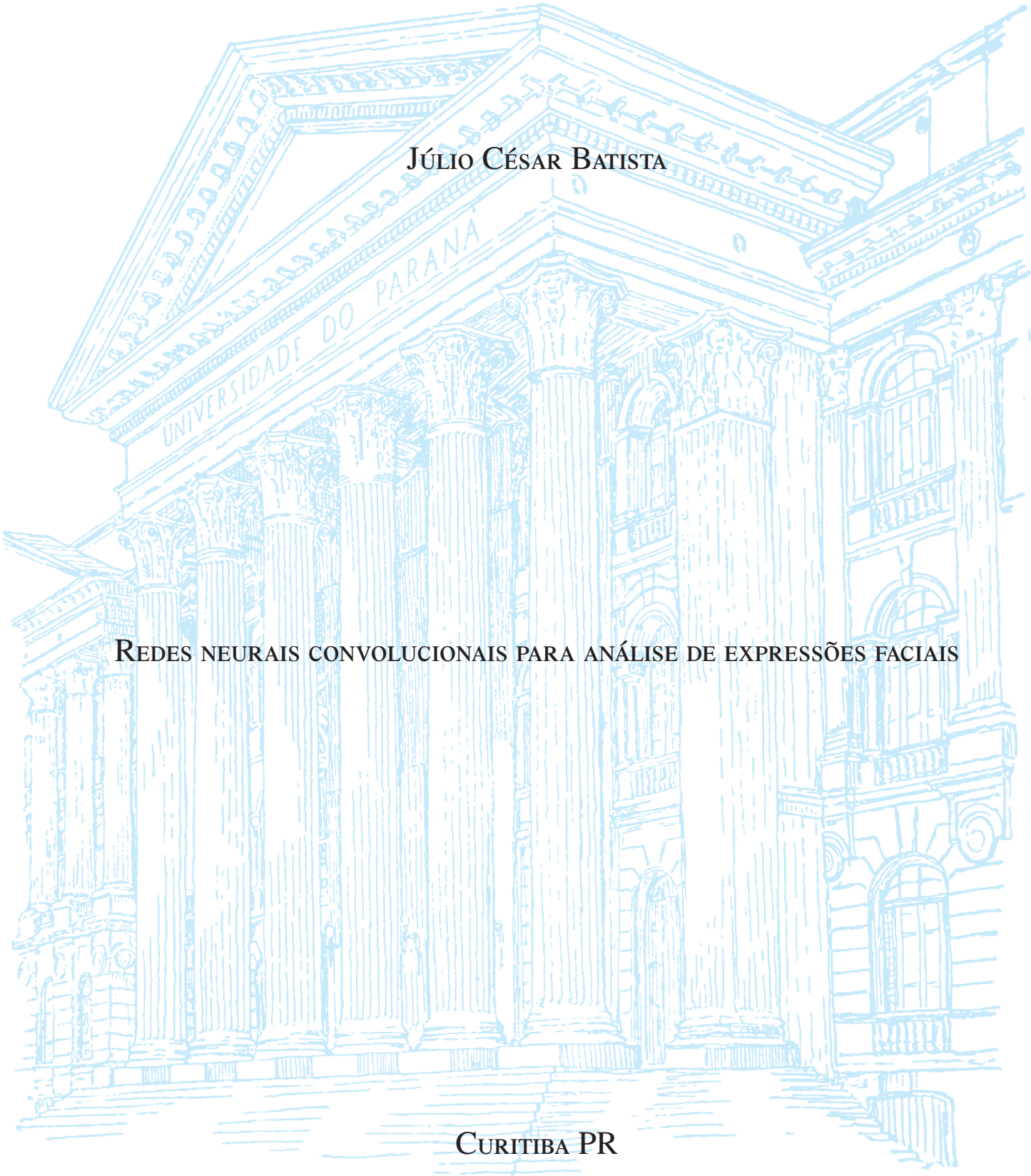
Universidade Federal do Paraná

Júlio César Batista



Redes neurais convolucionais para análise de expressões faciais

Curitiba PR
2018

Júlio César Batista

Redes neurais convolucionais para análise de expressões faciais

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Informática, no Programa de Pós-Graduação em Informática, setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: Luciano Silva.

Coorientador: Olga R. P. Bellon.

Curitiba PR
2018

FICHA CATALOGRÁFICA ELABORADA PELO SISTEMA DE BIBLIOTECAS/UFPR
BIBLIOTECA DE CIÊNCIA E TECNOLOGIA

B333r Batista, Júlio César
Redes neurais convolucionais para análise de expressões faciais / Júlio César Batista. – Curitiba, 2018.

Dissertação (Mestrado) - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-Graduação em Informática, 2018.

Orientador: Prof. Dr. Luciano Silva.
Coorientadora: Profa. Dra. Olga R. P. Bellon.

1. Expressões faciais. 2. Redes neurais convulcionais. 3. Imagem 3D. I. Universidade Federal do Paraná. II. Silva, Luciano. III. Bellon, Olga R. P. IV. Título.

CDD: 006.32

Bibliotecária: Romilda Santos - CRB-9/1214

MINISTÉRIO DA EDUCAÇÃO
SETOR SETOR DE CIENCIAS EXATAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO INFORMÁTICA

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em INFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **JÚLIO CÉSAR BATISTA** intitulada: **Redes neurais convolucionais para análise de expressões faciais**, após terem inquirido o aluno e realizado a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 31 de Agosto de 2018.

LUCIANO SILVA
Presidente da Banca Examinadora (UFPR)

OLGA REGINA PEREIRA BELLON
Coorientador - Avaliador Interno (UFPR)

AURORA TRINIDAD RAMIREZ POZO
Avaliador Interno (UFPR)

JOÃO PAULO PAPA
Avaliador Externo (UNESP)

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
Setor de Ciências Exatas
Coordenação de Pós-Graduação em Informática

Dedico este trabalho aos meus pais, Marina e Manoel, e também à Juliana, minha namorada.

Agradecimentos

Gostaria de agradecer à minha família. Principalmente, à meus pais, Marina e Manoel que sempre me incentivaram a estudar e a correr atrás de meus objetivos. Também gostaria de agradecer a minha namorada, Juliana, pelo carinho, apoio e compreensão enquanto caminhei por essa jornada.

Agradeço também ao meu orientador, professor Dr. Luciano Silva pelas várias discussões, conversas e ideias que compartilhamos. Não menos importante, agradeço à minha co-orientadora, professora Dra. Olga Bellon, por me fazer colocar as ideias em prática e no papel. Também aos meus amigos do IMAGO que compartilharam essa jornada comigo, com certeza vocês fazem parte dessa história e produziram inúmeras ideias e conversas sobre as mais diversas ideias.

Resumo

Este trabalho propõe uma rede neural convolucional (CNN) para efetuar a detecção e estimativa de intensidade de *Action Units* (AUs), de forma simultânea, em imagens de faces em poses arbitrárias. Na literatura existem vários métodos para detectar e estimar intensidades de AUs, entretanto, poucos lidam com as variações na pose e levam em consideração a correlação entre os AUs e as intensidades. Ainda, ao considerar a inferência conjunta surge o problema de desequilíbrio entre a quantidade de anotações para cada classe, o que dificulta o processo de otimização e generalização. Porém, é necessário lidar com essas restrições para que esses métodos possam ser utilizados em ambientes não controlados. Outro detalhe que dificulta a generalização para esses ambientes é a falta de bases de imagens anotadas. Nesse caso, é possível estender bases com modelos 3D para gerar poses arbitrárias de forma sintética assim como feito no *Facial Expression Analysis and Recognition Challenge* (FERA) 2017. Portanto, utilizando uma base de poses sintéticas, este trabalho propõe um modelo baseado em uma CNN, chamado AUMPNet, e aprendizado multi-tarefa para detectar e estimar a intensidade de AUs. Além do modelo para inferência conjunta, também é demonstrada uma abordagem para diminuir o desequilíbrio entre as intensidades dos AUs durante a otimização. O desempenho do modelo proposto, utilizando as bases do FERA 2015 e FERA 2017, é similar ao estado-da-arte, sendo superior para algumas AUs individualmente.

Palavras-chave: análise de expressões faciais, visão computacional, redes neurais convolucionais.

Abstract

This work presents a convolutional neural network (CNN) for joint Action Unit (AU) detection and intensity estimation on images of face in arbitrary head poses. There are a variety of approaches for AU detection and intensity estimation, however, few of them take into account head pose variations and the correlations among AUs and their intensities. Still, the problem of class imbalance appears when considering the joint inference of AUs, making optimization and generalization harder. Though, it is required to cope with these constraints in order to apply these methods in unconstrained environments. Another difficulty is the lack of labelled images in these conditions. In this case, it is possible to extend existing databases of 3D models to produce synthetic images in arbitrary head poses as in Facial Expression Recognition and Analysis Challenge (FERA) 2017. Thus, by using this database of synthetic head poses this work proposes a multi-task CNN based model, called AUMPNet, to detect AUs and estimate their intensity. Moreover, an approach to handle class imbalance among AUs during optimization is shown. The proposed model, when applied on the FERA 2015 and FERA 2017 databases, achieves average results comparable to the state-of-the-art, and surpasses them for some AUs individually.

Keywords: facial expression analysis, computer vision, convolutional neural networks.

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

3D	3 dimensões
AU	*Action Unit*
CNN	Rede Neural Convolucional
DINF	Departamento de Informática
FACS	*Facial Action Coding System*
FERA	*Facial Expression Recognition and Analysis Challenge*
HOG	*Histogram of Oriented Gradients*
LBP	*Local Binary Patterns*
PCA	*Principal Component Analysis*
RGB	*Red, Green, Blue*
SVM	*Support Vector Machine*
SVR	*Support Vector Regression*
UFPR	Universidade Federal do Paraná

Lista de Símbolos

α	*Learning rate*
μ	Média
σ	Desvio padrão
δ	Uma margem
$\propto$	Proporcial
λ	Um peso
π	Constante pi
$\mathbb{R}$	Conjunto dos números reais
$\frac{\partial f}{\partial x}$	Derivada de uma função f em relação ao parâmetro x

Sumário

Capítulo 1

Introdução

Pessoas utilizam expressões faciais diariamente para comunicação não verbal e para expressar sentimentos. Com a riqueza de informações que são transmitidas a partir de faces e expressões é possível trabalhar em melhores aplicações para as áreas de computação afetiva, interação humano computador e saúde. Essas expressões são, normalmente, categorizadas em um conjunto definido por: alegria, tristeza, nojo, medo, raiva, surpresa e também a face neutra. Entretanto, a análise de expressões utiliza o movimento de músculos da face, conhecidos como *Action Units* (AUs), para realizar essa categorização. Os AUs são definidos pelo *Facial Action Coding System* (FACS), proposto por Ekman et al. (2002), com base na anatomia da face. O FACS define um conjunto de 32 AUs, com os códigos e região da face onde ocorrem listados na Tabela 1.1, e as expressões faciais são criadas a partir da combinação de determinados AUs. Por exemplo, a expressão de alegria é determinada pela ocorrência do AU12 (movimento do canto dos lábios) e, eventualmente, do AU06 (movimento da bochecha) e AU25 (separação dos lábios). Conforme Du et al. (2014), o AU06 não precisa ocorrer para caracterizar um sorriso visto que muitas pessoas não demonstram esse AU ao sorrir. Por fim, a Figura 1.1 contém exemplos dos principais AUs e os que não aparecem neste figura menos comuns, por exemplo o AU 21 representa o movimento no pescoço (*Neck Tightener*).

Além da possibilidade da análise de expressões faciais a partir dos músculos da face, utilizar AUs para essa tarefa pode levar à um melhor entendimento de expressões faciais. Conforme Du et al. (2014); Du e Martinez (2015), as expressões faciais, e emoções, demonstradas diariamente não ocorrem em apenas seis categorias, mas podem ser compostas. Assim, determinar a expressão pela ocorrência de AUs é melhor do que elaborar todas as classes possíveis. Finalmente, Du e Martinez (2015) define a importância não só da ocorrência de AUs, mas também da intensidade com a qual são demonstrados nas expressões faciais.

O FACS também define uma escala ordinal de cinco níveis de intensidade para os AUs. Essa intensidade representa o nível de contração do músculo. A análise da intensidade de AUs permite aplicações como identificar se a expressão é verdadeira, proposto por Martinez et al. (2017); identificar se uma pessoa está sentindo dor, conforme Walecki et al. (2016); até mesmo,

Tabela 1.1: AUs definidos pelo FACS e a região da face onde eles ocorrem. Adaptado de: Martinez et al. (2017).

Região da face	AUs
Superior	1, 2, 4, 5, 6, 7, 43, 45, 46
Inferior	9, 10, 11, 12, 13, 14, 15, 17, 18, 20, 22, 23, 24, 25, 28, 27, 28

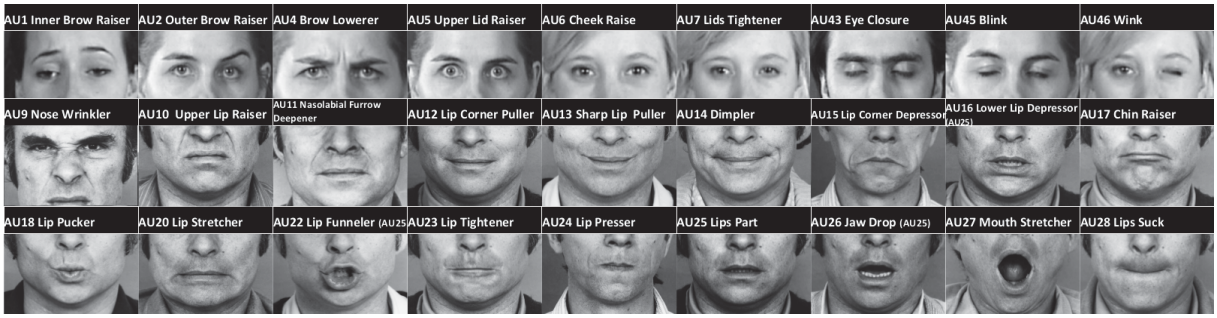


Figura 1.1: Exemplos de 27 AUs definidos pelo FACS. Fonte: Martinez et al. (2017).

como em Girard et al. (2015), identificar se a pessoa possui doenças psicológicas. Por fim, a Figura 1.2 demonstra um exemplo das intensidades de AUs.

A partir da Figura 1.2 é possível perceber uma propriedade dos AUs, a correlação entre AUs e intensidades. Nesse caso, quando um AU ocorre na face, existe a chance de outro AU também ocorrer, como é o caso dos AU06, AU12 e AU25 que formam um sorriso. Essa correlação ainda pode ser elevada ao nível das intensidades, conforme a Figura 1.2 demonstra, porque ao aumentar a intensidade do AU12, o AU25 surge naturalmente. A correlação entre os AUs também implica uma característica de adividade nos AUs. Nesse caso, AUs na mesma região da face são considerados não aditivos, ou seja, um AU interfere na aparência do outro com demonstrado na Figura 1.3. Quando os AUs são aditivos, a mudança em um AU não interfere no outro, como no exemplo dos AU12 e AU25 da Figura 1.2.

Em princípio a detecção de AUs é a a primeira tarefa na análise de expressões faciais a partir do FACS. Essa tarefa de indica se um determinado AU está presente na imagem ou não. Sendo que ela pode ser modelada com um classificador binário, conforme proposto por Baltrušaitis et al. (2015); Tősér et al. (2016), mas como uma imagem pode conter vários AUs presentes simultaneamente, é necessário utilizar um classificador binário para cada AU e efetuar a detecção com cada modelo para cada imagem. Entretanto, como a detecção de AUs é dada como uma tarefa de classificação binária, é possível utilizar um modelo de múltiplos rótulos, conforme Yüce et al. (2015); Tősér et al. (2016); Zhao et al. (2016), que efetua a inferência de todos os AU ao mesmo tempo, sem a necessidade de um modelo por AU.

Dado que as intensidades dos AUs constituem um conjunto $S = \{0,1,2,3,4,5\}$, sendo os cinco níveis definidos pelo FACS e também a intensidade 0 indicando que o AU não está presente na imagem. Essa definição permite modelar a tarefa de estimar a intensidade de AUs como classificação multiclasse, regressão e ranqueamento. Visto que os níveis de intensidade tem uma característica ordinal, ou seja $0 < 1 < 2 < 3 < 4 < 5$, os modelos de ranqueamento e regressão são as abordagens mais sugeridas. Um ponto importante a ser analisado é que esses métodos, normalmente, requerem um modelo por AU e não fazem uso de uma representação conjunta. Com isso, não é possível levar em consideração as correlações entre os AUs e suas

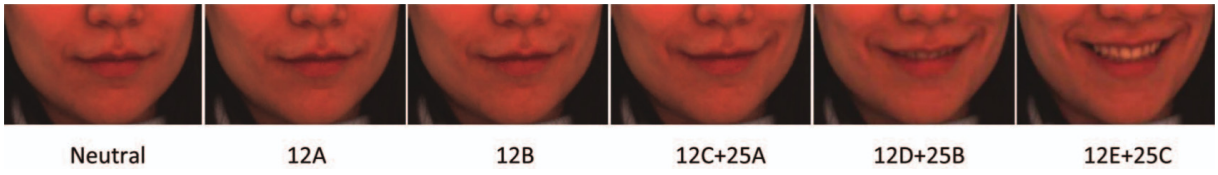


Figura 1.2: Níveis de intensidade de AUs. O primeiro quadro demonstra a face neutra que muda desde um movimento leve até uma contração maior do AU12 (canto dos lábios). É possível notar também a correção que existe com o AU25 (separação dos lábios). Fonte: Mavadati et al. (2013)

intensidades. Para superar essa limitação, modelos baseados em árvores latentes, utilizados por Kaltwang et al. (2015), ou em aprendizado estruturado, conforme Walecki et al. (2016, 2017), podem ser utilizados.

Além dos modelos de aprendizado de máquina que são utilizados na estimativa da intensidade de AUs, também é necessário uma representação eficiente da imagem. Normalmente, a representação é dada pela textura e pela geometria da face. A parte geométrica é computada a partir de *landmarks* e envolve o cálculo de ângulos e distâncias entre esses pontos. A textura da face pode ser computada utilizando descritores como em Baltrušaitis et al. (2015); Nicolle et al. (2016) que utilizaram *Histogram of Oriented Gradients* (HOG), ou em Kaltwang et al. (2015) que usou *Local Binary Patterns* (LBP) ou com filtros de Gabor usados em Benitez-Quiroz et al. (2016); Du et al. (2014). Outros modelos envolvem o uso de redes neurais convolucionais, como em Zhao et al. (2016); Tősér et al. (2016); Gudi et al. (2015); Batista et al. (2017), que otimizam uma representação intermediária com base em *loss function*.

Os parágrafos seguintes descrevem os trabalhos relacionados citados anteriormente que tem maior relação com o trabalho proposto. Para uma revisão completa dos métodos computacionais para análise de expressões faciais, o leitor é convidado a verificar as seguintes publicações Sariyanidi et al. (2015), Corneanu et al. (2016) e Martinez et al. (2017).

Baltrušaitis et al. (2015) propôs a detecção de AUs utilizando *Support Vector Machines* (SVMs) e a estimativa de intensidade utilizando *Support Vector Regression* (SVR). A entrada de ambos os modelos é mesma, consistindo de características geométricas computadas a partir de *landmarks*; características de textura computadas com HOG seguido de *Principal Component Analysis* (PCA) para redução de dimensionalidade. Além disso, também foi feito um *under-sampling* de AUs para reduzir o desequilívrio existente entre as classes. Finalmente, como este trabalho utiliza vídeos e não imagens separadas, os autores estimam um descritor base para cada vídeo a fim de ter uma referência da face neutra para cada sujeito.

Figura 1.3: Exemplo de combinação não-aditiva de AUs. Em todoas as imagens o AU 1 está presente, mas a aparência dele é modificada pelos outros AUs presentes. Fonte: Benitez-Quiroz et al. (2017).

Tősér et al. (2016) propôs uma rede neural convolucional (CNN) para detectar AUs em múltiplas poses. As variações de pose foram geradas a partir de transformações nos modelos 3D disponíveis na base. Esse trabalho apresenta um modelo simplificado com poucas camadas que filtram a imagem completa, obtendo uma representação holística da face. Além do modelo, a otimização também foi comparada ao otimizar um modelo para cada AU ou um modelo para todos os AUs. Nos resultados obtidos, a otimização de um modelo por AU obteve os melhores resultados. Por fim, a otimização foi feita utilizando Adamax com um critério de parada precoce caso o resultado de avaliação não melhore em m épocas.

Zhao et al. (2016) também propôs o uso de CNN para a detecção de AUs. Nesse caso, o modelo proposto de CNN possui uma camada intermediária que separa a face em 8×8 (64) regiões. Para cada uma dessas regiões, são aplicados filtros e não-linearidades e depois as regiões são agrupadas novamente. Essa separação é regiões é uma proposta para capturar vairações em regiões específicas da face, visto que os AUs são movimentos que geram mudanças em locais específicos da face. Finalmente, esse modelo foi otimizado para detecção de múltiplos AUs utilizando entropia cruzada, da mesma forma que Tősér et al. (2016).

Walecki et al. (2017) propôs um modelo de aprendizado estruturado (modelo de grafo probabilístico) para estimar a intensidade conjunta de AUs. Assim como em Tősér et al. (2016) e Zhao et al. (2016), a representação foi computada com o uso de CNNs. Entretanto, a inferência conjunta da intensidade de AUs foi feita com o uso de *Conditional Random Fields* (CRFs) com funções copula. O objetivo dessa abordagem é modelar os padrões de co-ocorência entre os AUs e suas intensidades. Além disso, esse modelo permite modelar a natureza ordinal das intensidades dos AUs, mantendo a relação que a intensidade 1 é menor que a 2, por exemplo. Por fima, também foi proposto um algoritmo para equilibrar a amostragem de intensidades nos *batches* usados para otimização.

Com base nos métodos existentes e nas limitações existentes em relação à imagens não-frontais e também das correlações entre os AUs, este trabalho propõe:

- Um modelo unificado para inferência conjunta da detecção e estimativa de intensidade de AUs em imagens de faces não-frontais;
- A adaptação para uma *loss function* de regressão para levar em consideração as características nas intensidades de AUs;
- Uma abordagem para amostrar imagens durante a otimização para diminuir o desequilíbrio entre os níveis de intensidade dos AUs.

Finalmente, esta dissertação resultou nas seguintes publicações: Batista et al. (2016), Batista et al. (2017) e Zavan et al. (2017).

Capítulo 2

AUMPNet: um modelo unificado para detecção e estimativa de intensidade de AUs

Neste capítulo é descrito o modelo proposto para análise de expressões faciais baseado na detecção e estimativa de intensidade de AUs. O capítulo está organizado da seguinte forma: a Seção 2.1 descreve a base de imagens e como ela foi organizada; na sequência, a Seção 2.2 descreve a arquitetura do modelo e a Seção 2.3 demonstra como ele foi otimizado. Finalmente, a Seção 2.4 descreve os experimentos e protocolo de avaliação; a Seção 2.5 apresenta e discute os resultados obtidos.

2.1 Organização das imagens e anotações

Dada uma base de imagens com anotações para treino, os dados foram organizados em conjunto $\mathcal{D} = \{\mathbf{X}, \mathbf{P}, \mathbf{O}, \mathbf{Y}\}$ com M amostras. A matriz $\mathbf{X} \in \mathbb{R}^{M \times 132 \times 132 \times 3}$ contém M imagens RGB de tamanho 132x132 normalizadas para média 0 e desvio padrão 1 usando $\mu = \sigma = 128$; a pose principal em cada imagem está no vetor $\mathbf{P}$ que contém M elementos e $\mathbf{P}_i \in \{1,2,3,4,5,6,7,8,9\}$ representa a pose da i-ésima imagem; as ocorrências de AUs estão na matriz $\mathbf{O}$, que tem formato $M \times Q$, e $\mathbf{O}_{iq} \in \{0,1\}$ indica a presença do q-ésimo AU na i-ésima imagem; as intensidades dos AUs estão na matriz $\mathbf{Y}$, que tem formato $M \times K$, e a intensidade do k-ésimo AU na i-ésima imagem é indicada por $\mathbf{Y}_{ik} \in \{0,1,2,3,4,5,9\}$. O caso $\mathbf{Y}_{ik} = 9$ indica que a intensidade é desconhecida.

2.1.1 Detecção de faces

A Faster R-CNN, proposto por Ren et al. (2015), foi utilizada na detecção de faces porque possui resultados estado-da-arte para detecçao de objetos. Além disso, Jiang e Learned-Miller (2017) realizou estudos de como esse detector deve ser utilizado para detecção de faces. Apesar que a Faster R-CNN é uma CNN proposta para detecção de objetos, a versão utilizada foi otimizada para detecção de faces. A detecção resulta em *bounding boxes* definidos pela tupla (x, y, w, h) que indicam, respectivamente, o canto superior esquerdo, a largura e a altura. Essas tuplas não possuem proporção definida, portanto, elas foram ajustadas para serem quadrados perfeitos no formato (x, y, max(w, h), max(w, h)). Após o recorte para o formato quadrado, as imagens foram redimensionadas para o tamanho 132x132 conforme requisitado na Seção 2.1. Como a Faster R-CNN efetua detecções em aproximadamente 5 *frames per second* (FPS), um rastreador

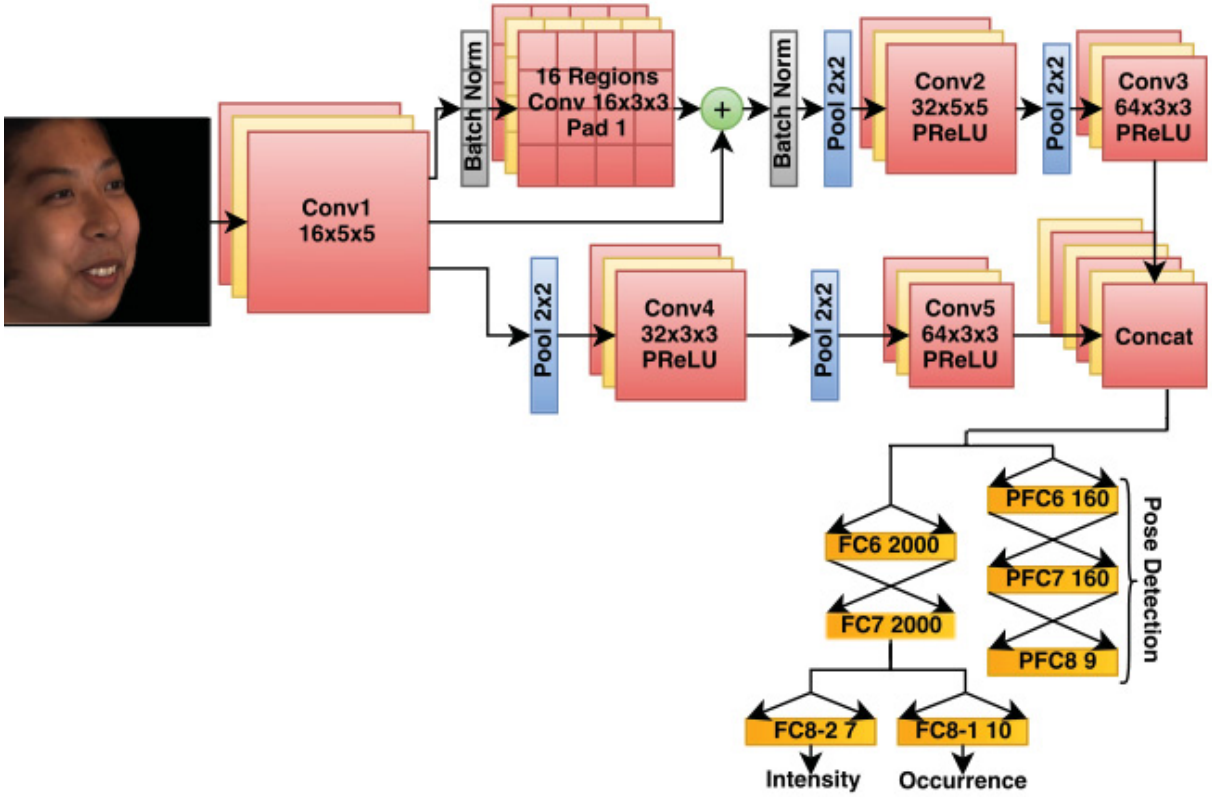


Figura 2.1: Visão geral do modelo proposto. Fonte: Batista et al. (2017).

de objetos genérico, disponível em King (2009), foi utilizado para agilizar este processo. O rastreador disponível é uma implementação do trabalho proposto por Danelljan et al. (2014).

2.2 Arquitetura do modelo

O modelo proposto consiste em uma CNN, nomeada AUMPNet, que efetua de forma simultânea a detecção e estimativa de intensidade de AUs. Nesse cenário, a detecção de AUs foi formulada como detecção de múltiplos rótulos e a estimativa de intensidade como regressão múltipla. Assim, é possível efetuar o aprendizado conjunto de forma que a representação intermediária do modelo seja otimizada para as duas tarefas. A otimização conjunta das duas tarefas é possível através do aprendizado multi-tarefa. Nesse modelo, a otimização é feita com base em múltiplas funções objetivo (*loss*), sendo uma para cada tarefa. Uma visão geral do modelo proposto pode ser vista na Figura 2.1.

A entrada do modelo é uma imagem RGB de tamanho 132x132x3 que é processada por uma série de operações de convolução, ativação e *pooling*, conforme demonstrado na Figura 2.1. Inicialmente, a imagem é filtrada por uma convolução no formato 5x5x16 e depois o modelo é dividido em duas partes. Uma parte utiliza o conceito de *region learning*, proposto por Zhao et al. (2016), para representar regiões da face, visto que AUs são movimentos locais na face. A outra parte aplica filtros em toda a região da face para aprender uma representação mais robusta considerando as possíveis mudanças de aparência que um AU pode causar em outros AUs. Na sequência, os filtros locais e globais são concatenados em uma representação única que então é utilizada em camadas totalmente conectadas. É importante notar o uso de uma terceira otimização para a pose, baseada no trabalho de Zavan et al. (2016). Entretanto, essa otimização só ocorre durante o processo de treino para gerar representações que sejam mais robustas em

relação as mudanças de pose. Em tempo de inferência essa camadas são removidas do modelo. Finalmente, o modelo termina com duas camadas totalmente conectadas que representam a detecção de AUs (FC8-1) e a estimativa de intensidade (FC8-2). Nesse caso, a estimativa das ocorrências $\hat{\mathbf{O}}$ possui Q unidades e a estimativa de intensidade $\hat{\mathbf{Y}}$ possui K unidades.

A inicialização dos filtros foi feita conforme o procedimento proposto por He et al. (2015). As camadas totalmente conectadas foram inicializadas a partir de distribuições Gaussianas $\mathcal{N}(0, 0.05)$ e $\mathcal{N}(0, 0.01)$. As funções de ativação de todas as camadas são *Parametric* ReLU (PReLU), descrita em He et al. (2015). É importante notar que as camadas totalmente conectadas relacionadas à estimativa da pose seguem o modelo proposto por Zavan et al. (2016). Finalmente, para auxiliar na regularização do modelo, uma camada de *dropout*, descrito por Srivastava et al. (2014), com probabilidade de 0.2 foi aplicada após a camada FC6.

2.3 Otimização do modelo

A otimização para detecção de AUs foi formulada como entropia cruzada de sigmóide. Essa formulação (Equação 2.1) foi utilizada devido a estimativa de múltiplos rótulos binários para cada imagem.

$$\mathcal{L}_{\text{occ}}(\hat{\mathbf{O}}, \mathbf{O}) = -\frac{1}{NQ}\sum_{n=1}^{N}\sum_{q=1}^{Q}\mathbf{O}_{nq}\log(\sigma(\hat{\mathbf{O}}_{nq})) + (1-\mathbf{O}_{nq})\log(1-\sigma(\hat{\mathbf{O}}_{nq})) \tag{2.1}$$

A função apresentada na Equação 2.1 recebe como entradas as estimativas do modelo ($\hat{\mathbf{O}}$) e as anotações ($\mathbf{O}$) conforme a Seção 2.1; σ representa a função sigmóide e a normalização é feita sobre a quantidade de imagens no *batch* (N) multiplicado pela quantidade de AUs (Q). Para inferência, a ocorrência de um determinado AU pode ser obtida utilizando 2.2, sendo que $1\{\cdot\}$ retorna 1 caso a condição for verdadeira e 0 caso contrário.

$$Th(\hat{\mathbf{O}}_q) = 1\{\hat{\mathbf{O}}_q > 0\} \tag{2.2}$$

Conforme mencionado na Seção 2.2, a estimativa de intensidade foi formulada como regressão múltipla. Nesse caso, a otimização foi feita utilizando a média da soma dos erros ao quadrado, apresentada na Equação 2.3.

$$\mathcal{L}_{\text{int}}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2N}\sum_{n=1}^{N}\|\mathbf{w}_n \circ (5\cdot\hat{\mathbf{Y}}_n - \mathbf{Y}_n)\|_2^2 \tag{2.3}$$

As entradas da função apresentada na Equação 2.3 são as estimativas do modelo ($\hat{\mathbf{Y}}$), normalizadas no intervalo (0, 1) utilizando a função sigmóide, e as anotações ($\mathbf{Y}$). A normalização também é feita utilizando o número de imagens (N) no *batch*. Porém, o escalar 2 é utilizado apenas como uma conveniência para facilitar o cálculo da derivada. Além disso, o vetor de pesos $\mathbf{w}$, definido na Equação 2.4, remove as intensidades com valor 9 da otimização, visto que esse valor indica que a intensidade é desconhecida. Finalmente, a Equação 2.5 mostra como as intensidades são obtidas durante a inferência.

$$\mathbf{w}_{nk} = 1\{\mathbf{Y}_{nk} \neq 9\} \tag{2.4}$$

$$\tau(\hat{\mathbf{Y}}) = [5\cdot\hat{\mathbf{Y}}] \tag{2.5}$$

A *loss function* multi-tarefa utilizada na otimização do modelo é demonstrada na Equação 2.6. Essa função é composta pela otimização da ocorrência de AUs, intensidade de AUs e otimização da pose. A otimização da pose ($\mathcal{L}_{\text{pose}}(\hat{\mathbf{P}}, \mathbf{P})$) é um *softmax*, conforme descrito em Zavan et al. (2016). Para cada uma das funções utilizadas existe um peso λ que indica o quanto deve ser ponderado para cada tarefa. Como o principal objetivo do modelo é a otimização para os AUs, $\lambda_{\text{occ}} = \lambda_{\text{int}} = 1$ e $\lambda_{\text{pose}} = 0,5$. Dessa forma, o modelo tem menos peso na otimização da pose e ela é utilizada como uma espécie de regularização para as variações de pose.

$$\mathcal{L}(\hat{\mathbf{O}}, \mathbf{O}, \hat{\mathbf{Y}}, \mathbf{Y}, \hat{\mathbf{P}}, \mathbf{P}) = \lambda_{\text{occ}}\mathcal{L}_{\text{occ}}(\hat{\mathbf{O}}, \mathbf{Y}) + \lambda_{\text{int}}\mathcal{L}_{\text{int}}(\hat{\mathbf{Y}}, \mathbf{Y}) + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}}(\hat{\mathbf{P}}, \mathbf{P}) \tag{2.6}$$

O modelo foi otimizado utilizando o algoritmo Adam, proposto por Kingma e Ba (2015), que é uma variação do *Stochastic Gradien Descent* (SGD). Nesse algoritmo, o *momentum* é calculado para cada uma das variáveis na otimização, ao invés de utilizar apenas um valor fixo. Além disso, esse algoritmo permite mais flexibilidade nos híper-parâmetros visto que ele adapta os valores conforme o aprendizado avança. A otimização foi efetuada utilizando um *learning rate* $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ e *batches* com 256 imagens. Um critério de parada precoce foi utilizado caso o *F1-score* (métrica de avalição para a detecção de AUs) não aumentasse por $\delta = 0,001$ em $m = 5$ épocas. Finalmente, caso esse critério não fosse atingido a otimização encerrava em 100.000 iterações que equivalem a aproximadamente 20 épocas.

2.4 Experimentos

O *Facial Expression Recognition and Analysis Challenge* (FERA) 2017, organizado por Valstar et al. (2017), abordou o tema de detecção e estimativa de intensidade de AUs em imagens de faces em múltiplas poses. Para detectção de AUs, foram disponibilizadas anotações binárias para os AUs 1, 4, 6, 7, 10, 12, 14, 15, 17 e 23, e para estimativa de intensidade foram disponibilizadas anotações em seis níveis (0 a 5) para os AUs 1, 4, 6, 10, 12, 14 e 17. Nesse caso, como não existe uma base com variações de pose, a base BP4D Zhang et al. (2014, 2016) foi estendida com nove poses sintéticas criadas a partir dos modelos 3D presentes na base original. A Figura 2.2 demonstra um exemplo das imagens com as poses sintéticas utilizadas na base. A base é composta por 61 sujeitos, sendo 21 para *Treino*, 20 para *Avaliação*, e 20 para *Teste*. Apenas os 41 sujeitos de *Treino* e *Avaliação* estão disponíveis, os 20 sujeitos de *Teste* ficam com os organizadores e a avaliação é feita pelos próprios organizadores que retornam o resultado obtido. O modelo proposto foi otimizado utilizando apenas as imagens disponíveis para *Treino*; a *Avaliação* foi utilizada apenas para avaliar os resultados localmente e definir qual modelo deveria ser enviado para obter o resultado final no conjunto de *Teste*. As métricas de avaliação de resultados definidas para essa base são o F_1-*score* para a detecção de AUs e o ICC(3, 1), apresentado em Shrout e Fleiss (1979), para a estimativa de intensidade. As duas métricas são comumente utilizadas para avaliar estes cenários e são robustas em relação à resultados aleatórios e desequilíbrio entre classes. Essas propriedades podem ser vistas nos estudos conduzidos por Jeni et al. (2013) e Werner et al. (2015).

2.5 Resultados

Esta seção apresenta os resultados obtidos, e comparativos, utilizando a base do FERA 2017. Os resultados estão divididos em duas sub-seções da seguinte maneira: a Seção 2.5.1

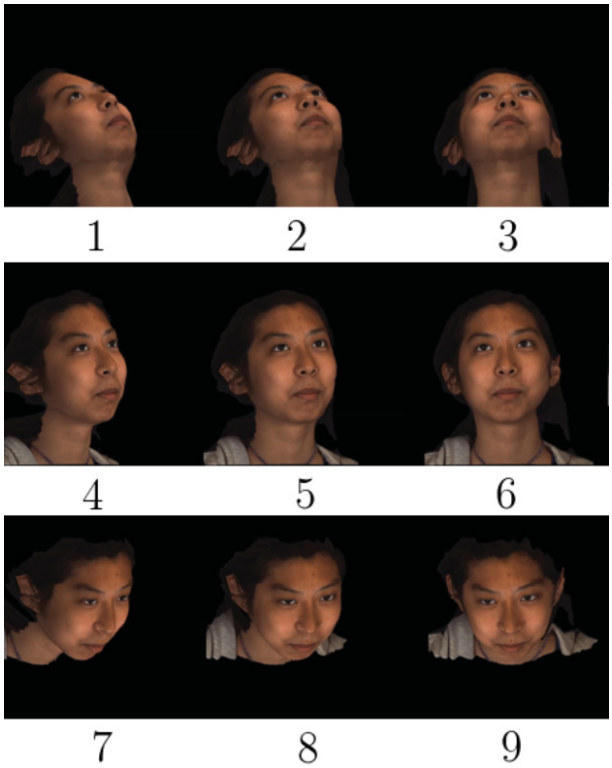


Figura 2.2: Exemplos, e identificadores, das nove poses disponíveis na base BP4D estendida para o FERA 2017. Fonte: Valstar et al. (2017).

descreve e argumenta sobre os resultados da detecção de AUs e a Seção 2.5.2 descreve e argumenta sobre os resultados obtidos para estimativa de intensidade de AUs.

2.5.1 Detecção de AUs

Esta seção apresenta os resultados obtidos para a detecção de AUs utilizando o modelo proposto. Os resultados gerais na *Avaliação* e *Teste* para todos os AUs inclusos na detecção são exibidos na Tabela 2.1. É importante notar que os métodos comparados dos dois conjuntos (*Avaliação* e *Teste*) são diferentes porque a *Avaliação* estava disponível e os comparativos foram utilizados para avaliar o desempenho do método proposto antes do envio para obter os resultados finais. O *Teste*, por outro lado, inclui o comparativo com os participantes do FERA 2017 que participaram na detecção de AUs.

A partir da Tabela 2.1 é possível perceber que o método proposto conseguiu ficar acima da média dos *baselines* na *Avaliação*. Observando os resultados para cada AU, Zhao et al. (2016) ficou acima no resultado individual apenas para os AUs 6 e 12. Observando os resultados finais no *Teste*, é possível perceber que o método proposto por Tang et al. (2017) ficou acima da média em relação a todos os métodos propostos. Entretanto, este método requer o treino e *fine-tuning* para cada AU, dessa forma tornando o método custoso em questão de treino e também de avaliação. Na sequência ficou o método proposto por He et al. (2017) que faz uso de informações temporais. Em terceiro está a AUMPNet, seguida pelo método proposto por Li et al. (2017) que combina características computadas com CNNs e características manuais. Ao observar os resultados individuais é possível observar que a AUMPNet obteve um resultado superior para os AUs 6 e 7. Ambos os AUs ocorrem na região dos olhos e portanto, o aprendizado por regiões, combinado com a estimativa conjunta dos AUs pode ter levado a um resultado superior para

Tabela 2.1: Resultados calculados com F_1-*score* para detecção de AUs.

		AUs										
		01	04	06	07	10	12	14	15	17	23	Média
Aval.	Valstar et al. (2017)	0.154	0.172	0.564	0.727	0.692	0.647	0.622	0.146	0.224	0.207	0.416
	Baltrušaitis et al. (2016)	0.246	0.216	0.572	0.675	0.666	0.673	0.576	0.237	0.321	0.231	0.441
	Zhao et al. (2016)	0.330	0.220	**0.710**	0.670	0.780	**0.770**	0.650	0.250	0.300	0.140	0.480
	AUMPNet	**0.345**	**0.278**	0.677	**0.794**	**0.785**	0.762	**0.692**	**0.267**	**0.364**	**0.250**	**0.521**
Teste	Valstar et al. (2017)	0.147	0.044	0.630	0.755	0.758	0.687	0.668	0.220	0.274	0.342	0.452
	Li et al. (2017)	0.215	0.044	0.755	0.805	0.810	0.753	0.750	0.208	0.286	0.356	0.498
	AUMPNet	0.219	0.056	**0.785**	**0.816**	0.838	0.780	0.747	0.145	0.388	0.286	0.506
	He et al. (2017)	0.198	0.043	0.747	0.784	0.816	0.809	0.691	0.208	0.398	0.374	0.507
	Tang et al. (2017)	**0.263**	**0.118**	0.776	0.808	**0.865**	**0.843**	**0.757**	**0.362**	**0.424**	**0.519**	**0.574**

Tabela 2.2: Resultados calculados com F_1-*score* para detecção de AUs em cada pose no conjunto de *Teste*.

AUs	Poses								
	1	2	3	4	5	6	7	8	9
AU 01	0.197	0.196	0.156	0.243	0.212	0.219	0.223	**0.300**	0.260
AU 04	0.070	0.064	**0.071**	0.043	0.039	0.058	0.060	0.065	0.030
AU 06	0.803	0.792	0.745	0.774	0.784	0.796	**0.809**	0.794	0.773
AU 07	0.822	**0.829**	0.823	0.808	0.812	0.822	0.806	0.803	0.817
AU 10	0.850	**0.852**	0.819	0.843	0.839	0.842	**0.852**	0.824	0.821
AU 12	0.764	0.779	0.766	0.780	0.784	**0.794**	0.791	0.788	0.776
AU 14	0.727	0.766	0.753	0.757	0.758	**0.767**	0.716	0.721	0.747
AU 15	0.093	0.116	0.117	**0.192**	0.132	0.164	0.159	0.169	0.147
AU 17	0.393	**0.433**	0.424	0.405	0.412	0.380	0.357	0.340	0.359
AU 23	0.168	0.203	0.271	0.304	0.316	0.311	0.291	0.324	**0.342**
Média	0.489	0.503	0.495	**0.515**	0.509	**0.515**	0.506	0.513	0.507

estes AUs. Outro ponto que pode ser observado é que os resultados da AUMPNet são similares entre os conjunto de *Avaliação* e *Teste*, exceto para o AU04. O mesmo efeito pode ser observado no *baseline* de Valstar et al. (2017). Como a principal característica desse AU é a textura gerada pelas rugas entre as sobrancelhas, o modelo das regiões utilizado pode ter influencia no efeito. Como essa região é pequena, o corte fixo das regiões pode ter separado essa região. Dessa forma, seria necessário um modelo com recortes em regiões específicas, possivelmente utilizando *landmarks* para obter regiões mais robustas. Além dos resultados gerais para os AUs, também é necessário investigar os resultados para cada uma das nove poses disponíveis. Esses resultados estão disponíveis, utilizando o conjunto de *Teste*, na Tabela 2.2.

Conforme é possível perceber a partir da Tabela 2.2, diferentes AUs obtiveram os melhores resultados em diferentes poses. Sendo que a pose frontal (6) e a pose lateral (4) obtiveram os melhores resultados médios. Ao analisar os resultados individuais, a pose 2 obteve os melhores resultados para os AUs 7, 10 e 17. Os AUs 10 e 17 estão presentes na região da boca, que fica em evidência nessa pose. Da mesma forma, o movimento da pálpebra inferior do olho (AU 7) fica mais evidente nesta pose. Outro ponto importante a ser notado é que a visão quase frontal, pose 5, não obteve nenhum resultado superior para ao menos um AU. Como esperado, os melhores resultados para os AU12 e AU14 são na pose frontal, visto que esses AUs ocorrem na região da boca. Em contrapartida, o AU23 que também ocorre na região da boca o melhor

Tabela 2.3: Resultados calculados utilizando ICC(3, 1) para estimativa de intensidade de AUs.

		AUs							
		01	04	06	10	12	14	17	Média
Aval.	Valstar et al. (2017)	0.082	0.069	0.429	0.434	0.540	0.259	0.005	0.260
	Baltrušaitis et al. (2016)	0.239	0.095	0.420	0.508	0.540	0.250	0.193	0.321
	AUMPNet	**0.381**	**0.227**	**0.658**	**0.685**	**0.755**	**0.458**	**0.327**	**0.499**
Teste	Valstar et al. (2017)	0.035	-0.004	0.461	0.451	0.518	0.037	0.020	0.217
	Amirian e Schwenker (2017)	0.169	0.021	0.509	0.590	0.615	-0.027	0.190	0.295
	AUMPNet	0.228	0.057	**0.702**	0.710	0.732	0.104	0.260	0.399
	Zhou et al. (2017)	**0.307**	**0.147**	0.671	**0.735**	**0.793**	**0.147**	**0.319**	**0.446**

resultado na pose 9. Da mesma forma, o AU17 obteve o melhor resultado na pose 2. Esses resultados indicam que outros efeitos, como a representação conjunta, auxiliaram nas estimativas assim como a possibilidade de *over-fitting* em algumas poses.

2.5.2 Estimativa de intensidade de AUs

Esta seção apresenta os resultados obtidos para a estimativa de intensidade de AUs. Da mesma forma que a Detecção de AUs (Seção 2.5.1), a Tabela 2.3 demonstra os resultados obtidos para *Avaliação* e *Teste*. Também foram utilizados diferentes *baselines* para *Avalição* e para *Teste*, visto que a *Avaliação* visa decidir se o desempenho estava de acordo para *Teste*. O comparativo no *Teste* consta com os participantes do FERA 2017 para estimativa de intensidade.

Os resultados da Tabela 2.3 demonstram que o método proposto ficou com média acima dos *baselines* na *Avaliação* e também nos resultados individuais para cada AU. Entretanto, no *Teste*, o método proposto por Zhou et al. (2017) obteve o melhor resultado. Apesar do modelo proposto por Zhou et al. (2017) também apresentar a estimativa de pose em um passo intermediário, existem modelos específicos para cada pose. Além disso, este modelo foi pré-treinado utilizando a ImageNet, proposta por Simon et al. (2016). O método proposto contém o segundo melhor resultado, seguido por Amirian e Schwenker (2017) que propôs um método de regressão baseado em *Support Vector Regression* (SVR). Outro ponto que é importante ser notado, nos resultados para detecção de AUs e estimativa de intensidade, é que o modelo proposto foi treinado apenas com a separação de *Treino*, sem utilizar a *Avaliação* quando foi avaliado no *Teste*. Esse detalhe com certeza teve influência nos resultados visto que um conjunto maior de sujeitos poderia levar a uma melhor generalização. Também é possível observar que o modelo proposto obteve o melhor resultado para o AU06. Esse resultado pode estar relacionado a representação conjunta dos AUs durante a regressão da intensidade. Em comparação com o método de Zhou et al. (2017), a diferença entre os resultados do AU06 e AU12 são menores na AUMPNet. Como esses AUs são correlacionados, a representação conjunta deve ter mantido as intensidades similares, asssim gerando melhores resultados. Finalmente, os resultados individuais entre poses no *Teste* são apresentados na Tabela 2.4.

A partir dos resultados apresentados na Tabela 2.4, é possível perceber que, diferentemente da detecção de AUs, a pose frontal (6) obteve o melhor resultado médio. Além disso, ao observar os resultados individuais, também fica claro que os AUs 1, 4, 10, 12 obtiveram os melhores resultados nessa pose. Novamente, os AUs da parte inferior do rosto obtiveram melhores resultados em poses que deixam a região da boca em evidência. Entretanto, o melhor resultado para o AU 6 foi na pose 2 (olhando para cima) enquando que para a detecção o melhor resultado foi na pose 7 (olhando para baixo). Esse comportamento pode ser um indicativo da

Tabela 2.4: Resultados calculados com ICC(3, 1) para estimativa de intensidade em cada pose no conjunto de *Teste*.

AUs	Poses								
	1	2	3	4	5	6	7	8	9
AU 01	0.173	0.209	0.182	0.263	0.275	**0.311**	0.169	0.254	0.232
AU 04	0.054	0.080	**0.127**	-0.001	0.029	0.098	0.044	0.042	-0.004
AU 06	0.723	**0.746**	0.710	0.684	0.705	0.721	0.712	0.694	0.650
AU 10	0.690	0.720	0.697	0.714	0.720	**0.741**	0.723	0.709	0.702
AU 12	0.728	0.740	0.740	0.693	0.748	**0.754**	0.737	0.734	0.732
AU 14	0.116	0.092	**0.139**	0.078	0.099	0.127	0.108	0.077	0.107
AU 17	0.248	0.301	**0.327**	0.296	0.314	0.252	0.238	0.204	0.199
Média	0.390	0.413	0.417	0.390	0.413	**0.429**	0.390	0.388	0.374

relação com o AU 12, visto que os resultados para o AU 12 foram melhores em poses olhando para baixo na estimativa de intensidade, enquanto que, para a detecção de AUs os melhores resultados foram em poses olhando para baixo. Da mesma forma que na detecção de AUs, o AU17 obteve o melhor resultado em uma pose não-frontal. Ainda em comparação com a detecção, o AU04 resultou em uma correlação negativa para as poses 4 e 9, que são similares os resultados baixos nas mesmas poses para detecção. Esse comportamento indica a relação da representação conjunta das duas tarefas, visto que a mesma representação é utilizada nos dois cenários, apenas alterando na última camada antes da saída final. Outro ponto importante que consta nos resultados do AU 4 que são baixos tanto para detecção de AUs quanto para estimativa de intensidade. Esse comportamento pode estar relacionado ao desequilíbrio entre as intensidades dos AUs conforme mostra a Figura 2.3. Observando a figura e os resultados obtidos na Tabela 2.4 é possível perceber que os AUs que apresentam uma melhor distribuição entre as intensidades obtiveram os melhores resultados. Para validar a hipótese da influência da distribuição entre os níveis de intensidade, a Tabela 2.5 demonstra a matriz de confusão para a estimativa de intensidade no conjunto de *Avaliação*.

A matriz de confusão demonstrada na Tabela 2.5 apresenta os resultados para cada nível de intensidade. Conforme pode ser visualizado, as intensidades mais frequentes na Figura 2.3 são as que apresentam os melhores resultados, especialmente o nível 0. Além disso, também é possível notar que 3% das anotações que correspondem ao nível 5 foram estimadas como nível 0, indicando um viés do modelo para as classes com maior quantidade de anotações. De forma geral, as estimativas seguiram a tendência das anotações que pode ser visto pela diagonal principal da matriz. É possível perceber que muitas estimativas de intensidade 3 foram para anotações de intensidades com nível 4 ou 5. Da mesma forma, estimativas de nível 4 foram comuns em anotações de nível 5. As intensidades com mais estimativas corretas foram as intensidade 0, 2 e 3. Como pode ser observado na Figura 2.3, esses são os níveis de intensidade mais comuns. Um resultado qualitativo que é visível a partir da matriz de confusão é a intensidade 1. Como pode ser visto na Figura 1.2, a diferença entre o nível 0, 1 e 2 é muito sutil, e mesmo assim o modelo obteve um bom resultado neste nível. Finalmente é possível perceber ainda na intensidade 1 que os principais erros foram para as intensidades 0 e 2.

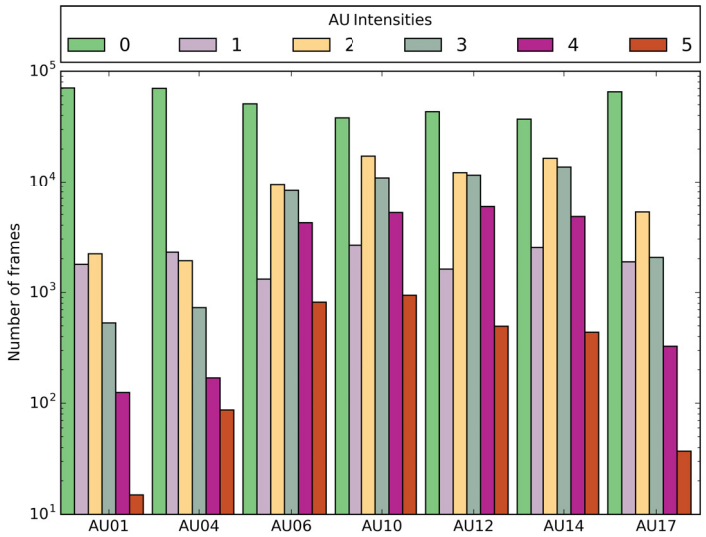


Figura 2.3: Distribuição de intensidades de AUs na base do FERA 2017. Fonte: Batista et al. (2017).

Tabela 2.5: Matriz de confusão dos níveis de intensidade de conjunto de *Avaliação*.

Anotações \ Estimativas	0	1	2	3	4	5
0	67.89	21.05	8.94	1.91	0.21	0.01
1	38.52	33.43	22.98	4.75	0.30	0.01
2	13.37	28.91	36.29	19.41	2.00	0.02
3	4.86	17.39	29.68	37.01	10.49	0.57
4	2.60	7.45	17.48	41.93	26.32	4.22
5	3.02	4.67	9.41	51.16	29.53	2.20

Capítulo 3

Modelo de regressão para lidar com o desequilíbrio entre classes

Como demonstrado na Seção 2.5, o desequilíbrio na quantidade de anotações foi um fator determinate nos resultados para estimativa de intensidade. Além desse ponto, também ficou evidente que a regressão múltipla, utilizando uma representação única para estimar a intensidade de AUs obteve bons resultados. Com isso, esse capítulo apresenta um modelo de regressão múltipla para tratar o desequilíbrio entre os níveis de intensidade dos AUs. Como o foco da abordagem é a estimativa da intensidade e o desequilibrio nas anotações, esse capítulo utiliza uma organização de dados e estrutura de experimentos diferentes do Capítulo 2. Portanto, este capítulo está organizado da seguinte maneira: a Seção 3.1 demonstra a organização dos dados utilizados para otimização e avaliação do modelo, Seção 3.2 apresenta o modelo proposto e como foi feita a otimização. Na sequência, a Seção 3.3 descreve a organização dos experimentos e protocolos de avaliação. Finalmente, a Seção 3.4 apresenta e discute os resultados obtidos.

3.1 Organização das imagens e anotações

Dado que $\mathcal{D} = \{\mathbf{X}, \mathbf{Y}, \mathbf{P}\}$ seja uma base de imagens com anotações de AUs utilizada para treinamento e avaliação. A matriz $\mathbf{X} \in \mathbb{R}^{M \times 224 \times 224 \times 3}$ contém M imagens RGB de faces normalizadas para $\mu = 0$ e $\sigma = 1$. $\mathbf{Y}$ é uma matriz de $M \times Q$ contendo a intensidade de Q AUs em uma imagem. Nesse caso, $\mathbf{Y}_{ik} \in \{0, 1, 2, 3, 4, 5\}$ representa a intensidade do k-ésimo AU na i-ésima imagem. O vetor $\mathbf{P} \in \mathbb{R}^M$ contém as probabilidades para amostragem de uma determinada imagem durante a otimização. Portanto, $0 \leq \mathbf{P}_i \leq 1$ sendo que $\mathbf{P}_i$ é a probabilidade das anotações $\mathbf{Y}_i$ ocorrerem. O cálculo do vetor ocorre conforme a Equação 3.1 que representa a probabilidade conjunta de um vetor de intensidades.

$$\mathbf{P}_i = \prod_{j=1}^{Q} P(i,j) = P(i,1) \cdot P(i,2) \cdot \ldots \cdot P(i,Q-1) \cdot P(i,Q) \tag{3.1}$$

A Equação 3.2 demonstra o cálculo da probabilidade $P(i,j)$. Essa probabilidade é calculada como o complemento da ocorrência da intensidade $\mathbf{Y}_{ij}$ no conjunto de treino. Com isso, intensidades mais frequentes terão uma probabilidade menor do que as intensidades menos frequentes. Dado que cada imagem possui uma probabilidade de ocorrência calculada pelas anotações, é possível usar um método de amostragem para selecionar imagens durante a otimização. Com isso, imagens que possuem uma probabilidade maior podem ser repetidas

Figura 3.1: Exemplos de faces da base utilizadas para treino e avaliação.

várias vezes durante uma época enquanto que imagens com menor probabilidade podem ser utilizadas uma vez, ou até nenhuma. Logo, há um método de amostragem que considera as múltiplas anotações de uma imagem ao invés de anotações isoladas. Finalmente, como as imagens com anotações menos frequentes tem chance de serem repetidas, essa abordagem é considerada como um *over sampling* de imagens que acaba diminuindo o desequilíbrio entre as classes durante a otimização.

$$P(i,j) = 1 - \frac{\sum_{k=1}^{M} [\mathbf{Y}_{kj} = \mathbf{Y}_{ij}]}{M} \tag{3.2}$$

3.1.1 Detecção de faces

As faces foram previamente detectadas da mesma forma como na Seção 2.1.1. Entretanto, os *bounding boxes* não foram transformados em quadrados, mas as imagens foram redimensionadas para que o menor lado tenha o tamanho de 224 *pixels*. Com isso, durante a otimização é feito um recorte aleatório para que a imagem fique no tamanho de 224 x 224. Esse recorte aleatório só ocorre durante a otimização, como uma forma de *augmentation*. Durante a inferência é feito um corte central de 224 x 224. A Figura 3.1 demonstra um exemplo de imagens de faces utilizadas para treino/avaliação com o tamanho de 224 pixels para o menor lado.

3.2 Otimização do modelo

Como o objetivo é a estimativa de intensidades de AUs com o uso de regressão múltipla, o modelo utilizado não é a AUMPNet, demonstrada no Capítulo 2. Para validar a abordagem proposta o modelo utilizado é uma adaptação da VGG16, proposta por Simonyan e Zisserman (2015). Entretanto, qualquer modelo poderia ser utilizado como base visto que as abordagens propostas não alteram a arquitetura do modelo. Originalmente, a VGG16 foi utilizada para classificar imagens na base Imagenet. Portanto, a saída do modelo é de 1.000 elementos contendo as probabilidades para as classes na ImageNet. Neste trabalho, a arquitetura da VGG16 foi alterada para que a última camada seja de Q elementos, com a intensidade dos AUs, e não 1.000 classes. Além disso, uma camada de *dropout* foi adicionada entre o último vetor de representação, com 4.096 elementos, e a saída com as intensidades dos AUs. A saída do modelo é linear, sem

nenhuma função de ativação para restringir as saídas no intervalo [0, 5]. A saída contínua e sem restrição é utilizada na otimização do modelo. Durante a inferência, essa saída é convertida em um dos níveis de intensidade utilizando a transformação demonstrada na Equação 3.3.

$$\hat{\mathbf{Z}}_{ij} = \max\{0, \min\{5, [\frac{1}{2}\hat{\mathbf{Y}}_{ij}]\}\} \tag{3.3}$$

A equação 3.3 demonstra a restrição das saídas do modelo, $\hat{\mathbf{Y}} \in \mathbb{R}^{Q}$, no intervalo [0, 5]. Inicialmente, as estimativas do modelo são reduzidas pela metade, devido a *loss function* utilizada, demonstrada na equação 3.4. Após essa etapa, as estimativas são convertidas de valores contínuos para inteiros utilizando o arredondamento para o inteiro mais próximo, representado por $[\cdot]$. Finalmente, *max* restringe valores negativos para 0 e *min* limita o valor máximo como 5.

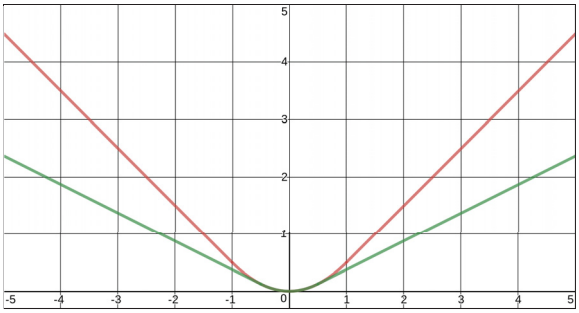
O modelo foi otimizado utilizando SGD, a partir de um pré-treino na ImageNet, com a *loss function* apresentada na equação 3.4.

$$\mathcal{L}(\hat{\mathbf{Y}}, \mathbf{Y}) = \mathcal{H}(\hat{\mathbf{Y}} - 2 \cdot \mathbf{Y}) \tag{3.4}$$

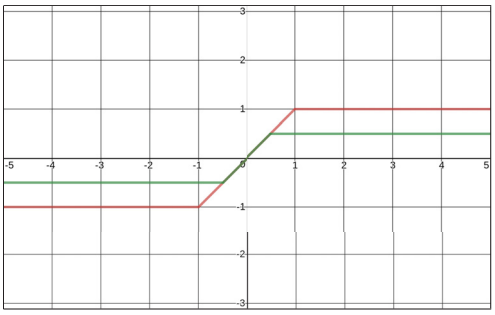
$$\mathcal{H}(x) = \begin{cases} \frac{1}{2}x^2 & \text{se} |x| \leq \delta \\ \delta(|x| - \frac{1}{2}\delta) & \text{demais casos} \end{cases} \tag{3.5}$$

Na *loss function* apresentada na equação 3.4, $\mathcal{H}(\cdot)$ representa a *Smooth L1 loss* descrita por Girshick (2015) e demonstrada na Equação 3.5. A entrada dessa função é o erro calculado entre a saída do modelo ($\hat{\mathbf{Y}}$) e as anotações ($\mathbf{Y}$). As anotações são multiplicadas por 2 devido a utilização do arredondamento na inferência (Equação 3.3). Por exemplo, dado uma anotação y e uma saída do modelo $\hat{y}$, a inferência $\hat{y}$ estará correta se $y - 0,5 \leq \hat{y} < y + 0,5$, ou seja $[\hat{y}] = y$. Entretanto, conforme Girshick (2015), $\mathcal{H}$ assume uma margem $\delta = 1$. Assim, $y \pm 1$ está dentro da margem de inferência correta, mas como visto no exemplo anterior essa propriedade não é válida e deveria ser $y \pm 0,5$. Uma forma de lidar com esse detalhe é $\delta = 0,5$ em $\mathcal{H}$. Porém, como pode ser visto na Figura 3.2, ao alterar a margem para $\delta = 0,5$ algumas propriedades da *loss* também são alteradas. Uma propriedade que pode ser vista na Figura 3.2(a) é que o erro quando $\delta = 0,5$ é a metade de quando $\delta = 1$. Esse ponto fica evidente ao observar o gráfico da derivada de $\mathcal{H}$ na Figura 3.2(b). Como pode ser observado nessa figura, a derivada também é reduzida pela pela metade quando $\delta = 0,5$. Visto que o modelo é otimizado utilizando SGD, um algoritmo que ajusta os pesos do modelo pelo gradiente (derivada) da *loss function*, ter uma derivada com valor menor que 1 pode fazer com que as camadas iniciais do modelo não sejam otimizadas. Essa propriedade contém o mesmo motivo para utilização de ReLU, conforme demonstrado em Krizhevsky et al. (2012). Portanto, uma forma de evitar utilizar $\delta = 0,5$ em $\mathcal{H}$ é multiplicar as anotações por 2. Assim, amplia-se a margem para 1 como esperado em $\mathcal{H}$ quando $\delta = 1$. Finalmente, durante a inferência é necessário ajustar a saída $\hat{\mathbf{Y}}$ para a escala de intensidade com a multiplicação por $\frac{1}{2}$ conforme a Equação 3.3.

A diferença na *loss function* utilizada neste modelo, para a descrita na Seção 2.3 é que ela é mais robusta em relação à pontos fora da curva. A função $\mathcal{H}$ foi proposta por Huber (1964) em uma categoria denominada *M-estimators*. Conforme Huber (1964), a minimização da função definida na Equação 2.3 tende a média das observações. Essa propriedade somente é válida quando a distribuição é Gaussiana. Entretanto, essa propriedade não pode ser verificada porque a



(a) Gráfico de $\mathcal{H}$ com $\delta = 1$ em vermelho e com $\delta = 0,5$ em verde.



(b) Gráfico de $\frac{\partial \mathcal{H}}{\partial x}$ com $\delta = 1$ em vermelho e com $\delta = 0,5$ em verde.

Figura 3.2: Gráficos de $\mathcal{H}$ e $\frac{\partial \mathcal{H}}{\partial x}$ ao alterar δ.

representação utilizada na regressão é otimizada em conjunto com o modelo. Portanto, a *loss function* definida na Equação 3.5 é mais indicada do que a definida na Equação 2.3.

Para a otimização do modelo são utilizadas 41.328 imagens, quantidade total de imagens da base removendo detecções erradas de face e anotações desconhecidas conforme descrito na Seção 3.3.2. Porém, essas imagens são amostradas utilizando uma distribuição multinomial utilizando as probabilidades em **P**, conforme a Seção 3.1. Assim, é possível que imagens com baixa probabilidade nunca sejam utilizadas e imagens com maior probabilidade sejam utilizadas mais de uma vez. Como a probabilidade em **P** é inversa a frequência com que as anotações ocorrem na base, anotações muito frequentes na base serão pouco usadas enquanto que anotações pouco frequentes serão utilizadas mais de uma vez. Portanto, esse método de amostragem tende a gerar *batches* mais homogêneos em relação aos níveis de intensidade utilizados em cada iteração, facilitando a otimização.

Por fim, antes da última camada do modelo proposto existe uma camada de *dropout* com probabilidade de 0.6 para desativar uma determinada unidade. Essa camada foi adicionada para diminuir a quantidade de unidades utilizadas para gerar uma intensidade sem a necessidade de uma nova camada totalmente conectada. Ao analisar o vetor de representações com 4.096 elementos gerado pela VGG16, utilizando *Principal Component Analysis* (PCA), notou-se que era possível manter uma variância de 95% com aproximadamente 2.300 elementos. Portanto, se apenas 40% dos 4.096 elementos da representação forem utilizados, a representação diminui para 1.639 elementos. Assim, ao utilizar *dropout*, a representação é reduzida enquanto que o modelo é regularizado. A outra abordagem seria a adição de uma camada totalmente conectada com 4.096 elementos de entrada e 2.300 elementos de saída. Porém, ao adicionar essa camada, o modelo teria mais $4.096 \cdot 2.300 = 9.420.800$ parâmetros para otimizar. Ou seja, o uso de *dropout* tem uma função similar na redução da representação sem o custo da adicção de parâmetros.

3.3 Experimentos

O Capítulo 2 apresentou um modelo unificado para detectar e estimar a intensidade de AUs em imagens de faces não-frontais. Entretanto, como este capítulo visa um modelo para trabalhar com o desequilíbrio existente nos níveis de intensidade dos AUs, a base utilizada para experimentos foi alterada. Assim, esta seção descreve o modelo *baseline*, a base de imagens e o protocolo utilizados na avaliação do modelo proposto. Além do protocolo padrão da base, todos os modelos foram otimizados e testados três vezes de forma independente. Assim, os resultados apresentados são a média das cinco melhores épocas de cada teste. Essa metodologia

foi adotada para diminuir o efeito da aleatoriedade durante a otimização e obter um resultado mais consistente.

3.3.1 *Baseline*

O modelo *baseline* consiste na VGG16 sem nenhum tipo de pré-treino. As modificações consistem na alteração da última camada para que a saída seja de Q unidades sem restrição para gerar as estimativas de intensidade e a função utilizada na otimização, que é a mesma descrita na Equação 3.4 sem a multiplicação por 2. Assim, esse modelo é otimizado assumindo que a alteração na margem não influencia o resultado. Para inferência, foi utilizada a mesma abordagem apresentada na Equação 3.3 sem o ajuste de $\frac{1}{2}$. Os demais híperparâmetros consistem em: *learning rate* (α) inicial de 0,01, *momentum* de 0,9 e *weight decay* de 0,0005. Esse modelo foi otimizado em cem épocas com o α sendo atualizado para $\alpha := \frac{\alpha}{10}$ a cada vinte épocas. A única diferença dos híperparâmetros do *baseline* para o modelo proposto é o α inicial que é 0,001. Os demais valores foram mantidos para ambos os modelos.

3.3.2 Base de imagens e protocolo de avaliação

Os experimentos foram conduzidos com o subconjunto da BP4D, introduzida por Zhang et al. (2014, 2016), utilizado no *Facial Expression Recognition and Analysis Challenge* (FERA) 2015, proposto por Valstar et al. (2015). Essa base é similar a do FERA 2017, descrita na Seção 2.4, com a principal diferença sendo que as imagens são sempre frontais, sem a mudança na pose. As imagens são em alta resolução com 21 sujeitos para treino e de 20 sujeitos para avaliação. No total, existem anotações de intensidade variando entre 0 e 5 para cinco AUs: AU06, AU10, AU12, AU14 e AU17. Entretanto, nem todas as imagens estão anotadas com todos os AUs. Em algumas imagens os AUs são anotados com um valor 9 indicando que a intensidade é desconhecida. As imagens que possuem algum AU com essa anotação foram removidas do conjunto de treino visto que a abordagem proposta utiliza regressão múltipla e requer todos os AUs anotados para uma imagem. Ao remover as imagens com anotações desconhecidas, o conjunto de treino resultante ficou com 41.328 imagens. No conjunto de avaliação existe o mesmo caso com anotações desconhecidas, porém, nesse caso, as anotações desconhecidas foram desconsiderados ao calcular o resultado final. Finalmente, os modelos são avaliados utilizando o ICC(3, 1), que é a medida utilizada no FERA 2015 para avaliar os modelos na estimativa de intensidade de AUs.

Assim como na base do FERA 2017, o FERA 2015 possui desequilíbrio entre os níveis de intensidade dos AUs. A Figura 3.3 demonstra a quantidade de anotações para cada intensidade de cada AU nessa base. Como é possível perceber, a intensidade 0 é predominante para todos os AUs. Da mesma forma, as intensidades mais altas são menos frequentes. Um ponto que fica evidente é pouca quantidade de amostras para a intensidade 5 do AU 17. Finalmente, a intensidade 1 é bem representada, mas possui menos imagens que as intensidade 2 e 3.

3.4 Resultados

O comparativo entre o modelo proposto, o *baseline* e o estado-da-arte é apresentado na Tabela 3.1. A partir dessa tabela é possível perceber que o modelo proposto superou o *baseline* em três, de cinco, AUs e também no resultado médio. Além disso, o modelo proposto superou o estado-da-arte para o AU10 e obteve um resultado similar para o AU17. Em relação ao *baseline*, o resultado para o AU06 é superior ao do modelo proposto, e também ao estado da arte, e

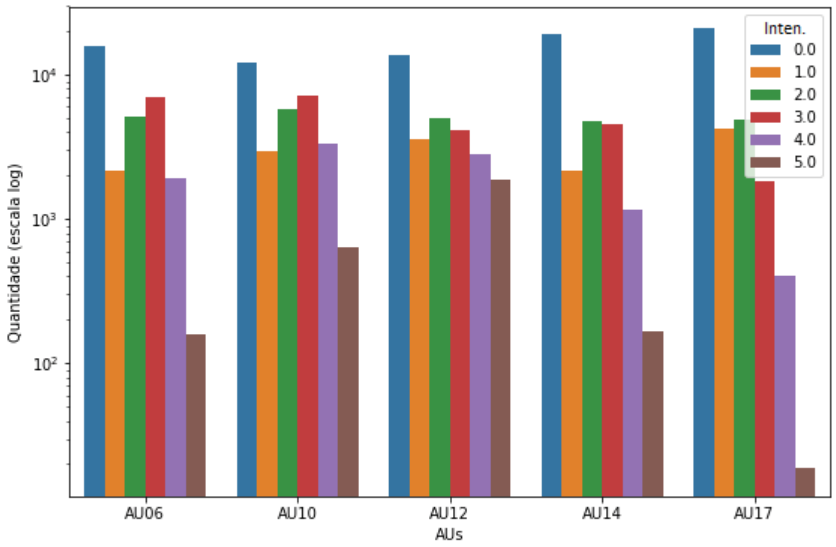


Figura 3.3: Quantidade de imagens para cada nível de intensidade de cada AU no conjunto de avaliação.

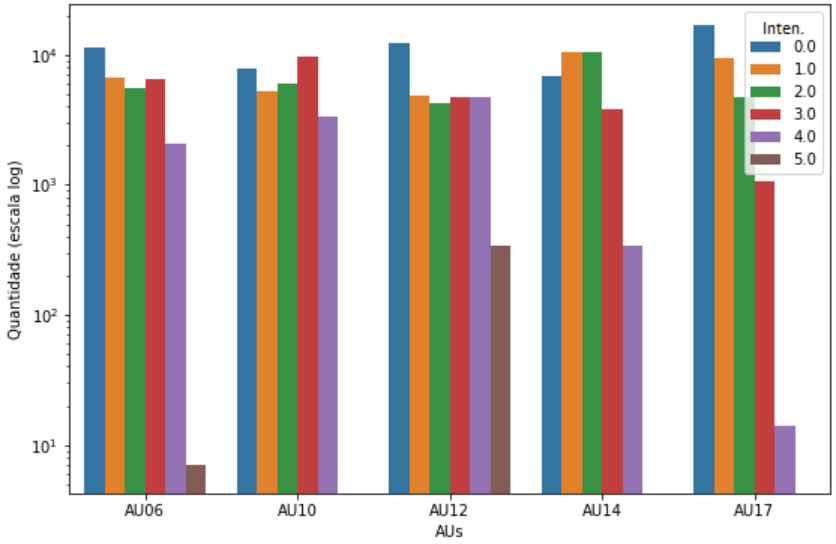
Tabela 3.1: Comparativo para estimativa de intensidade entre AUs utilizando ICC(3, 1). Os resultados com * são valores iguais porque Walecki et al. (2017) usou duas casas decimais.

Action Units	Modelo			
	Baseline	Proposto	Walecki et al. (2017)	VGG16 Walecki et al. (2017)
AU 06	**0,770**	0,734	0,750	0,630
AU 10	0,679	**0,704**	0,690	0,610
AU 12	0,783	0,824	**0,860**	0,730
AU 14	0,385	0,385	**0,400**	0,250
AU 17	0,433	**0,449***	**0,450***	0,310
Média	0,610	0,619	**0,630**	0,510

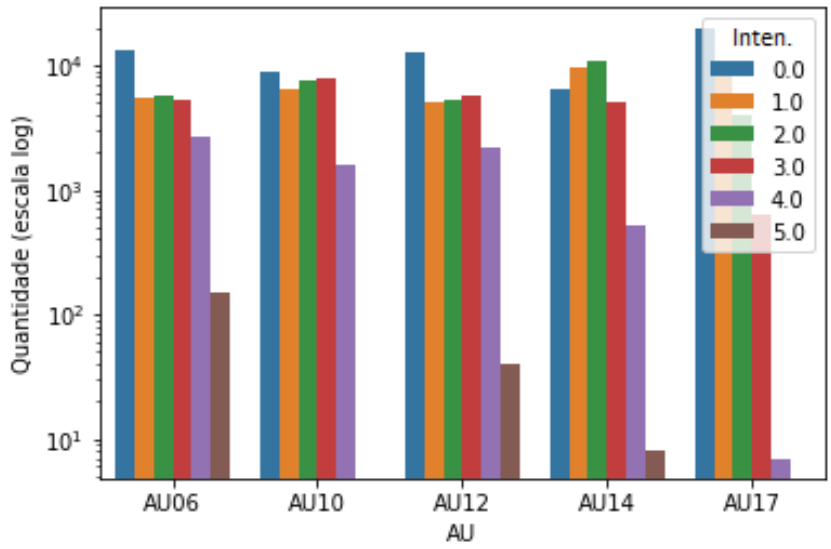
o AU14 obteve um resultado similar ao modelo proposto. Para comparativo entre *baselines*, foi incluído o resultado da VGG16 demonstrado em Walecki et al. (2017). Como é possível notar, o *baseline* proposto, utilizando regressão, superou o *baseline* em Walecki et al. (2017). Esse comparativo indica a melhora resultante no uso de regressão ao invés de classificação na estimativa de intensidade. Devido a natureza ordinal da intensidade dos AUs, faz sentido que a regressão obtenha um resultado melhor que a classificação visto que as saídas são contínuas e estabelecem essa ordem.

O resultado médio reportado na Tabela 2.3 é em relação à melhor iteração em três experimentos independentes que foram executados, mencionado na Seção 3.3.2. Ao considerar as cinco melhores iterações de cada experimento, o *baseline* tem um resultado médio de 0,588 e o modelo proposto tem um resultado médio de 0,603. Portanto, o modelo proposto é mais robusto na variação entre treinamentos gerando um resultado mais homogêneo entre várias iterações e sem sofrer muito com o efeito da aleatoriedade.

A Figura 3.4(a) apresenta a quantidade de imagens estimadas para cada intensidade de cada AU utilizando o modelo proposto. A partir dela é possível perceber que o modelo proposto obteve uma distribuição similar à do conjunto de avaliação demonstrado na Figura 3.3.



(a) Quantidade de imagens inferidas por intensidade com o modelo proposto.



(b) Quantidade de imagens inferidas por intensidade com o *baseline*.

Figura 3.4: Distribuição de intensidades inferidas pelo modelo proposto e pelo *baseline*.

Entretanto, o modelo falhou ao estimar o nível de intensidade 5 para os AU10, AU14 e AU17. Para comparação, a Figura 3.4(b) demonstra a distribuição das estimativas utilizando o *baseline*. De forma geral, os dois modelos resultaram em distribuições similares, com maiores variações nas intensidades mais altas. Por exemplo, enquanto o modelo proposto falhou na intensidade nível 5 para os AU10, AU14 e AU17, o *baseline* falhou no mesmo nível de intensidade para os AU10 e AU17. É importante notar que os dois modelos falharam na estimativa da intensidade 5 para os AU10 e AU17. Para entender melhor os resultados obtidos, a Seção 3.4.1 faz uma análise utilizando matrizes de confusão.

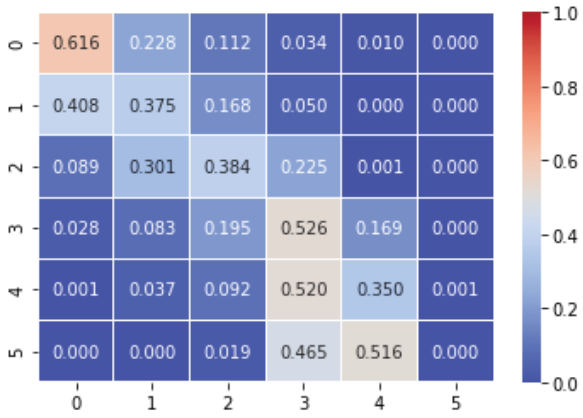
3.4.1 Análise dos resultados para os níveis de intensidade

Ao observar as matrizes de confusão, Figura 3.5, é possível notar que para todos os AUs a intensidade 0 está entre as melhores estimativas. Esse resultado, em conjunto com os erros de estimativas na intensidade 5 indicam que ainda é necessário investigar uma estratégia de amostragem para equilibrar os *batches* utilizados para otimização. Mesmo assim, como pode ser visto na matriz de confusão para o AU12, Figura 3.5(c), é possível verificar que mesmo com erros, as estimativas são próximas dos níveis esperados. Alguns casos excedem esse padrão, como o de algumas imagens que deveriam ter o AU12 com intensidade 0, mas foram estimados com intensidade 5. Como é possível observar nas matrizes de confusão na Figura 3.5, mesmo com a estratégia de amostragem para diminuir o desequilibrio das anotações, a intensidade 5 não obteve bons resultados. De forma geral, apenas o AU14 (Figura 3.5(d)) e o AU17 (Figura 3.5(e)) não obtiveram bons resultados nas intensidades 4 e 5.

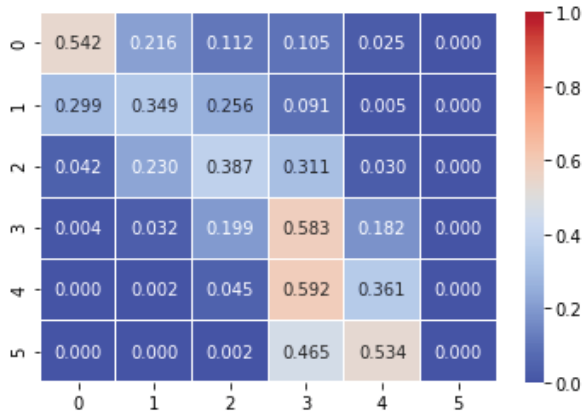
Outro ponto a ser observado nas matrizes de confusão das Figura 3.5(a), Figura 3.5(b) e Figura 3.5(c) é que os erros nas intensidades 3, 4 e 5 foram para intensidades próximas. Esse detalhe indica que a abordagem utilizada para arredondamento das estimativas pode ser melhorada. Ao invés de utilizar um simples arredondamento para o inteiro mais próximo, é possível aplicar algum tipo de transformação mais elaborada. No mais, as divergências nos níveis de intensidade podem estar relacionadas às correlações entre os AUs que o modelo aprendeu durante a otimização que são apresentadas na Figura 3.7 e discutidos na Seção 3.4.2.

Como pode ser observado nas matrizes de confusão dos AU06 (Figura 3.5(a)), AU10 (Figura 3.5(b)) e AU12 (Figura 3.5(c)) as estimativas seguem uma tendência com as anotações pela diagonal principal. Entretanto, para o AU17 (Figura 3.5(e)) e para o AU14 (Figura 3.5(d)) as estimativas ficaram concentradas nas intensidades mais baixas. Esse efeito pode estar relacionado a regressão múltipla, visto que ele também ocorre no *baseline*, demonstrado nas Figura 3.6(e) e Figura 3.6(d), respectivamente. O motivo da regressão múltipla influenciar nestes casos é a correlação entre os AUs, discutida em mais detalhes na Seção 3.4.2. Principalmente, no caso do AU17 que não possui correlação com os demais AUs. Dessa forma, se a representação possui uma evidência mais forte para um determinado AU, pode afetar a estimativa do AU17.

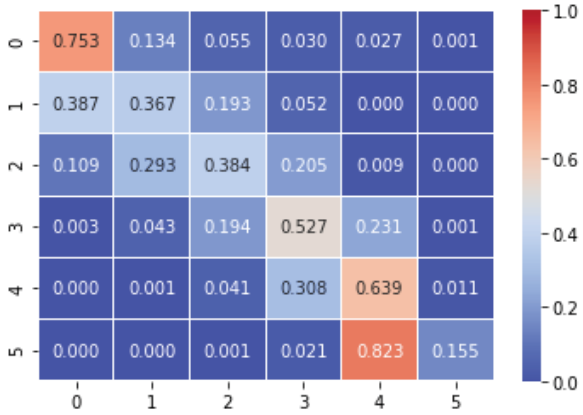
Para comparar a melhora em relação ao *baseline*, a Figura 3.6 apresenta as matrizes de confusão para o *baseline*. De forma geral, os resultados são similares aos do modelo proposto, mas com alguns pontos que ficam evidentes. Por exemplo, o *baseline* inferiu muitas intensidades de nível 2 quando era esperado o nível 5 para o AU17, como pode ser observado na Figura 3.6(e). O modelo proposto, como pode ser visto na Figura 3.5(e), obteve um viés para a intensidade 2, mas foi muito mais atenuado em comparação com o *baseline*. Seguindo na mesma linha, pode ser observado que para os AU10 e AU12, o *baseline* inferiu a intensidade 3 quando o esperado era 5. Em contrapartida, o modelo proposto inferiu a intensidade 4 nesses casos, indicando uma melhora nas estimativas. Outro efeito que pode ser observado nas Figura 3.5, Figura 3.6 e Tabela 2.5 são as estimativas para a intensidade nível 1. Como já foi observado na Tabela 2.5, quando a anotação é intensidade 1 o modelo prevê tanto a intensidade 0 quanto a intensidade 2. Esse efeito está mais relacionado a capacidade do modelo de observar pequenas variações a textura e geometria da face do que na correlação que existe entre os AUs e intensidades. De forma geral, esse erro para níveis de intensidade próximos da anotação se repete para todos os AUs conforme demonstram as Figura 3.5 e Figura 3.5. Finalmente, um ponto que pode ter levado às melhoras entre o modelo proposto e o *baseline* são as correlações entre os AUs que foram exploradas com a diminuição do desequilíbrio entre as classes e esses resultados são analisados na Seção 3.4.2.



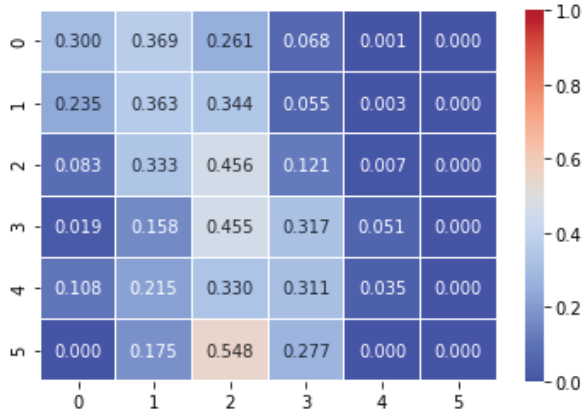
(a) Matriz de confusão do AU06.



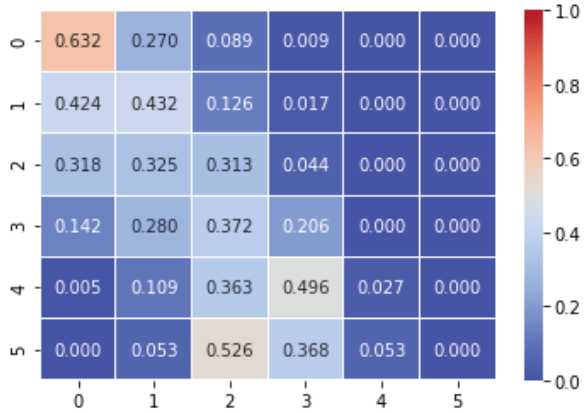
(b) Matriz de confusão do AU10.



(c) Matriz de confusão do AU12.



(d) Matriz de confusão do AU14.

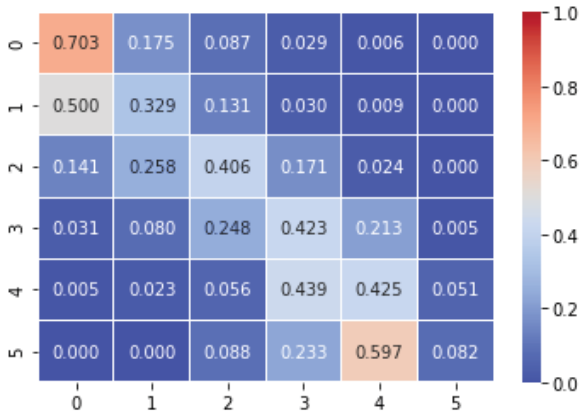


(e) Matriz de confusão do AU17.

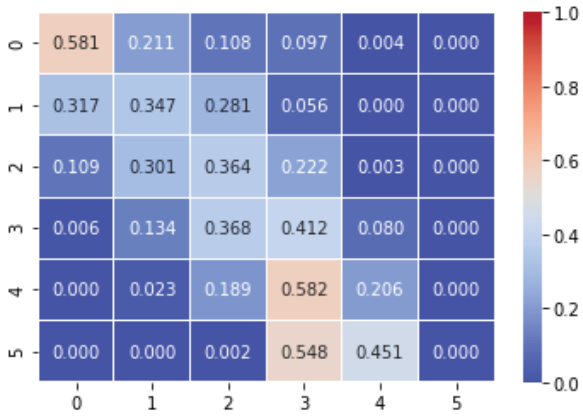
Figura 3.5: Matrizes de confusão para as saídas do modelo proposto.

3.4.2 Análise dos resultados de correlação entre AUs

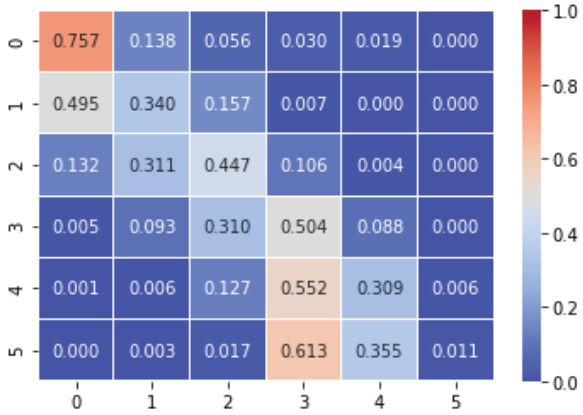
Um dos objetivos de realizar a regressão múltipla para a intensidade dos AUs é explorar a correlação entre os AUs em uma representação intermediária do modelo. A correlação existente entre os AUs no conjunto de avaliação pode ser vista na Figura 3.7(a) e nas estimavas do modelo proposto pode ser vista na Figura 3.7(b). Como é possível perceber, o modelo conseguiu capturar a tendência entre os AUs. Por exemplo, existe uma correlação mais forte entre AU06, AU10, AU12 e AU14 enquanto o AU17 não tem correlação com os demais. Entretanto, apesar das



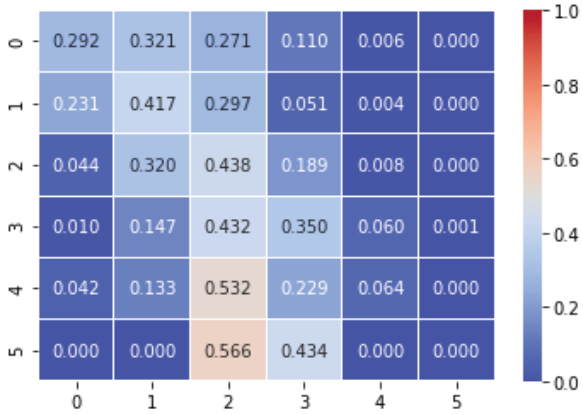
(a) Matriz de confusão do AU06.



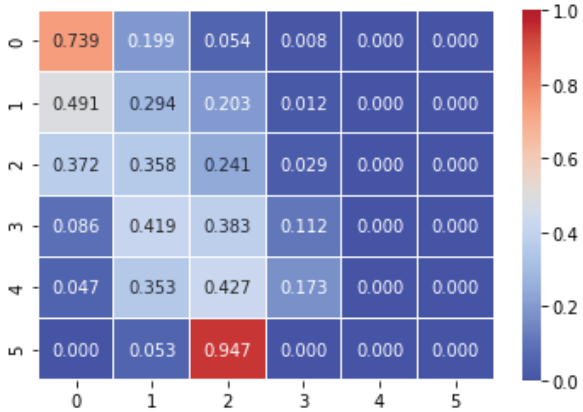
(b) Matriz de confusão do AU10.



(c) Matriz de confusão do AU12.



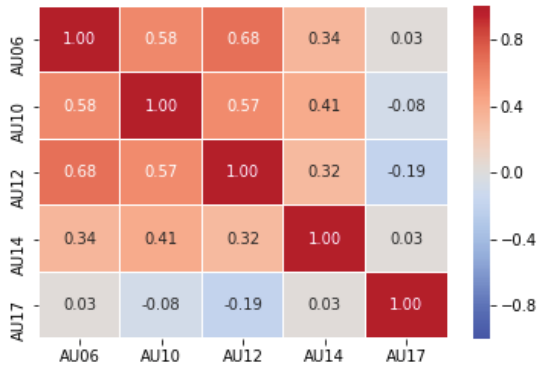
(d) Matriz de confusão do AU14.



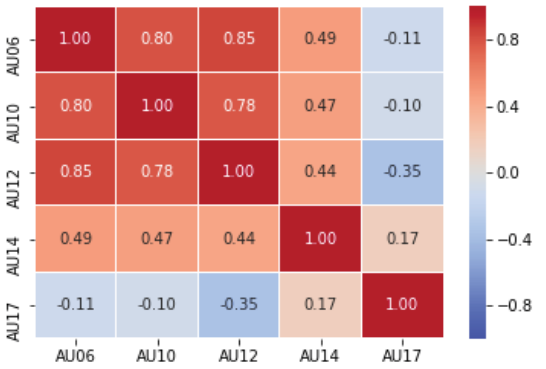
(e) Matriz de confusão do AU17.

Figura 3.6: Matrizes de confusão para as saídas do *baseline*.

estimativas seguirem os mesmos padrões, elas tem correlação maior que as conjunto de avaliação. Portanto, essas correlações fortes entre os AUs que foram geradas pelo modelo podem influenciar os casos da intensidade esperada ser 0, mas o modelo estimar o nível 5. O mesmo vale para a falha do modelo em estimar a intensidade 5 para os AU10, AU14 e AU17. Além das correlações do modelo proposto, a fim de comparação, a Figura 3.7(c) demonstra as correlações entre os AUs utilizando o *baseline*. Como é possível perceber, o próprio *baseline* foi capaz de capturar a tendência existente nas anotações. Entretanto, o modelo proposto obteve correlações mais próximas dos valores esperadas como, por exemplo, entre os AU14 e AU17 onde a correlação nas



(a) Correlações entre os AUs no conjunto de avaliação.



(b) Correlações entre os AUs nas saídas do modelo proposto.

(c) Correlações entre os AUs nas saídas do *baseline* para comparação com o modelo proposto.

Figura 3.7: Mapa de calor indicando as correlações entre os AUs no conjunto de avaliação e nas saídas do modelo proposto.

anotações é 0,03, o modelo proposto obteve 0,17 e o *baseline* obteve 0,38. Porém, ao observar as correlações entre os AU12 e AU17 o *baseline* obteve um resultado mais próximo ao esperado; da mesma forma para a correlação entre os AU06 e AU17. De forma geral, as correlações demonstradas na Figura 3.7 demonstram que o modelo proposto foi capaz capturar a correlação entre as intensidades dos AUs a partir da otimização conjunta com regressão múltipla.

Capítulo 4

Considerações gerais e trabalhos futuros

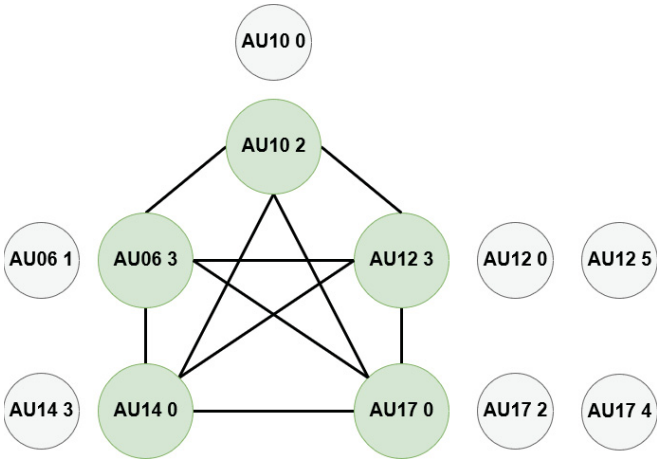
Os Capítulo 2 e Capítulo 3 apresentaram modelos para inferência conjunta de AUs utilizando redes neurais convolucionais. O modelo AUMPNet é um modelo unificado que demonstrou obter bons resultados utilizando a otimização conjunta para pose, detecção de AUs e estimativa de intensidade de AUs. Conforme observados nos resultados, a estimativa de intensidade sofreu com o efeito do desequilíbrio existente entre as níveis de intensidade nas anotações. Portanto, a abordagem descrita no Capítulo 3 visou resolver este problema. Com os resultados demonstrados na Seção 3.4 ficou evidente que as melhorias sugeridas sobre um *baseline* foram eficientes. Entretanto, os resultados também demonstraram que alguns pontos precisam ser refinados.

O primeiro ponto é que o modelo utilizado no Capítulo 3, diferentemente da AUMPNet, utiliza uma representação holística da face. Dessa maneira, toda a imagem é utilizada na estimativa de intensidade de AUs e essa representação pode levar a ruídos no modelo. Uma abordagem melhor seria o uso de regiões conforme a AUMPNet, mas com foco nas regiões onde AUs ocorrem para evitar o efeito de varições na pose. Essa tarefa pode ser trabalhada com o uso de um detector de *landmarks*, como o proposto por Bulat e Tzimiropoulos (2017), em um passo anterior para recortar regiões pré-definidas. Outra abordagem seria um modelo *end-to-end* que faz a detecção de regiões da face estima a intensidade de AUs. Um ponto inicial para esse modelo seria a utilização de técnicas como *Region Proposal Networks*, introduzida em Ren et al. (2015), para gerar regiões candidatas para estimar a intensidade de AUs.

O outro ponto que ficou com possibilidade de melhoria é a estratégia para redução no desequilíbrio entre os níveis de intesidade. A abordagem proposta consiste na amostragem de imagens com base no complemento da probabilidade com que elas ocorrem no conjunto de treino, descrito na Seção 3.1. Mesmo com essa abordagem o modelo ainda sofreu com efeitos do desequilíbrio entre os níveis de intensidade. Uma sugestão de melhoria utilizando a técnica proposta seria atualizar as probabilidades a cada época considerando os erros cometidos durante o treino na época anterior. Aqui a regra de Bayes pode ser utilizada para atualizar a distribuição de probabilidade posterior com base em uma distribuição anterior (probabilidade atual) e uma nova evidência (erros cometidos pelo modelo na época atual). Outro ponto que pode levar a melhorias é a utilização de *curriculum learning*, conforme proposto por Gui et al. (2017). Nesse cenário, pode levar em consideração que as intensidades 0, 3 e 5 são mais fáceis de estimar do que as intensidades 1, 2, e 4. Essa propriedade ficou evidente com os resultados demonstrados na Seção 3.4.1. Entretanto, seria necessário propor uma função de complexidade que considera a intensidade de todos os AUs em uma única imagem.

Por fim, como o modelo estado-da-arte para estimativa de intensidade de AUs utiliza o conceito de aprendizado estruturado (modela a relação entre os AUs como um grafo), um *baseline*

(a) Imagem da BP4D com os AUs e intensidades: AU06 3, AU10 2, AU12 3, AU14 0 e AU17 0.



(b) Grafo que representa a conectividade dos AUs e intensidades da imagem. Em verde estão os AUs e intensidades presentes na imagem e as arestas indicam a adjacência entre eles. Em cinza estão alguns AUs e intensidades que não estão presentes na imagem. Para clareza da imagem, alguns AUs que não estão presentes foram omitidos.

Figura 4.1: Grafo representando os AUs e suas intensidades presentes na imagem.

com essa propriedade foi considerado. Para explorar esse modelo é necessário considerar as anotações de intensidades de AUs na imagem como um grafo. A Figura 4.1 demonstra como essa relação pode ser modelada. Com isso, as anotações podem ser consideradas como uma matriz de adjacência $\mathbf{A} \in \mathbb{R}^{6Q \times 6Q}$ entre Q AUs com seis níveis de intensidade. Cada elemento $\mathbf{A}_{ij} \in \{0,1\}$ indica se existe uma aresta conectando dois determinados em um determinado nível de intensidade. É importante notar que essa matriz é simétrica, portanto $\mathbf{A}_{ij} = \mathbf{A}_{ji}$. Com essa definição, é possível alterar a última camada de um modelo base, como a VGG16, para que a última camada tenha $6Q \cdot 6Q = 36Q^2$ unidades que representam os elementos da matriz de adjacência $\mathbf{A}$. Para que essas saídas representem uma probabilidade de ocorrência de uma arestas entre AUs e intensidades, uma ativação sigmóies é utilizada. Assim, esse modelo pode ser otimizado utilizando a *loss function* descrita na Equação 2.1. Portanto, esse modelo é otimizado para reconstruir um grafo de adjacência com a probabilidade de uma determinada aresta existir no grafo ou não. Como a saída desse modelo contém $36Q^2$ unidades, é necessário aplicar uma transformação para obter os níveis de intensidade estimados. Essa transformação é feita utilizando as Equação 4.1 e Equação 4.2.

$$\mathbf{p}_j = \prod_{i=1}^{6Q} \hat{\mathbf{A}}_{ij} \propto \log(\mathbf{P}_j) = \sum_{i=1}^{6Q} \log(\hat{\mathbf{A}}_{ij}) \tag{4.1}$$

$$\hat{\mathbf{y}}_i = \underset{j \in \{1,\ldots,Q\}}{\operatorname{argmax}} \hat{\mathbf{P}}_{ij} \tag{4.2}$$

A Equação 4.1 recebe a saída do modelo de $36Q^2$ unidades reorganizadas como uma matriz de adjacência $\hat{\mathbf{A}} \in \mathbb{R}^{6Q \times 6Q}$. Com essa matriz, é possível usar o conceito de que as colunas

representam as arestas que chegam em determinados nós do grafo. Portanto, é possível calcular a probabilidade de um determinado nó ocorrer multiplicando as probabilidades das arestas chegarem neste nó, que é representado pela Equação 4.1. Para evitar problemas de instabilidade numérica, visto que são feitas multiplicações de números menores que 1, a Equação 4.1 faz uso da propriedade que a multiplicação é equivalente a soma dos logs. A saída da Equação 4.1 é um vetor de $\mathbf{P} \in \mathbb{R}^{6Q}$ elementos que é reorganizado para uma matriz $\hat{\mathbf{P}} \in \mathbb{R}^{Q \times 6}$ onde cada linha representa um AU da imagem e cada coluna representa uma intensidade. Como o elemento $\hat{\mathbf{Y}}_{ij}$ contém um valor proporcional à uma probabilidade, é possível estimar a intensidade final de cada AU obtendo o índice j com maior valor de cada linha, como demonstra a Equação 4.2. Finalmente, o vetor $\hat{\mathbf{y}}$ contém o nível de intensidade estimado para uma determinada imagem.

Ao otimizar este modelo utilizando os mesmos critérios descritos na Seção 3.3, o melhor resultado foi um ICC(3, 1) de 0,571 e com as cinco melhores épocas em três otimizações, o resultado médio foi de 0,560. Como esses resultados ficaram abaixo do *baseline* proposto com regressão, esse modelo foi desconsiderado nas comparações. Entretanto, essa abordagem pode ser promissora ao investigar melhores modelos para trabalhar com grafos. Esse modelo é mais simples do que o proposto por Walecki et al. (2017) e também considera as relações entre AUs como grafos. Porém, esse modelo não considera a natureza ordinal dos AUs, o que pode ter feito modelo obter resultados abaixo do *baseline* com regressão. Finalmente, essa abordagem com grafos fica como um possível ponto a ser explorado em trabalhos futuros.

Capítulo 5

Conclusão

Este trabalho apresentou um modelo baseado em redes neurais convolucionais para análise de expressões faciais. O modelo proposto consiste de uma arquitetura única para detectar e estimar a intensidade de AUs em imagens com múltiplas poses da cabeça. O modelo unificado é possível através do uso do aprendizado multi-tarefa e da formulação da detecção como rotulação múltipla da imagem, assim como a regressão múltipla para a estimativa de intensidade. Os resultados obtidos com o modelo são comparáveis com o estado da arte, sendo que ao observar alguns AUs isoladamente, os resultados são superiores. Posteriormente, notou-se que o desequilíbrio entre as classes durante a otimização afetou os resultados obtidos. Portanto, também foi apresentada uma técnica para amostragem de imagens para treino considerando o contexto de múltiplas anotações por imagem. O modelo proposto foi capaz de aprender correlações entre os níveis de intensidade a partir de uma representação intermediária. Entretanto, algumas correlações levaram o modelo a gerar estimativas fora do esperado. Mesmo assim, o modelo foi capaz de gerar resultados médios similares ao estado da arte. Ao analisar os resultados para cada AU, o modelo conseguiu melhorar o estado da arte para o AU10 e obteve um resultado similar para o AU17. Em suma, a hipótese de que o uso de regressão é mais apropriado para a estimativa de AUs, dada a natureza ordinal das intensidades, foi validada, visto que o *baseline* com regressão superou a classificação. Por fim, com base nos resultados obtidos, sugestões de trabalhos futuros foram propostas.

Referências

Amirian, M. e Schwenker, F. (2017). Support vector regression of sparse dictionary-based features for view-independent action unit intensity estimation. páginas 854–859.

Baltrušaitis, T., Mahmoud, M. e Robinson, P. (2015). Cross-dataset learning and person-specific normalisation for automatic action unit detection. Em *IEEE Conference on Automatic Face and Gesture Recognition*.

Baltrušaitis, T., Robinson, P., Morency, L.-P. et al. (2016). Openface: an open source facial behavior analysis toolkit. Em *IEEE Workshop on Applications of Computer Vision*.

Batista, J. C., Albiero, V., Bellon, O. R. P. e Silva, L. (2017). Aumpnet: Simultaneous action units detection and intensity estimation on multipose facial images using a single convolutional neural network. Em *IEEE Conference on Automatic Face and Gesture Recognition*, páginas 866–871.

Batista, J. C., Bellon, O. R. e Silva, L. (2016). Landmark-free smile intensity estimation. Em *SIBGRAPI Conference on Graphics, Patterns and Images*.

Benitez-Quiroz, C. F., Srinivasan, R., Feng, Q., Wang, Y. e Martinez, A. M. (2017). Emotionet challenge: Recognition of facial expressions of emotion in the wild.

Benitez-Quiroz, C. F., Srinivasan, R. e Martinez, A. M. (2016). Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. Em *IEEE Conference on Computer Vision and Pattern Recognition*.

Bulat, A. e Tzimiropoulos, G. (2017). How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). *arXiv preprint arXiv:1703.07332*.

Corneanu, C. A., Simón, M. O., Cohn, J. F. e Guerrero, S. E. (2016). Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1548–1568.

Danelljan, M., Häger, G., Khan, F. e Felsberg, M. (2014). Accurate scale estimation for robust visual tracking. Em *British Machine Vision Conference*.

Du, S. e Martinez, A. M. (2015). Compound facial expressions of emotion: from basic research to clinical applications. *Servier Dialogues in Clinical Neuroscience*, 17:443–455.

Du, S., Tao, Y. e Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 2014.

Ekman, P., Friesen, W. e Hager, J. (2002). *Facial Action Coding System (FACS): Manual*. A Human Face.

Girard, J. M., Cohn, J. F. e la Torre", F. D. (2015). Estimating smile intensity: A better way. *Pattern Recognition Letters*.

Girshick, R. (2015). Fast R-CNN. Em *IEEE International Conference on Computer Vision*.

Gudi, A., Tasli, H. E., den Uyl, T. M. e Maroulis, A. (2015). Deep learning based facs action unit occurrence and intensity estimation. Em *IEEE Conference on Automatic Face and Gesture Recognition*.

Gui, L., Baltrušaitis, T. e Morency, L. P. (2017). Curriculum learning for facial expression recognition. Em *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, páginas 505–511.

He, J., Li, D., Yang, B., Cao, S., Sun, B. e Yu, L. (2017). Multi view facial action unit detection based on cnn and blstm-rnn.

He, K., Zhang, X., Ren, S. e Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Em *IEEE International Conference on Computer Vision*.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*.

Jeni, L. A., Cohn, J. F. e De La Torre, F. (2013). Facing imbalanced data–recommendations for the use of performance metrics. Em *IEEE Humaine Association Conference on Affective Computing and Intelligent Interaction*.

Jiang, H. e Learned-Miller, E. (2017). Face detection with the faster r-cnn. Em *IEEE Conference on Automatic Face and Gesture Recognition*.

Kaltwang, S., Todorovic, S. e Pantic, M. (2015). Latent trees for estimating intensity of facial action units. páginas 1–9.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*.

Kingma, D. P. e Ba, J. L. (2015). Adam: a Method for Stochastic Optimization. *International Conference on Learning Representations*.

Krizhevsky, A., Sutskever, I. e Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Em *Conference and Workshop on Neural Information Processing Systems*.

Li, X., Chen, S. e Jin, Q. (2017). Facial action units detection with multi-features and -aus fusion.

Martinez, B., Valstar, M. F., Jiang, B. e Pantic, M. (2017). Automatic analysis of facial actions: A survey. *IEEE Transactions on Affective Computing*, PP(99):1–1.

Mavadati, S. M., Member, S., Mahoor, M. H., Bartlett, K., Trinh, P. e Cohn, J. F. (2013). Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 6(1):1–11.

Nicolle, J., Bailly, K. e Chetouani, M. (2016). Real-time facial action unit intensity prediction with regularized metric learning. *Image and Vision Computing*.

Ren, S., He, K., Girshick, R. e Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. Em *Conference and Workshop on Neural Information Processing Systems*.

Sariyanidi, E., Gunes, H. e Cavallaro, A. (2015). Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1113–1133.

Shrout, P. e Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

Simon, M., Rodner, E. e Denzler, J. (2016). Imagenet pre-trained models with batch normalization. *arXiv preprint arXiv:1612.01452*.

Simonyan, K. e Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations*.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. e Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Tang, C., Zheng, W., Yan, J., Li, Q., Li, Y., Zhang, T. e Cui, Z. (2017). View-independent facial action unit detection. páginas 1–5.

Tősér, Z., Jeni, L. A., Lőrincz, A. e Cohn, J. F. (2016). Deep learning for facial action unit detection under large head poses. Em *Springer European Conference on Computer Vision Workshops*.

Valstar, M., Lozano, E. S., Cohn, J. F., Jeni, L. A., Girard, J. M., Yin, L., Zhang, Z. e Pantic, M. (2017). Fera 2017 - addressing head pose in the third facial expression recognition and analysis challenge. *arXiv preprint arXiv:1702.04174*.

Valstar, M. F., Almaev, T., Girard, J. M., McKeown, G., Mehu, M., Yin, L., Pantic, M. e Cohn, J. F. (2015). Fera 2015 - second facial expression recognition and analysis challenge. Em *IEEE Conference on Automatic Face and Gesture Recognition Workshops*.

Walecki, R., Rudovic, O., Pavlovic, V. e Pantic, M. (2016). Copula ordinal regression for joint estimation of facial action unit intensity. Em *IEEE Conference on Computer Vision and Pattern Recognition*.

Walecki, R., Rudovic, O., Pavlovic, V., Schuller, B. e Pantic, M. (2017). Deep structured learning for facial action unit intensity estimation. Em *IEEE Conference on Computer Vision and Pattern Recognition*.

Werner, P., Saxen, F. e Al-Hamadi, A. (2015). Handling data imbalance in automatic facial action intensity estimation. *British Machine Vision Conference*.

Yüce, A., Gao, H. e Thiran, J.-P. (2015). Discriminant multi-label manifold embedding for facial action unit detection. Em *IEEE Conference on Automatic Face and Gesture Recognition*.

Zavan, F. H. d. B., Gasparin, N., Batista, J. C., Silva, L. P. e., Albiero, V., Lucio, D. R., Bellon, O. R. e Silva, L. (2017). Beyond flat faces: Facial image analysis and processing in the wild. Em *SIBGRAPI Conference on Graphics, Patterns and Images*.

Zavan, F. H. d. B., Nascimento, A. C. P., e Silva, L. P., Bellon, O. R. P. e Silva, L. (2016). 3d face alignment in the wild: A landmark-free, nose-based approach. Em *Springer European Conference on Computer Vision Workshops*.

Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., Liu, P. e Girard, J. M. (2014). Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*.

Zhang, Z., Girard, J. M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H., Cohn, J. F., Ji, Q. e Yin, L. (2016). Multimodal spontaneous emotion corpus for human behavior analysis. Em *IEEE Conference on Computer Vision and Pattern Recognition*.

Zhao, K., Chu, W.-S. e Zhang, H. (2016). Deep region and multi-label learning for facial action unit detection. Em *IEEE Conference on Computer Vision and Pattern Recognition*.

Zhou, Y., Pi, J. e Shi, B. E. (2017). Pose-independent facial action unit intensity regression based on multi-task deep transfer learning. páginas 872–877.