Investigando a plausibilidade de classificadores de imagens aplicados à detecção de motoristas distraídos

Flávio Rosendo da Silva Oliveira
Universidade de Pernambuco / Instituto Federal de Pernambuco
Recife, Brasil
frso@ecomp.poli.br

Felipe Costa Farias
Instituto Federal de Pernambuco
Campus Paulista
Paulista, Brasil
felipe.farias@paulista.ifpe.edu.br

Fernando Buarque de Lima Neto
Universidade de Pernambuco
Escola Politécnica
Paulista, Brasil
fbln@ecomp.poli.br

***Abstract*—O número de mortes anuais em acidentes de trânsito, no Brasil, ultrapassa 40 mil segundo dados da Organização Mundial da Saúde. Entre as principais causas de acidentes encontram-se as distrações ao volante, cuja gravidade é aumentada em até 4 vezes pelo uso do celular. Este artigo apresenta método destinado a aferir a plausibilidade de classificadores de imagens, aplicados à detecção de motoristas distraídos. São apresentados experimentos investigando a plausibilidade de diferentes arquiteturas de aprendizagem profunda, utilizadas nesta tarefa. Os resultados sugeriram que o método apresentou-se como uma alternativa viável para auxiliar a investigação da plausibilidade de classificadores de imagens, e neste problema, auxiliou a observação de alguns aspectos não triviais no comportamento dos modelos investigados.**

Keywords—visão computacional; aprendizagem profunda; motoristas distraídos; plausibilidade

I. INTRODUÇÃO

Segundo relatório publicado pela Organização Mundial da Saúde [1], mais de 40 mil pessoas perdem a vida no trânsito no Brasil, todos os anos. Estudos indicam que até 80% do volume dos acidentes de trânsito podem ser correlacionados com distrações do condutor [2]. Outro estudo identifica o uso do celular ao volante como causador de aumento em até 4 vezes da probabilidade do condutor se envolver em um acidente grave [3].

Neste contexto, as técnicas de Visão Computacional [4] poderiam ser empregadas para, mediante análise de imagens de um condutor, avaliar se o mesmo encontra-se distraído ou concentrado ao volante. A partir da percepção de métricas acerca dos episódios de distração ocorridos, um condutor poderia se conscientizar sobre suas práticas ao volante e eventualmente tomar medidas no sentido de minimizar a quantidade de distrações ocorridas, quer ocorram voluntariamente ou não.

Considerando que vasta quantidade de modelos de aprendizagem de máquina são ditos modelos de caixa-preta, nos quais não é possível inferir as causas de uma determinada classificação a partir da inspeção direta do seu processamento interno, estudos têm sido realizados no sentido de obter explicações acerca da classificação destes modelos visando melhorar a sua confiabilidade e ampliar sua adoção no mercado. Esta área de pesquisa tem sido denominada Inteligência Artificial Explicável (XAI) [5].

Neste artigo, foi realizado estudo visando investigar o quão plausíveis são classificadores de imagens aplicados à tarefa de identificar condutores distraídos. Adicionalmente, se busca verificar se arquiteturas diferentes baseiam-se nas mesmas áreas de atenção para realizar suas inferências e se estas áreas são plausíveis para as inferências realizadas. Para tanto, utilizou-se como referência, áreas da imagem que seriam consideradas importantes por um especialista humano para realizar a classificação. O método proposto foi então empregado para investigar, mediante análise comparativa, o grau de plausibilidade de diferentes classificadores utilizando aprendizagem profunda, quando aplicados à detecção de motoristas distraídos.

O restante deste artigo está estruturado conforme a seguir: (i) na seção 2, são apresentados os modelos de aprendizagem profunda utilizados na análise comparativa e informações relacionadas à área de Inteligência Artificial Explicável (XAI); (ii) na seção 3 é apresentada a metodologia, descrevendo como a plausibilidade da classificação foi quantificada neste estudo e também é caracterizada a base de dados utilizada; (iii) na seção 4 a configuração experimental e os resultados são apresentados e por fim; (iv) na seção 5 é apresentada a discussão e conclusão do trabalho.

II. REFERENCIAL TEÓRICO

A. Aprendizagem Profunda

As Redes Neurais Convolucionais (RNCs) tem ganho destaque pelo bom desempenho obtido quando utilizadas em tarefas de Visão Computacional. Tais redes possuem duas partes essenciais: (i) um conjunto de camadas convolucionais, responsáveis por decompor as imagens de entrada em características adequadas para a classificação e (ii) camadas densamente conectadas, que são empregadas para realizar o aprendizado no sentido de relacionar tais características e as classes abordadas no problema. Normalmente, o principal diferenciador entre as arquiteturas de RNC é a forma como as camadas convolucionais e densas são conectadas, e eventualmente o emprego de estruturas internas especiais. Neste trabalho, foram empregadas quatro arquiteturas de RNCs com diferentes características, visando fornecer diversidade à

investigação. Naturalmente esta lista não é exaustiva, mas estas quatro arquiteturas apresentam características bastante representativas do estado da arte em RNC. São elas: ResNet, Densenet, Inception e Mobilenet.

A arquitetura Inception v3, proposta por Szegedy et al. [6], foi projetada para superar problemas de performance observados em RNCs anteriores, como as arquiteturas VGG. Para tanto, foram empregadas diversas estratégias, tais como: (i) evitar gargalos representacionais, especialmente em camadas iniciais; (ii) utilizar representações com alta dimensionalidade, para permitir melhor fluxo de informações no modelo e (iii) balancear largura e profundidade da rede.

A arquitetura Resnet foi proposta por He et. al [7] e é caracterizada por aumentar a profundidade da rede, acrescentando mais camadas convolucionais. Por exemplo, as arquiteturas VGG possuem tipicamente 16 ou 19 camadas convolucionais. As Resnet apresentadas em He et. al, chegam a ter 152 camadas convolucionais. Apesar deste aumento de profundidade a Resnet utiliza um número menor de filtros convolucionais, proporcionando aumento de precisão acompanhado de expressiva redução na complexidade computacional. Além disso, são utilizadas propagações residuais do sinal entre camadas não consecutivas a fim de reduzir o problema de *vanishing gradients* e aumentar a convergência de treinamento.

A arquitetura Densenet, proposta por Huang et al. [8], amplia ainda mais a profundidade da rede acrescentando mais camadas convolucionais. Nesta arquitetura, em seu artigo original, foram testadas redes com até 264 camadas convolucionais obtendo desempenho compatível com o estado-da-arte à época, sem aumento significativo no custo computacional. Para lidar com o problema de *vanishing gradients*, é realizada a conexão das saídas das camadas convolucionais às entradas de todas as camadas posteriores, propiciando melhor fluxo de informação (fase *forward*) e melhor propagação de gradientes (base *backwards*), acelerando a convergência dos algoritmos de treinamento.

Por fim, a arquitetura Mobilenet [9] utiliza convoluções em profundidade seguidas de convoluções pontuais, reduzindo de maneira expressiva a quantidade de cálculos a serem realizados durante o treinamento, o que o acelera e torna a arquitetura apta a ser executada em dispositivos com poder computacional modesto, como dispositivos móveis.

B. Inteligência Artificial Explicável

Segundo a Agência de Defesa e Projetos de Pesquisa Avançada (DARPA), o próximo passo na evolução da Inteligência Artificial é torná-la auditável ou explicável [10]. Neste sentido, vários trabalhos já foram realizados aplicando XAI a problemas de Visão Computacional.

Hendricks et. al [11] produziram um modelo capaz de classificar uma imagem e na sequência, gerar uma explicação textual que discrimina as principais propriedades de uma imagem. Este resultado foi obtido combinando uma Rede Neural Convolucional (RNC) para realizar a classificação e uma combinação de módulos *Long Short Term Memory* (LSTM) para gerar o texto da explicação.

Yang e Shafto [12] propuseram abordagem de explicação com base em exemplos, utilizando os princípios de suas pesquisas com Ensino Bayesiano. Nesta abordagem, o objetivo é selecionar um subconjunto dos padrões contidos numa base de dados, a partir do qual o modelo sendo treinado seria capaz de derivar conclusões acerca do restante da base de dados. Dentre as vantagens que os autores alegam sobre este método, está a possibilidade de utilizá-lo em conjunto com outros modelos treinados com algoritmos supervisionados, não supervisionados ou por reforço.

No trabalho de Bau et. al [13], foi proposto um *framework* geral intitulado Dissecação de Rede, utilizado para quantificar a interpretabilidade das representações latentes em RNCs. O método proposto avalia as unidades ocultas e as correlaciona com imagens contidas num banco de imagens, contendo objetos, cenas, texturas, cores e materiais. Estas unidades então recebem rótulos com base na sua correlação com as imagens.

A proposta de Ribeiro et. al [14] consiste em derivar um modelo linear visando explicar o comportamento local de um classificador mais complexo, possivelmente não-linear. Segundo os autores, ao utilizar o algoritmo proposto neste trabalho, intitulado LIME (*Linear Interpretable Model-Agnostic Explanations*), é possível obter insights acerca do funcionamento interno de qualquer modelo verificando quais são os atributos ou características mais importantes para certa classificação numa base de dados tabular ou quais super-pixels são mais relevantes para a classificação numa base de imagens.

C. Trabalhos Relacionados

Trabalhos anteriores exploraram o tema da detecção de imagens de condutores distraídos. Hssayeni et al. [15] empregou engenharia de características e comparou a performance de classificadores utilizando tais características com RNC. Nestes experimentos, as RNC superaram os classificadores treinados com características geradas manualmente. Masood et al. [16] utilizou RNC e comparou as métricas de performance de modelos inicializados randomicamente e modelos pré-treinados em outra base. Nestes experimentos, as redes pré-treinadas superaram as anteriores. No trabalho de Oliveira e Farias [17] foi apresentada investigação acerca do impacto do uso de diferentes metodologias de transferência de aprendizagem, no qual a metodologia de transferência com refinamento de toda a rede apresentou desempenho superior quando comparado ao refinamento apenas das camadas densamente conectadas e superior a classificadores tradicionais treinados com características geradas a partir dos filtros convolucionais de RNC.

A diferenciação mais marcante deste trabalho em relação aos anteriores é seu enfoque nas características de explicabilidade exploradas no problema, que não foi observado nos outros trabalhos relacionados à detecção de motoristas distraídos.

III. Materiais e Métodos

A. Abordagem utilizada na investigação

Visando investigar detalhes acerca do funcionamento interno de classificadores de imagem, propõe-se um método

que objetiva condensar informação acerca do quão plausível é a operação de dado classificador em relação ao que seria considerado razoável por um certo especialista humano. Na Figura 1 pode-se verificar o algoritmo que elucida o método proposto:

- Dada uma base de dados *DB* contendo tuplas (*I*,*y*) nas quais *I* é uma imagem destinada à classificação e *y* é sua classe verdadeira
- Dado um conjunto *M*, contendo tuplas (m_i,c_i) nas quais m_i é uma máscara de atenção com informações necessárias para demarcar uma região na imagem que é considerada plausível por um especialista humano como sendo importante para a determinação de cada classe c_i e;
- Dado também um classificador treinado previamente *C*.

1.	**função** calcula_plausibilidade(*DB*,*M*,*C*): plausibilidade_percentual
2.	**para** cada tupla (*I*,*y*) em DB **faça**
3.	calcule a predição *y'* submetendo a imagem *I* ao classificador *C*
4.	**se** *y* igual a *y'* **então**
5.	determine o conjunto de pixels *P*, mais determinante para que *I* seja classificado como *y'*
6.	Acumule em *AC* o percentual de pixels *P* contidos na máscara *M*
7.	**fim-se**
8.	**fim-para**
9.	retorne *AC* / (nº de Imagens em DB)
10.	**fim-função**

Fig. 1. Algoritmo da função utilizada no cálculo da plausibilidade

A relevância das análises que podem ser realizadas com a Plausibilidade Percentual depende em certa medida de como foram geradas as máscaras de atenção para cada classe e também do método utilizado para determinar os pixels mais importantes para a classificação. Estes dois elementos podem ter seus parâmetros ajustados, notadamente no método utilizado para a determinação dos pixels mais relevantes. Destaca-se também que dois especialistas podem determinar máscaras de atenção significativamente diferentes para as classes do mesmo problema, já que a determinação possui componente subjetivo, pois a demarcação das máscaras de atenção deriva da experiência e conhecimento do especialista.

Em função da necessidade de determinar uma máscara para a área de atenção previamente, que é fixa para cada classe, o método proposto não deve ser utilizado com classificadores que foram treinados utilizando métricas de *data augmentation* que manipulam o zoom, transladam ou rotacionam as imagens. Embora tais métodos sejam amplamente utilizados visando evitar *overfitting* no treinamento de arquiteturas baseadas em aprendizagem profunda, é possível que causem problemas na avaliação da Plausibilidade Percentual, a menos que a máscara seja ajustada automaticamente em cada imagem, de acordo com a transformação adotada.

Considerando que o objetivo de tal métrica não é aferir o quanto um classificador acerta as classes e nem compreender o nível de confusão entre elas, o cálculo da métrica ocorre apenas sobre os padrões classificados corretamente, conforme a linha 4. Um classificador com baixa acurácia tende a ter baixa Plausibilidade Percentual pela correção gerada no cálculo da média conforme linha 9, que considera todas as imagens contidas em *DB*. Um classificador com baixa acurácia, pode ter percentuais altos calculados em *AC* (ver linha 6) para cada imagem correta avaliada. No entanto o valor da média tenderá a ser reduzido (ver linha 9).

B. Caracterização da Base de Dados

A base de dados utilizada foi produzida a partir daquela originalmente gerada pela empresa State Farm [18]. A base originalmente contém 22424 imagens, sendo estas distribuídas entre 9 tipos de distração ao volante e a classe que representa os condutores concentrados.

Para confeccionar a base de imagens, foram produzidos vídeos curtos nos quais os condutores realizaram os 10 tipos de ações considerados. A partir daí, alguns *frames* de vídeo foram selecionados como exemplos de cada classe. Os condutores apresentados nas imagens são de diferentes grupos étnicos e possuem diferentes idades sendo parte do sexo masculino ou feminino, conforme pode ser visto na Figura 2. No total, a base de dados possui 26 condutores diferentes.

Em função da relevância dos acidentes causados pelo uso do celular, este estudo enfocou apenas 5 classes, sendo 4 tipos de distração utilizando o celular e o estado de condução concentrada.

IV. Experimentos e Resultados

Para a execução dos experimentos relatados nesta seção, foi empregada instância de servidor na plataforma Amazon AWS p2xlarge, cuja configuração corresponde a: 4 vCPUs, 61GB de memória RAM e GPU K80, da qual são disponibilizados 4096 cuda cores e 11GB de memória RAM. Foram construídos scripts com a linguagem Python, empregando o Scikit Image e Keras.

A. Configuração Experimental

Com o propósito de simular o papel do especialista, foram escolhidas aleatoriamente 5 imagens do conjunto de treinamento para cada classe e em cada imagem foi realizada a marcação de um retângulo correspondendo à área considerada plausível para realizar a classificação.

Fig. 2. Imagens de exemplo de 4 das 5 classes. Condutora concentrada (superior esquerda), condutora segurando celular na mão direita (superior direita), condutor teclando com mão esquerda (inferior esquerda) e condutora segurando celular com mão esquerda (inferior direita).

As coordenadas destas máscaras foram extraídas e depois a média aritmética das coordenadas de cada imagem em cada classe foi utilizada como máscara de atenção para avaliar a plausibilidade de classificação de cada classificador conforme Figura 1, linha 6. Abaixo na Tabela I, pode-se visualizar as coordenadas utilizadas como máscara de atenção em cada classe.

TABELA I. COORDENADAS DAS MÁSCARAS UTILIZADAS PARA CADA CLASSE AVALIADA

Classe	Coord. X1	Coord. X2	Coord. Y1	Coord. Y2
c0	113	202	61	190
c1	136	173	89	163
c2	51	104	63	130
c3	117	174	87	160
c4	74	119	50	110

Foram selecionadas quatro arquiteturas com diferentes características internas a fim de realizar a investigação. Todas estão disponíveis na biblioteca Python Keras: ResNet, DenseNet, Inception V3 e Mobilenet. As três primeiras arquiteturas foram utilizadas previamente em Oliveira e Farias [17] e se mostraram factíveis para classificação neste tipo de problema, utilizando transferência de conhecimento. Foi agregada a arquitetura Mobilenet a fim de investigar o comportamento de uma arquitetura projetada para execução em dispositivos com poder computacional reduzido.

Para a transferência de conhecimento foi realizada a metodologia *hold-out*, empregando a seguinte estratégia: para cada classe foram selecionadas 64 imagens para treinamento e 32 imagens para testes. Em cada classe, condutores selecionados para o conjunto de treinamento não são utilizados no conjunto de testes, visando obter uma avaliação de generalização mais precisa. As arquiteturas originalmente treinadas na base de dados ImageNet tiveram suas camadas totalmente conectadas substituídas por duas novas camadas a serem treinadas. A primeira camada recebe como entradas as saídas das camadas convolucionais, contendo 1024 neurônios. A segunda camada, de saída, contém 5 neurônios, a mesma quantidade de classes. Estas saídas são submetidas a uma camada *Softmax* que determina a classe de um dado padrão de acordo com a maior ativação contida entre os 5 neurônios de saída. Estas arquiteturas tiveram todos os pesos ajustados por até 15 épocas máximas, executando o algoritmo SGD com taxa de aprendizagem inicial 0,001 e *momentum* 0,9.

Após a etapa de *transfer learning*, foram selecionadas as duas arquiteturas com maior acurácia para a investigação de plausibilidade. Para esta etapa, além das máscaras de atenção mencionadas anteriormente, foi utilizado o algoritmo LIME. A priori, qualquer metodologia para a inferência dos pixels mais importantes poderia ter sido utilizada, conforme Figura 1, linha 5. A escolha pelo LIME se deu pela sua integração simples à pilha de software utilizada. Com o LIME, foi realizada a extração dos 5 super-pixels mais importantes para a classificação de cada imagem, conforme Figura 1, linha 5. O LIME foi parametrizado de maneira a gerar 1000 padrões derivados a partir do padrão em análise, a fim de estabelecer o modelo localmente interpretável que propicia a determinação dos pixels mais relevantes. Os demais parâmetros foram utilizados conforme sua configuração padrão.

O cálculo da plausibilidade percentual foi realizado para as quatro arquiteturas, mas a geração das matrizes de confusão e mapas de calor foram apenas realizados com as duas arquiteturas com melhores plausibilidades e acurácias.

B. Resultados

Os resultados em termos de Acurácia de treinamento e testes, além da plausibilidade percentual das arquiteturas testadas podem ser vistos na Tabela II.

TABELA II. ARQUITETURAS, ACURÁCIAS E PLAUSIBILIDADES AVALIADAS

Arquitetura	Acurácia de Treinamento (%)	Acurácia de Teste (%)	Plausibilidade Teste (%)
Resnet 50	100,00	46,88	6,32
Densenet 121	99,69	75,00	25,84
Mobilenet	98,44	80,62	28,52
Inception v3	100,00	51,25	7,24

É possível observar que nenhuma das redes apresentou plausibilidade superior a 29%, o que a priori pode ser considerado uma taxa baixa.

O baixo volume de dados e/ou pequena quantidade de épocas permitidas durante o *transfer learning*, provavelmente impactou negativamente o aprendizado das arquiteturas Inception V3 e Resnet50. Sua alta taxa de acertos no treinamento e resultado muito inferior no conjunto de testes sugere que ocorreu *overfitting*.

A seguir é realizada uma análise mais aprofundada acerca das arquiteturas Densenet e Mobilenet que apresentaram resultados mais promissores em termos de Acurácia e plausibilidade. A diferença entre a acurácia de treinamento e testes, mesmo nas arquiteturas Densenet e Mobilenet pode ser explicada pela forma como foi realizada a separação entre os conjuntos de treinamento e testes, conforme explicado na seção 3.B.

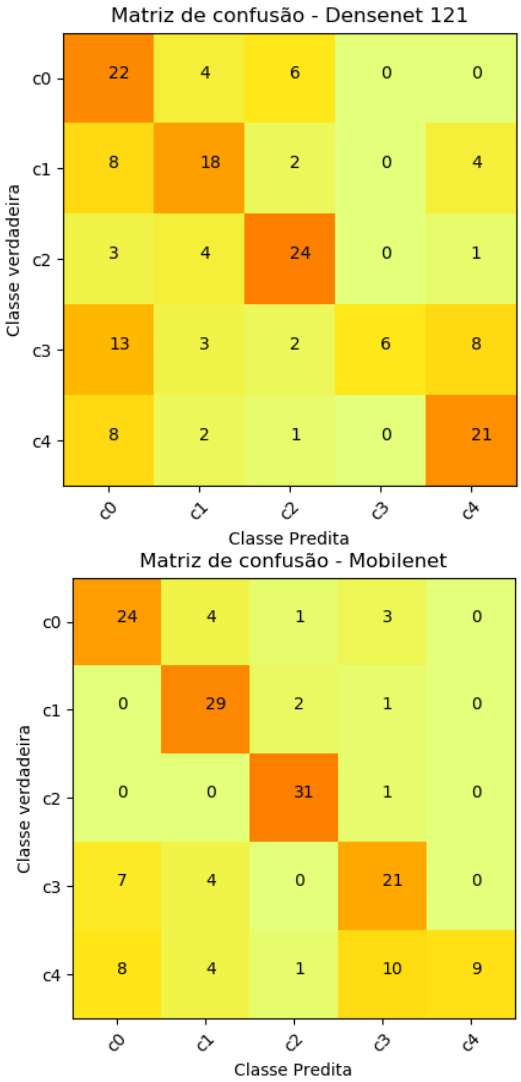


Fig. 3. Matrizes de confusão no conjunto de testes das arquiteturas Densenet121 (superior) e Mobilenet (inferior).

Quanto às matrizes de confusão das arquiteturas com maior Acurácia e maior plausibilidade percentual, pode-se observar algumas diferenças significativas , que podem ser vistas na Figura 3. Mobilenet acertou mais imagens nas classes c0,c1,c2 e C3, enquanto Densenet foi superior na classe c4.

Uma das contribuições deste trabalho, relacionada à área de XAI, é que sem inspecionar a plausibilidade, provavelmente se aceitariam as arquiteturas Densenet ou Mobilenet para utilização após o término da etapa de modelagem. No entanto, mesmo estas redes apresentaram níveis relativamente baixos de plausibilidade, o que sugeriria a necessidade de uma inspeção mais detalhada do seu funcionamento interno.

Na Tabela III, disposta ao final do artigo, podem ser vistos os mapas de calor em relação aos pixels mais utilizados como explicações para a classificação das classes c0 a c4 nas arquiteturas Densenet e Mobilenet, além de uma imagem selecionada aleatoriamente em cada classe. Vale ressaltar que o retângulo em amarelo representa uma das máscaras criada por especialista e os retângulos vermelhos representam as máscaras utilizadas para cálculo da plausibilidade percentual, conforme explanado na seção 4.A, Tabela I.

Ao observar a Tabela III, é possível inferir que:

- No caso da classe c0, os dois modelos apresentaram comportamento similar. Os pixels mais importantes estiveram em grande medida de acordo com a máscara usada como referência. Estes pixels estão mais localizados na região entre o corpo do condutor e o volante, sendo áreas plausíveis para avaliar a concentração dos condutores.
- No caso da classe c1, o número de imagens foi bem inferior na Densenet 121 e não se formou para esta uma área coerente. No caso da Mobilenet, é possível observar uma área que está próxima ao corpo do condutor e no que parece ser um dos braços. Embora não tenha havido coerência com as máscaras de referência, a Mobilenet aparenta ter se baseado em áreas da imagem plausíveis para realizar a classificação quando o usuário está teclando com a mão direita.
- No caso da classe c2, Mobilenet superou Densenet em 7 imagens. Apesar disso, a Densenet 121 teve os pixels mais relevantes onde parece estar o corpo mais próximo da cabeça do condutor, que é plausível para identificar o condutor falando ao celular com a mão direita. A Mobilenet parece considerar esta área e também a região próxima ao volante, provavelmente realizando uma verificação mais abrangente.
- A classe c3, parece ter sido a mais desafiadora para ambos os modelos, considerando o alto nível de confusão desta com a classe 0. A Densenet 121 acertou apenas 6 imagens mas os pixels relevantes foram bastante coerentes com a área entre o corpo do condutor e o volante – local onde se está digitando no celular com a mão esquerda. A Mobilenet superou a última em 15 imagens, mas os seus pixels mais relevantes para a classificação não formaram uma área coerente no mapa de calor, apresentando-se de forma difusa.
- Por fim, na classe c4, na qual Densenet 121 superou Mobilenet em 12 imagens, há uma formação de concentração de pixels relevantes em região próxima ao corpo do condutor. No caso da Mobilenet, isto não ocorre, sendo revelado um padrão similar à c1 e c2.

V. Conclusão

Considerando a quantidade de acidentes no trânsito do Brasil e a relevância das distrações do condutor neste quantitativo, este trabalho objetivou investigar o quão plausíveis são classificadores de imagens aplicados a identificar tais distrações.

Após realizar *transfer learning*, adequando quatro arquiteturas treinadas na base Imagenet, foi possível aprofundar o estudo nas duas com melhores acurácias. Nestas, apesar de apresentarem valores relativamente baixos na métrica plausibilidade percentual, foi possível identificar pontos positivos em ambas arquiteturas que poderiam inclusive ser combinadas em um comitê. A combinação dos modelos poderia propiciar bons níveis de acurácia, aliados à capacidade de explicar suas inferências a partir das áreas de atenção.

Além disso, a análise exibida na Tabela III, mostrou-se oportuna no sentido de verificar padrões gerais observados na Densenet e Mobilenet, tendo sido possível observar que embora apresentem níveis próximos de acurácia de testes e de plausibilidade, há diferenças significativas na maneira como utilizam as áreas de atenção. Além do método proposto, esta é a principal contribuição desta investigação.

É possível observar que foram obtidos baixos níveis de Plausibilidade Percentual. Estes níveis baixos podem ter sido obtidos em função de: (i) máscaras de referência definidas de forma muito rígida e/ou utilizando área muito restrita; (ii) parametrização do LIME que não foi explorada em profudidade, (iii) resultados do treinamento das RNC que não leva em consideração informações do especialista, visando apenas reduzir o erro de classificação e (iv) a metodologia de cálculo que considera conjuntamente todas as classes. A determinação de qual destes fatores teve maior relevância requer aprofundamento do estudo.

A própria utilização do LIME poderia ser objeto de discussão. Embora seja um método de extração de pixels relevantes, ou de explicabilidade de classificadores, não é o único mecanismo disponível para realizar tal tarefa. Outras alternativas além do LIME merecem ser investigadas.

Por fim, o uso das arquiteturas de RNC pode ser considerada uma limitação do estudo. Primeiramente porque não representam a totalidade das arquiteturas disponíveis e depois por serem técnicas projetadas para problemas muito mais complexos em termos de quantidade de classes e volume de padrões. No entanto, esta seleção de arquiteturas e de técnica de treinamento é um ponto de partida, sobre o qual estudos posteriores podem ser realizados. Foi possível observar diferenças no tocante à plausibilidade das arquiteturas, o que se pode considerar junto às propostas, como resultados iniciais, porém promissores.

Com o encorajamento de tais resultados, os trabalhos futuros apontam na direção de: (i) estudar o impacto de outros tipos de máscaras em outros tipos de problema, (ii) empregar outras arquiteturas de aprendizagem profunda, (iii) utilizar outros mecanismos de determinação das áreas de atenção, (iv) aprimorar o cálculo da plausibilidade, possivelmente investigando-a no nível de cada classe e (v) avaliar os resultados com outras estratégias de treinamento, por exemplo utilizando classificadores binários independentes para cada classe.

Vale destacar ainda, que tais resultados apontam na direção de ser necessário incorporar aspectos subjetivos no treinamento e avaliação de RNCs, notadamente em situações nas quais haja alto impacto sobre a utilização dos modelos. Isto ocorre frequentemente em problemas de apoio a decisão, que são campo de aplicação para XAI. Este tipo de algoritmo de treinamento híbrido, envolvendo aspectos objetivos e subjetivos, sendo os últimos fortemente relacionados às expectativas dos usuários, serão explorados em trabalhos futuros.

Agradecimentos

Os autores agradecem ao PPGEC/UPE e ao IFPE Campus Paulista pelo apoio na realização deste trabalho.

References

[1] Organização Mundial da Saúde, Segurança Viária no Brasil, Accessado em 10/09/2019 https://www.paho.org/bra/index.php?option=com_content&view=article&id=5147:acidentes-de-transito-folha-informativa&Itemid=779

[2] National Highway Traffic Safety Administration, “The impact of driver inattention on near-crash/crash risk: an analysis using the 100-car naturalistic driving study data”, Department of Transportation, NHTSA, Washington, 2006.

[3] S.P. Mcevoy et al., “Role of mobile phones in motor vehicle crashes resulting in hospital attendance: a case-crossover study”. Br Med J331:428–430, 2005.

[4] Backes, A. “Introdução à visão computacional usando Matlab”, 1ª Edição. Brasil: Elsevier, 2016. 290 pgs.

[5] Dosilovic, F. K., Brcic, M. e Hlupic , N. “Explainable artificial intelligence: A survey”, In: Anais do 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2018.

[6] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. e Wojna, Z. “Rethinking the Inception architecture for computer vision”. http://arxiv.org/abs/1512.00567.

[7] He, K. , Zhang, X., Ren, S. e Sun, J. “Deep residual learning for image recognition”, In Proceedings of CVPR, pages 770–778, 2016. http://arxiv.org/abs/1512.03385.

[8] Huang, G., Liu, Z. e Weinberger, K.Q. “Densely connected convolutional networks”, http://arxiv.org/abs/1608.06993.

[9] Howard, A. G. “Mobilenets: Efficient convolutional neural networks for mobile vision applications”, 2017. arXiv:1704.04861

[10] DARPA. Explainable Artificial Intelligence - XAI. Disponível em https://www.darpa.mil/attachments/XAIProgramUpdate.pdf . Acessado em 10/09/2019.

[11] Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B. e Darrell, T. “Generating Visual Explanations”, In: Anais do 14th European Conference on Computer Vision, pp. 3-19, 2016.

[12] Yang, S.C.H. e Shafto, P. Explainable Artificial Intelligence via Bayesian Teaching, In: Proceedings of 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017.

[13] Bau, D., Zhou, B., Khosla, A., Oliva, A. e Torralba, A. “Network Dissection: Quantifying Interpretability of Deep Visual Representations”. In: Anais do 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017.

[14] Ribeiro, M.T., Singh, S. e Guestrin, C. “Why Should I Trust You? Explaining the predictions of any classifier”. In: Anais do 22nd Conference of Knowledge Discovery and Data Mining, 2016.

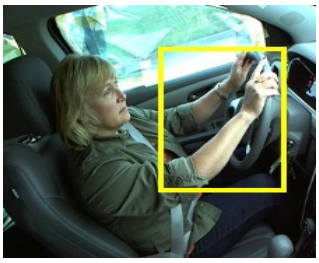
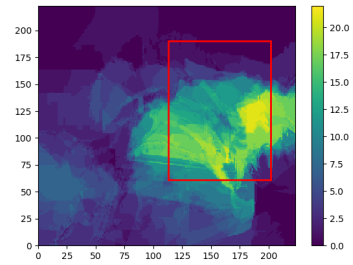
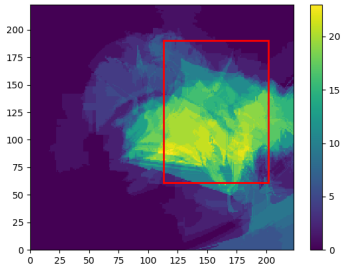
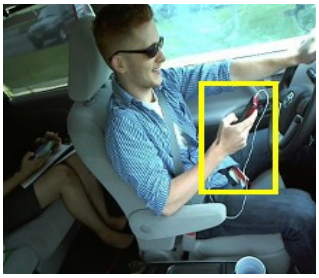
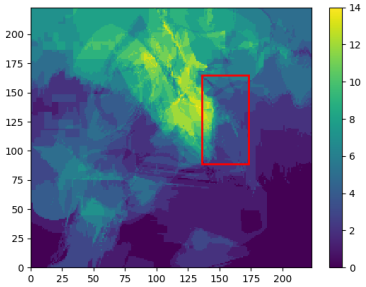
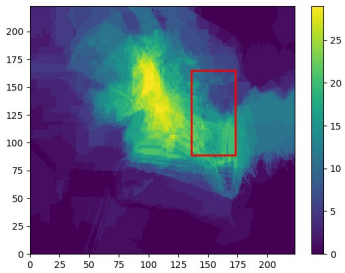
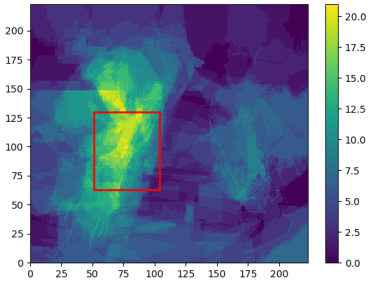
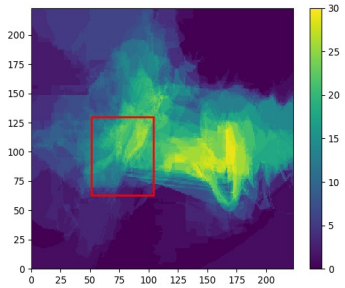
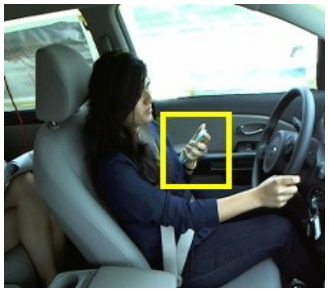
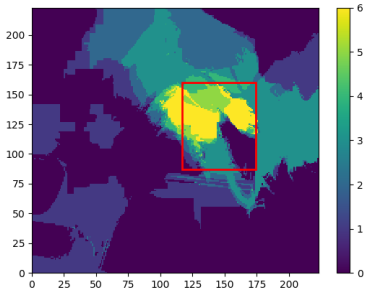
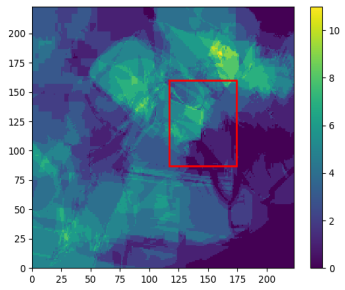
[15] Hssayeni, M., Saxena, S., Ptucha, R. e Savakis, A. “Distracted Driver Detection: Deep Learning vs Handcrafted Features”, In Imaging and Multimedia Analytics in a Web and Mobile World 2017, pp. 20-26(7).

[16] Masood, S., Rai, A., Aggarwal, A., Doja, M. e Ahmad, M. “Detecting Distraction of drivers using Convolutional Neural Network”, Pattern Recognition Letters, 2018.

[17] Oliveira, F. R. S. e Farias, F. C. “Comparing transfer learning approaches applied to distracted driver detection”. In: Anais do 2018 IEEE Latin American Conference on Computational Intelligence (LA-CCI), 2018, Guadalajara, Jalisco, México. Proceedings of 5th IEEE Latin American Conference on Computational Intelligence (LA-CCI), 2018.

[18] State Farm Distracted Drivers Dataset, acessado em 10/09/2019. https://www.kaggle.com/c/state-farm-distracted-driver-detection/data

TABELA III. IMAGENS ORIGINAIS COM A MÁSCARA ESPECÍFICA DO ESPECIALISTA E MAPAS DE CALOR OBTIDOS PARA DENSENET121 E MOBILENET, COM A MÁSCARA MÉDIA. IMAGEM MELHOR VISTA COLORIDA.

Classe	Imagem original com máscara	Mapa de calor Densenet	Mapa de calor Mobilenet
c0			
c1			
c2			
c3			
c4