ESTUDO COMPARATIVO ENTRE REDES NEURAIS CONVOLUCIONAIS SIAMESAS COM MÉTODOS CLÁSSICOS DE REDES NEURAIS PROFUNDAS EM UMA APLICAÇÃO DE RASTREAMENTO E IDENTIFICAÇÃO DE PESSOAS

COMPARATIVE STUDY BETWEEN METRIC LEARNING AND CLASSIC METHODS OF DEEP NEURAL NETWORKS IN A PEOPLE TRACKING AND IDENTIFICATION APPLICATION

Leite, Fabricio T. [a]. Lochter, Johannes V. [b]

[a]Centro Universitário Facens - Sorocaba, SP, Brasil
fabricio.torquato1@gmail.com
[b]Centro Universitário Facens - Sorocaba, SP, Brasil
johannes.lochter@facens.br

Submetido em: 30 Jan. de 2022. Aceito em: --

RESUMO

Quando se têm um ambiente vigiado por câmeras de segurança onde precisa determinar se uma pessoa na qual está aparecendo na imagem já esteve presente no vídeo anteriormente ou se é sua primeira aparição é chamado dentro da área de pesquisa de inteligência artificial como re-identificação de pessoas. Esse problema é apresentado como um desafio entre pesquisadores uma vez que tais imagens coletadas por câmeras de segurança podem apresentar diversos ruídos visuais além de obstruções parciais das pessoas. Desse modo, inspirado no ensinamento do historiador grego Heródoto, pensar o passado para compreender o presente e idealizar o futuro, esse trabalho vem apresentar através de métricas entre a FaceNet e MobileNetV2 quando exposta ao dataset da WiseNet voltado para o problema de re-identificação, sendo que ambas redes tiveram desempenhos de F-Medida e acurácia muito próximos, entretanto quando comparado ao tempo de inferência do experimento e o de treinamento a FaceNet mostra suas vantagens, na qual com esses dados inferidos esperasse ter ajudado a contribuir para futuras evoluções e melhorias nos algoritmos da área.

Palavras-chave: Re-identificação de Pessoa, Vídeo Analytics, Avaliação por Métrica.

ABSTRACT

When they have an environment monitored by security cameras where they need to determine if a person who is appearing in the image was already present in the video before or if it is their first appearance, it is called within the field of artificial intelligence research as re-identifying people. This problem is presented as a challenge among researchers since such images collected by security cameras can present various visual noises in addition to partial obstructions by people. Thus, inspired by the teaching of the Greek historian Herodotus, thinking about the past to understand the present and idealizing the future, this work is presented through metrics between FaceNet and MobileNetV2 when exposed to the WiseNet dataset facing the problem of re-identification, being that both networks had very close F-Measure performances and accuracy, however when compared to the inference time of the experiment and the training time, FaceNet shows its advantages, in which with these inferred data it hoped to have helped to contribute to future evolutions and improvements in the algorithms of the area.

Keywords: Person Re-Identification, Video Analytics, Evaluation Metric.

1 INTRODUÇÃO

Vezzani (VEZZANI; BALTIERI; CUCCHIARA, 2013) define re-identificação como a tarefa de atribuir o mesmo identificador para todas as instâncias de um indivíduo detectado em uma série de imagens e vídeos, inclusive após uma lacuna significativa de espaço ou tempo. Quando olhado dentro do cenário de segurança, a re-identificação é grande valia para reconhecer e identificar uma pessoa, sendo em cenários reais de investigação essa pessoa pode ser um assaltante em uma empresa, ou até mesmo uma pessoa desaparecida, nesses casos a análise de vídeo chega ser de alta importância para o rumo da investigação.

Utilizando das palavras de Bedagkar-Gala e Shah (2014), o problema de re-identificar pessoas em imagens digitais é considerado desafiador. Mesmo o foco desse trabalho ser em imagens digitais coletadas por câmeras de segurança, a área de re-identificação possui diversas ramificações como rastreamento, recuperação de trajetória, segurança entre outras que surgem a cada dia com a evolução da tecnologia. O problema é considerado ainda sem solução dentre os pesquisadores, entretanto existem diversas técnicas publicadas para a re-identificação de pessoas (BEDAGKARGALA; SHAH, 2014).

Visando a área de vigilância e segurança, hoje o método que é mais aplicado em diversos países é a metodologia manual, em razão de ser uma técnica mais barata, uma vez que não necessite de um software. Nesse cenário utiliza-se de uma pessoa que fará a análise vídeo a vídeo anotando o tráfego das pessoas de modo a identificar se o indivíduo que ele está observando no frame atual já apareceu na cena ou não.

Todavia, quando têm um vídeo de câmera de segurança com pouco tempo de duração e com poucas pessoas trafegando, o trabalho manual não se torna algo complexo, entretanto quando uma pessoa é colocada para analisar um vídeo com diversas pessoas ou com tempo de duração elevado, um pequeno momento de cansaço ou distração pode comprometer a re-identificação (NAZARE; SCHWARTZ, 2016).

Com base nessas informações, esse trabalho apresenta uma análise comparativa baseada em métricas para entender o funcionamento de dois algoritmos de identificação de pessoas por rosto em um cenário multi câmeras e assim compreender quais os prós e contras de cada algoritmo. Com isso, analisando o passado podem contribuir para um avanço científico na área de re-identificação de pessoas.

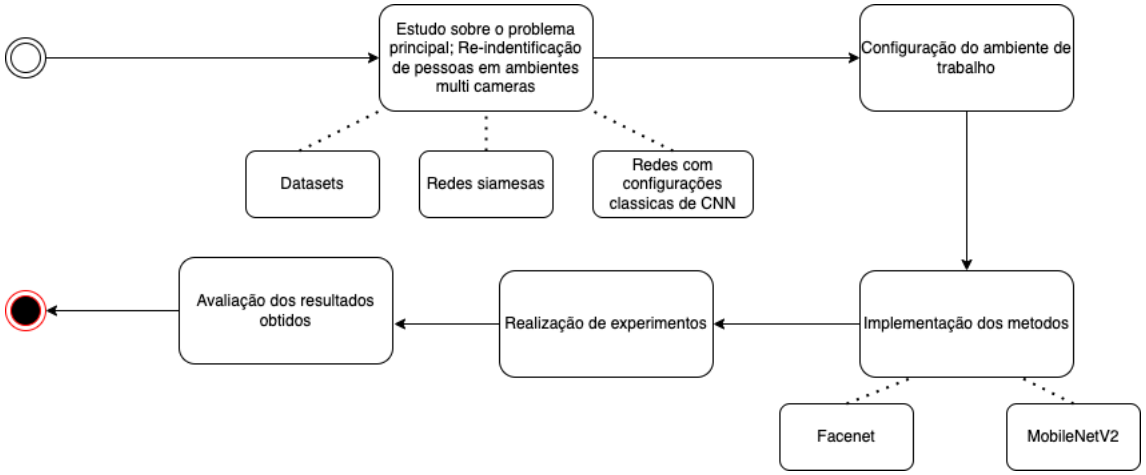
O objetivo principal deste trabalho é construir uma análise de duas arquiteturas que usam técnicas distintas de identificação de pessoas em problemas do tipo de re-identificação, onde a contribuição para área se vem na aplicação de métricas de comparação entre as técnicas voltado a buscar os prós e contras e melhorias em seus algoritmos.

2 FUNDAMENTAÇÃO TEÓRICA

Este trabalho consiste em cinco partes (Figura 1). A primeira fica sendo as pesquisas das arquiteturas utilizadas e datasets de re-identificação de pessoas, a segunda sendo a configuração do ambiente, a terceira a

implementação do código, a quarta a realização dos experimentos e por último seria análise dos resultados obtidos e uma apresentação dos métodos validando os prós e contras junto a seus resultados de suas métricas e assim propondo melhorias em seus algoritmos visando uma contribuição para a área de inteligência artificial.

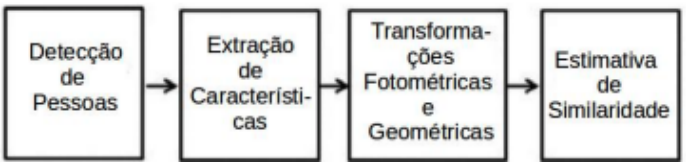
Figura 1 - Etapas adotadas para o desenvolvimento deste trabalho.



Fonte: Autoria própria.

Observando a Figura 2 proposta por Enembreck (ENEMBRECK, 2020), a re-identificação de pessoas possui 4 etapas basicamente, que podem sofrer pequenas alterações em seu fluxo principal dependendo do método aplicado. Inicialmente deve-se detectar a pessoa em cena, em seguida realizar a extração de características que servem para identificar e diferenciar cada pessoa nas imagens, podendo haver a fase de transformações geométricas e por final a estimativa de similaridade da pessoa em cena. No caso deste trabalho está sendo utilizado a identificação da pessoa em nível de reconhecimento facial, deste modo, a extração de características junto a transformações geométricas estaria sob encargo da rede neural e a estimativa de similaridade seria adaptada a dependendo da saída da rede neural utilizada no processo.

Figura 2 – Principais etapas de re-identificação de pessoas em imagens.



Fonte: ENEMBRECK, 2020

Considerando o problema de re-identificação de pessoas em vídeos de câmera de segurança, neste trabalho foram desenvolvidas duas abordagens para tratá-lo. A primeira é uma rede neural siamesa, constituída por duas sub-redes idênticas compostas por uma rede neural convolucional e uma rede autoencoder. A rede neural siamesa tem como função decidir se duas imagens de rostos diferentes referem-se à mesma pessoa, comparando duas imagens

de entrada, através de características extraídas de cada uma pelas sub-redes. A outra técnica desenvolvida é uma MobileNetV2, que é composta pela mesma ideologia da MobileNetV1 (HOWARD, 2017) ao utilizar também a convolução em profundidade como blocos de construção, processo que será detalhado a seguir.

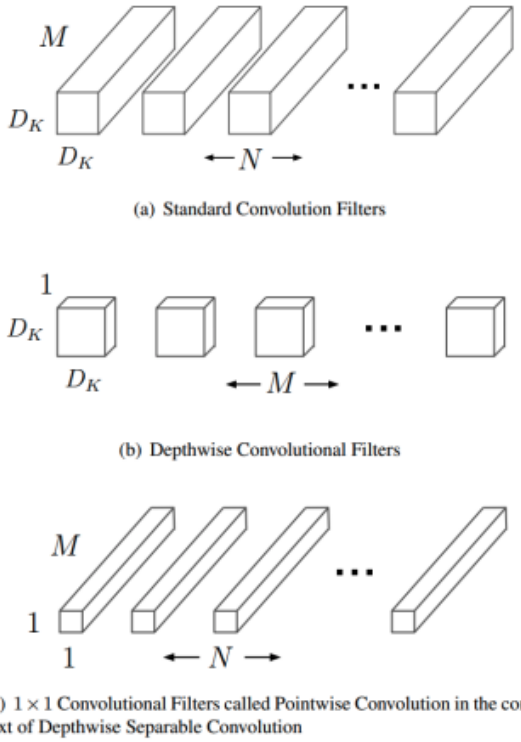
2.1 MobileNet

A MobileNet foi desenvolvida com uma proposta de ser uma Rede Neural Convolucional leve para aplicações móveis e embarcadas de visão computacional (BARROS, 2020). Redes neuronais convolucionais, também conhecidas como CNNs, do inglês Convolutional Neural Networks, são uma classe de redes neurais utilizadas no processamento de dados. Estas redes têm sido bem eficazes em diversos problemas da área de aprendizado de máquina, entre elas temos o destaque para desafios com imagens digitais.

Seu nome vem devido a sua estrutura que é montada em cima da operação matemática de convolução, que é um caso especial das operações lineares (LECUN, 1995). As CNN's são estruturadas por blocos de convolução, que na abordagem clássica sua configuração é a primeira camada sendo uma de entrada e a última uma de saída, e no seu intermédio possuem camadas de convolução, composta por um grupo de filtros, onde retornam um mapa de características, pela camada de agrupamento, responsável por substituir a saída num determinado local da rede e a camada totalmente conectada, tal como o próprio nome indica, interconecta todos os neurónios de uma determinada camada com todos os neurónios da camada seguinte.

Com basicamente essas camadas, foi descrito na literatura diversas organizações e diferentes tipos de agrupamento, sendo que dependendo da arquitetura construída pelo autor é atribuída à rede propriedades de interpretação de características baseadas nos dados de entrada. Seguindo esse conceito surge a MobileNet onde é baseada no conceito de Depthwise Separable Convolution (HOWARD, 2017), que é uma forma de convolução fatorada. Segundo Barros, esse tipo de técnica separa o processo em duas partes: a convolução em profundidade e a convolução pontual, na Figura 3 é possível observar a comparação entre a convolução padrão utilizada por redes como a VGG-16 (UL HASSAN, 2018) e a convolução fatorada da MobileNet.

Figura 3 – A convolução padrão (a) é substituída por duas camadas: convolução em profundidade em (b) e convolução pontual em (c) para construir o conceito de Depthwise Separable Convolution.



(a) Standard Convolution Filters

(b) Depthwise Convolutional Filters

(c) 1×1 Convolutional Filters called Pointwise Convolution in the context of Depthwise Separable Convolution

Fonte: HOWARD, 2017

A convolução em profundidade, executa uma única convolução em cada canal de cor, em vez de combinar os três canais como nas Redes tradicionais. Segundo Howard et al. a convolução em profundidade é muito eficiente em relação à convolução padrão, ela se concentra nas filtragens dos canais de entrada e não os combina para criar novas features.

Na Tabela 1, é possível observar a arquitetura da MobileNetV2, na qual foi utilizada neste trabalho, onde contém uma camada convolucional com 32 filtros, seguida de 19 camadas residuais chamadas bottleneck, Tabela 2.

Tabela 1 – Arquitetura Geral do MobileNetV2, onde t: fator de expansão, c: número de canais de saída, n: número de repetição, s: stride.

Input	Operator	t	c	n	s
$224^2 \times 3$	conv2d	-	32	1	2
$112^2 \times 32$	bottleneck	1	16	1	1
$112^2 \times 16$	bottleneck	6	24	2	2
$56^2 \times 24$	bottleneck	6	32	3	2
$28^2 \times 32$	bottleneck	6	64	4	2
$14^2 \times 64$	bottleneck	6	96	3	1
$14^2 \times 96$	bottleneck	6	160	3	2
$7^2 \times 160$	bottleneck	6	320	1	1
$7^2 \times 320$	conv2d 1x1	-	1280	1	1
$7^2 \times 1280$	avgpool 7x7	-	-	1	-
$1 \times 1 \times 1280$	conv2d 1x1	-	k	-	

Fonte: SANDLER, Mark et al., 2018

Tabela 2 – Bloco residual bottleneck transformando k canais para k', com slide igual a “s” e fator de expansão “t”.

Input	Operator	Output
$h \times w \times k$	1x1 conv2d , ReLU6	$h \times w \times (tk)$
$h \times w \times tk$	3x3 dwise s=s, ReLU6	$\frac{h}{s} \times \frac{w}{s} \times (tk)$
$\frac{h}{s} \times \frac{w}{s} \times tk$	linear 1x1 conv2d	$\frac{h}{s} \times \frac{w}{s} \times k'$

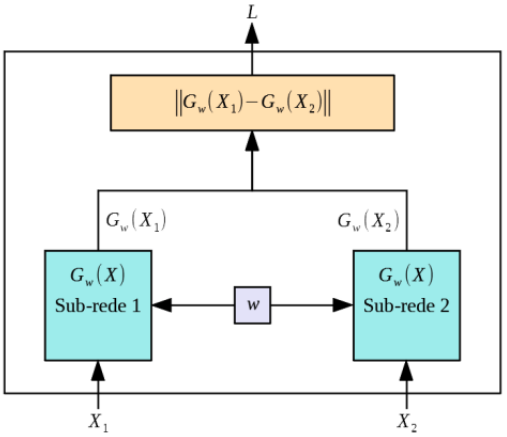
Fonte: SANDLER, Mark et al., 2018

2.2 FaceNet

Por muito tempo o estado da arte de problemas de reconhecimento de rosto, foram redes neurais do tipo convolucionais clássicas, projetadas para processar sinais em 2 dimensões, como imagens, através de operações de convolução, pooling e ativação, entre esse tipo de redes tem a VGG-16 (UL HASSAN, 2018), ResNet-50 (HE, 2018), Inception (SZEGEDY, 2016) e a própria MobileNet.

Entretanto, nos últimos anos começa a surgir um outro paradigma de redes neurais que difere das CNN 's com abordagens clássicas, principalmente quando comparado suas últimas camadas, as chamadas redes neurais siamesas. Segundo Enembreck (ENEMBRECK, 2020), na Figura 4 é apresentado um exemplo de rede siamesa, sendo X_1 e X_2 um par de imagens de entrada, recebem um rótulo binário Y para identificar se são semelhantes ou não, o conjunto de funções $G(X)$, um vetor de parâmetros W compartilhado e $G_w(X_1)$ e $G_w(X_2)$o mapeamento das entradas dadas X_1 e X_2, respectivamente. Com isso, a rede neural siamesa durante o seu treinamento busca por um valor do parâmetro W que encontre uma menor diferença entre X_1 e X_2.

Figura 4 - Macroestrutura de uma Rede Neural Siamesa.



Fonte: ENEMBRECK, 2020

Na abordagem inicial das redes siamesas a Equação 1, é utilizada para calcular a diferença entre as saídas da sub-redes na qual é passada para uma função de perda contrastiva (BROMLEY et al., 1994; KOCH; ZEMEL; SALAKHUTDINOV, 2015).

$$E_w = \left| G_w(X_1) - G_w(X_2) \right| \quad (1)$$

Entretanto com o avanço dos estudos das redes siamesas, começaram a imergir redes que utilizam ao invés da equação de perda contrastiva a função de erro triplo, e por sua vez essas acabaram conquistando resultados bem expressivos em diversas competições de dataset, a arquitetura Deepface (TAIGMAN, 2014) alcançou 97.35% de acurácia no dataset Labeled Faces in the Wild (HUANG, 2008), a VGG-Face (SIMONYAN, 2013) e a FaceNet (SCHROFF, 2015) alcançaram acurácia de 98.95% e 99.63%, respectivamente, no Labeled Faces in the Wild.

Para validar a proposta das redes siamesas para o problema de re-identificação de pessoas em um ambiente multi-câmeras, este trabalho optou em adotar a arquitetura da FaceNet, por principalmente dois motivos sendo o primeiro na qual segundo Alencar conseguem angariar resultados extremamente satisfatórios, mapeando imagens faciais em um espaço euclidiano compacto (ALENCAR, 2020) e o segundo foi pelo fato de que a FaceNet é gera um vetor de caracterısticas $X \in \mathbb{R}^{128}$ (BIESSECK, 2021), o que representa um baixo custo de espaço para armazenamento.

2.3 Dataset

Para a realização desse estudo foi utilizado o dataset WiseNET (MARROQUIN et al., 2019). Esse conjunto de dados é utilizado para detecção e rastreamento de pessoas, composto por 6 câmeras internas no terceiro andar com informações contextuais e anotações, do prédio Institut Marey et Maison de la Metallurgie(I3M) localizado em Dijon, França, conforme apresentado nas Figura 5 e 6 (MARROQUIN et al., 2019).

Figura 5- Mapa de localização das câmeras de segurança da WiseNET.

Fonte: MARROQUIN et al., 2019

Figura 6 - Exemplos de imagens da WiseNET.

Fonte: MARROQUIN et al., 2019

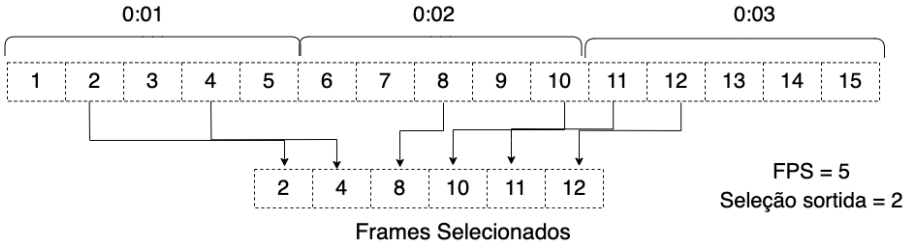
Os conjuntos de vídeos foram gravados com 5 a 6 câmeras simultaneamente em um total de 11 conjuntos distintos. Os vídeos capturaram diferentes ações humanas, como andar, ficar em pé, sentar, ficar imóvel, entrar, sair de um espaço.

2.4 Protocolo de avaliação

Para melhor entender o processo de avaliação deste trabalho, é necessário contextualizar que o problema em questão de re-identificação de pessoas por identificação facial, possui duas redes neurais sendo executadas em conjunto. Sendo a primeira uma rede focada em localizar rosto em imagens que se trata de um problema de reconhecimento, e a segunda uma rede que busca fazer a identificação facial para num terceiro momento realizar o rastreamento dessa pessoa.

Como o foco do trabalho é realizar um relatório de rastreamento das pessoas durante um período de tempo, baseado em mais de um vídeo de segurança e ainda ser um sistema que possui valia para organizações com poucos recursos, o autor optou por realizar uma estratégia de seleção sortida de frames baseado no tempo em segundo de vídeos, isso é, o dataset WiseNet foi gravado num fps de 30, ao invés de processar os 30 frames foi realizado uma seleção sortida desses frames voltado a diminuir a quantidade de frames observados e assim melhorar o relatório final de rastreamento baseado em tempos em segundos, na Figura 7 é possível observar um exemplo da metodologia aplicada, na qual o FPS é 5, a seleção sortida é igual a 2 e os números dentro dos retângulos é a representação dos index de cada frame de um vídeo de 3 segundos.

Figura 7 – Diagrama da seleção sortida de frames.

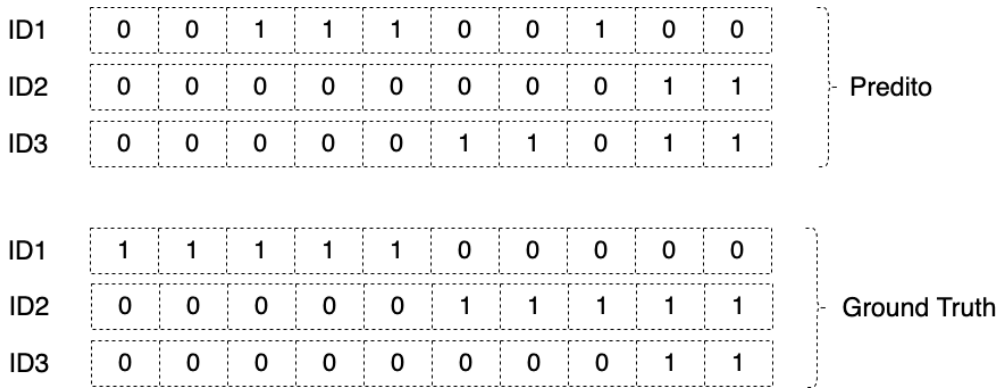


Fonte: Autoria própria.

Um dos desafios desse projeto foi como realizar a avaliação da rede de reconhecimento, uma vez que o dataset WiseNet, não possui o mapeamento das caixa delimitadora dos rosto durante a gravação para então gerar o grau de assertividade da rede, em razão disso foi realizada uma métrica por aproximação de resultados, isso é, para cada frame analisado em questão foi verificado quantos rosto era para ter naquele frame e quantos rostos a rede de identificação fez o mapeamento. Mesmo não sendo a melhor estratégia para medir e avaliar uma rede de reconhecimento de objetos, a segunda métrica que irá avaliar o resultado da rede subjacente de identificação supre os pontos fracos de uso da métrica por aproximação em uma rede de reconhecimento.

Para a rede neural de identificação facial, foi utilizado a abordagem de matriz de confusão multiclasses, sendo que para cada rosto analisado pela rede ela possui N +1 possibilidades de respostas, sendo N a quantidade de rosto de pessoas que a rede foi treinada mais uma classe de indivíduo desconhecido. Para gerar a matriz de confusão foi abordado o problema em nível de frame, Figura 8, sendo que para cada frame do ground truth, foi verificado se houve a predição correta no mesmo frame da pessoa correta, caso não tivesse ocorrido a predição certa, era avaliado o mesmo frame para as demais predições, Na Figura 9 é possível ver como ficaria a matriz de confusão da Figura 8.

Figura 8 – Array de frames, sendo 1 quando tem aquele usuário naquele frame e 0 quando não tem.



Fonte: Autoria própria.

Figura 9 – Matriz de confusão multi classes.

	ID1	ID2	ID3
ID1	3	1	0
ID2	0	2	0
ID3	0	2	2

Fonte: Autoria própria.

Para fazer a verificação da eficácia dos modelos da MobileNetV2 e da Facenet foi utilizada a F-medida e Acurácia, Tabela 3. A F-medida utiliza as métricas de precisão (Pi) e revocação (Ri), onde i representa a classe recorrente, as quais podem ser obtidas por meio de outras métricas extraídas da matriz de confusão multiclasses. A F-medida é definida como uma média harmônica entre a precisão (Pi) e revocação (Ri), Tabela 3.

Tabela 3 – Tabela com as métricas utilizadas juntamente com suas fórmulas.

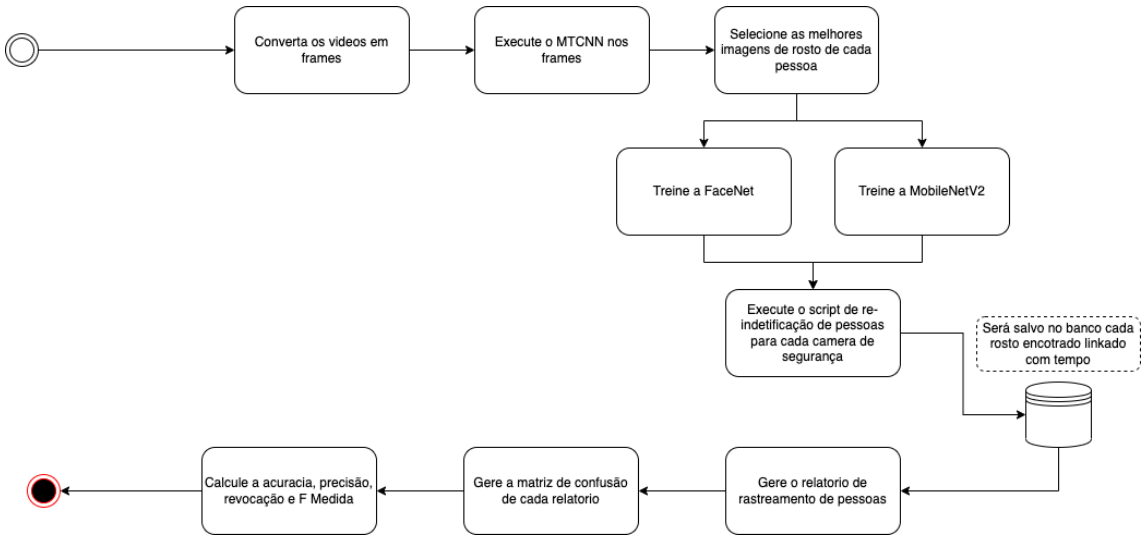
Métrica	**Fórmula**
F-medida	$F1_i = \left(2P_iR_i\right)/\left(P_i + R_i\right)$
Acurácia	$Acc = \sum_i MC_{ii}/total$
Precisão	$P_i = \left(MC_{ii}\right)/\sum_j MC_{ij}$
Revogação	$R_i = \left(MC_{ii}\right)/\sum_j MC_{ji}$

Fonte: Autoria própria.

2.5 Experimentos

Como dito anteriormente, o propósito deste projeto é corroborar com diversas pesquisas para com organizações que possuem pouco recursos computacionais, entretanto que necessitam de algoritmos de re-identificação de pessoas. Olhando esse contexto junto ao ambiente multi-câmeras do problema em questão, o algoritmo foi estruturado para executar em threads em paralelo com mais de um vídeo de câmera de segurança, e ao mesmo tempo fazendo a identificação das pessoas por reconhecimento facial e mapeando suas localizações em um banco de dados local, após a execução do algoritmo principal é iniciado um script na qual faz a interpretação das pessoas e gera o relatório de tracking dos vídeos, ver Figura 10.

Figura 10 – Fluxo de execução do algoritmo.



Fonte: Autoria própria.

Para executar e avaliar os experimentos foi utilizado uma máquina com as seguintes configurações de hardware e software:

- GPU – Nvidia GeForce 930MX.
- CPU – Intel Core i7 -7600U.
- Memória RAM – 16Gb.
- Sistema Operacional Ubuntu 20.04 LTS.

Vale destacar que a escolha da biblioteca tensorflow foi escolhida devido ser uma ferramenta de código aberto desenvolvido e mantido pelo Google, na qual possui uma documentação rica e abrangente facilitando a construção e manipulação de tensores dentro dos modelos de redes neurais (ABADI, 2018), enquanto a biblioteca Keras foi escolhida devido sua grande facilidade no desenvolvimento e treinamento de redes neurais tanto em nível de CPU ou GPU além do grande suporte da comunidade para o desenvolvimento de softwares open-sources fazendo com que o conhecimento seja cada vez mais compartilhado (CHOLLET, 2018).

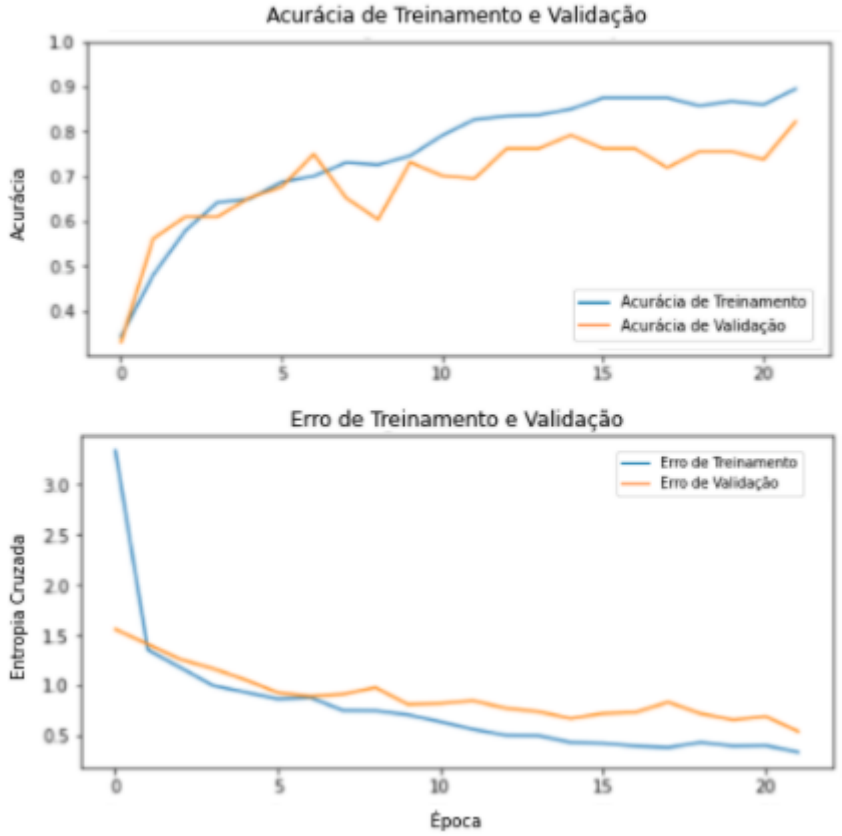
3 RESULTADOS

Com base no contexto e metodologia descrita nos capítulos anteriores, a execução dos experimentos se fez através de 3 conjuntos de vídeos do Dataset WiseNet, na qual cada conjunto é referente a um dia de gravação com 5 câmeras de segurança do mesmo período de tempo. Desse modo, foi utilizado os outros 8 conjuntos encontrados no WiseNet para gerar o Dataset de treinamento da FaceNet e MobileNetV2, devido a pouca variância de posições dos rosto entre as pessoas que foram utilizadas no treinamento, foi selecionado 100 imagens dos rosto de cada pessoa, sendo 5 classes referente

a cada pessoa, referenciado dos resultados sendo ID1, ID3, ID12, ID15 e ID16, e uma classe sendo de pessoa desconhecida, também chamado nos resultados com UNK.

Durante o treinamento foi notado algumas diferenças expressivas entre as redes de FaceNet e MobileNetV2, na qual para 6 classes com 100 imagens para cada a rede siamesa teve um tempo de treinamento de 13 segundos enquanto a rede de abordagem clássica de CNN teve um tempo de 342 segundos, já a FaceNet teve uma acurácia no conjunto de teste de 92% enquanto a Mobilenet teve uma acurácia de 83%, veja Figura 11, aproximadamente 10% a menos e aproximadamente 26 vezes a mais do tempo de treinamento da Facenet.

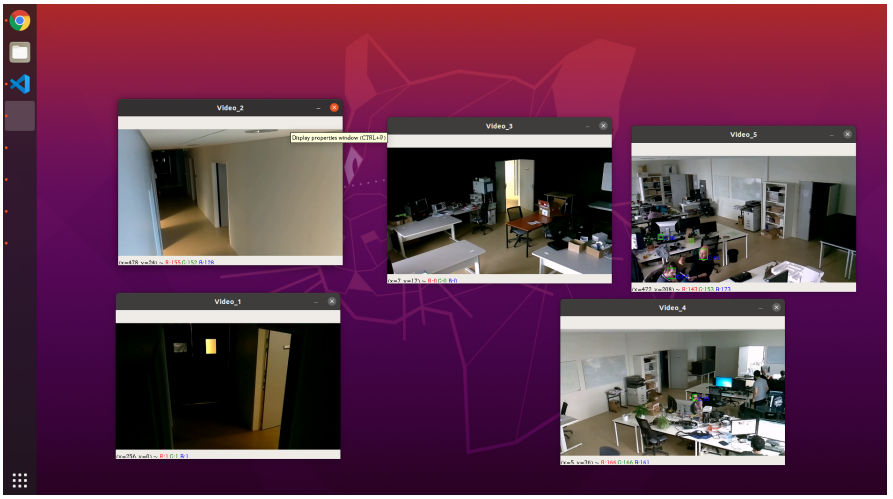
Figura 11 – Acurácia e Loss durante o treinamento e validação da MobileNetV2.



Fonte: Autoria própria.

Após os modelos treinados, foi executado o script de re-identificação de pessoas, na qual cada experimento se baseia na execução paralela das 5 câmeras de segurança de cada conjunto de vídeo, Figura 12. Mesmo cada vídeo sendo executado de forma otimizada por Threads, a velocidade de execução fica atrelada ao tempo de inferência de cada modelo, e assim acarretando no tempo total de execução script.

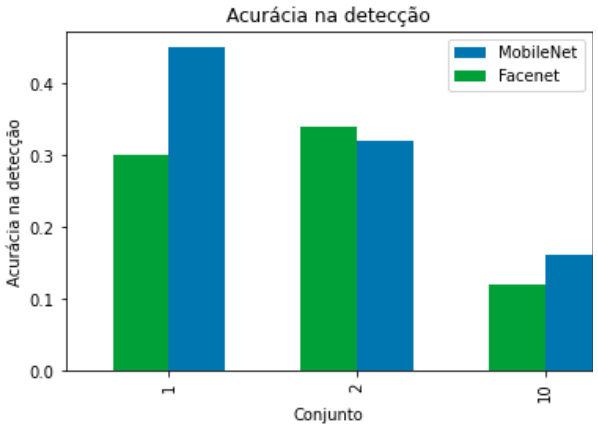
Figura 12 – Execução da re-identificação de pessoas com 5 vídeos de câmeras em paralelo.



Fonte: Autoria própria.

Desse modo, foi notado uma diferença nos tempos de cada modelo para os conjuntos 1, 2 e 10 que possuem 60, 118 e 40 segundos e tiveram para FaceNet o tempo de 972, 1636 e 230 segundos e MobileNetV2 1059, 1916 e 493 segundos, aproximadamente 15 vezes o tempo de vídeo para FacaNet e 17 vezes o tempo do vídeo para a MobileNetV2, entretanto vale destacar o outlier no tempo do conjunto 10, isso se dá devido a baixa taxa de acurácia do modelo de detecção utilizado no sistema, no caso a MTCNN, Figura 13.

Figura 13 – Representação da acurácia do modelo de detecção de rosto por cada conjunto de vídeos de segurança.



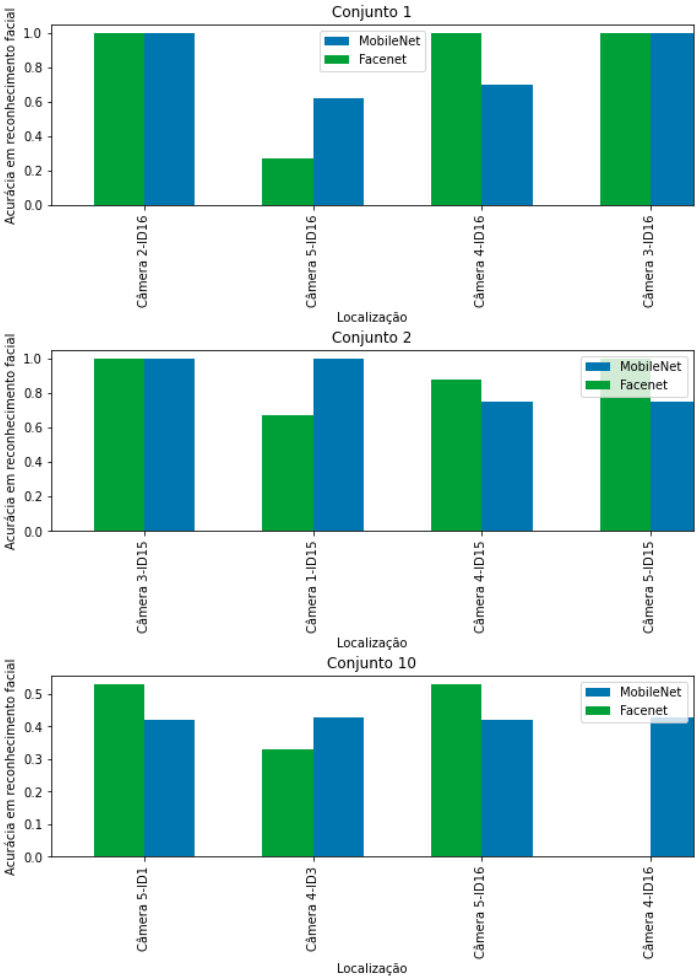
Fonte: Autoria própria.

Mesmo a acurácia do modelo de detecção ter sido metade no conjunto 10, vale destacar que em todos os 3 conjuntos a taxa está bem baixa de um esperado por um sistema que deverá realizar o rastreamento de pessoas, desse modo durante o experimento foi notado diversos casos na qual a pessoa passou pelas câmeras com rosto sem obstruções, entretanto a MTCNN não conseguiu fazer seu reconhecimento, fazendo com que o sistema como um

todo não chegue no seu melhor potencial. Vale destacar que para a FaceNet foi utilizado a MTCNN pré-treinada disponibilizada pelo autor do código da FaceNet David Sandberg (SANDBERG, 2017), enquanto para a Mobilenet Foi utilizado a rede pré-treinada oferecida pela biblioteca mtcnn (MTCNN, 2019).

Entretanto mesmo com a baixa acurácia do modelo de detecção, os modelos de reconhecimentos tiveram resultados interessantes, conforme apresentado na Figura 14, sendo que para todos frames que foram detectados pela MTCNN que iam para a FaceNet ou para MobileNetV2 foi registrados altos valores de acurácias, sendo que para o conjunto 1 a Facenet obteve 3 cenários com 100% de acurácia e a Mobilnet obteve 2 cenários com 100%.

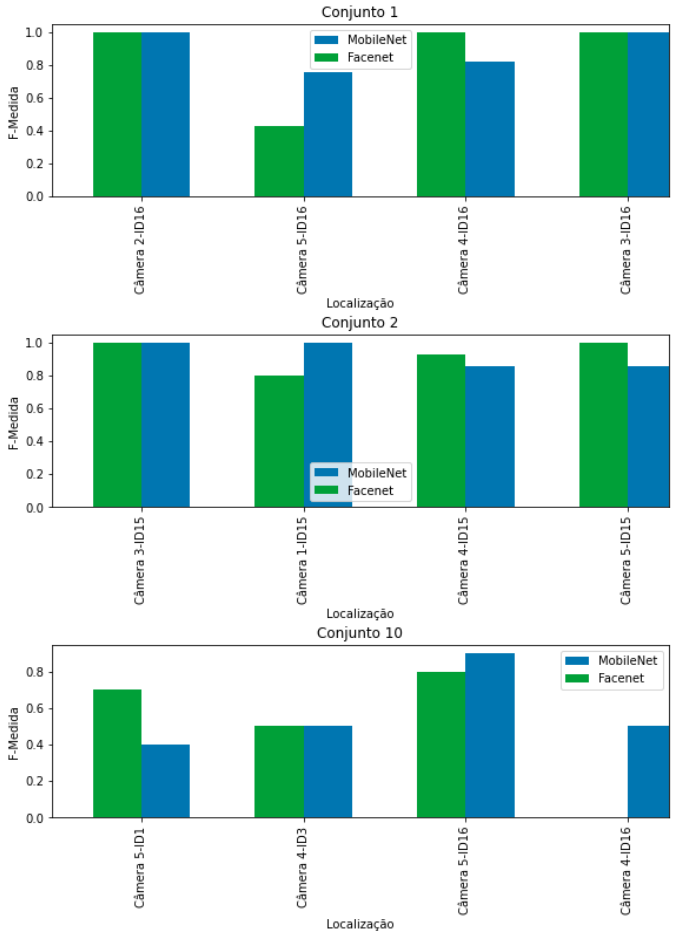
Figura 14 – Representação da acurácia do modelo de reconhecimento de rosto por cada pessoa e espaço e conjunto de vídeos.



Fonte: Autoria própria.

Na Figura 15 é apresentado a F-Medida de cada cenário, desse modo quando relacionamos com a Figura 14 nota-se que em casos que a acurácia foi alta a F-Medida também está com valores elevados, indicando que ocorreu poucas interpretações errôneas de pessoas nos frames analisados. Com isso conclui-se que a utilização dessas redes são válidas para resolver problemas de re-identificação de pessoas e gerar um relatório de rastreamento.

Figura 15 – Representação da F Medida do modelo de reconhecimento de rosto por cada pessoa e espaço e conjunto de vídeos.



Fonte: Autoria própria.

Outros dois pontos são relevantes, sendo o primeiro a baixa taxa de acurácia no conjunto 10 para os dois modelos e a exclusão de apenas a MobileNetV2 conseguir reconhecer o usuário 16 no espaço 4. Para o primeiro ponto, pode ser levantado dois questionamentos sendo, qual a correlação e limite entre taxa de acurácia baixa do modelo de detecção para ele começar a afetar a taxa de sucesso do sistema, e em segundo questionamento, o vídeo 10 é aquele que possui mais pessoas trafegando pelo espaço, qual o limite entre pessoas em um mesmo ambiente que pode acabar atrapalhando a viabilidade desse tipo de solução. Para o segundo ponto, foi notado que o ID16 durante as imagens de treinamento estava de touca, enquanto durante a gravação do video do conjunto 10 ele estava sem, sendo que justamente a câmera do espaço 4 quando aparece o usuario 16, ele está longe da câmera com bastante ruído luminoso, diferente da câmera do espaço 5, que ele está mais perto e com menos ruído luminoso, desse modo é possível levantar o questionamento de qual o limite de interpretação de variações de características que a rede siamesa consegue aprender, uma vez que ela utiliza o espaço latente de treinamento como referência.

4 CONCLUSÃO

Este trabalho teve como principal objetivo desenvolver e comparar duas arquiteturas de redes neurais para um problema de re-identificação de pessoas em ambientes multi câmeras utilizando algoritmos que não depende de grande poder de processamento para serem executados. Uma das principais contribuições foi a construção de uma análise apresentando as vantagens e desvantagens de cada técnica perante o ambiente multi câmeras.

Como trabalhos futuros, será avaliado a troca do modelo de detecção que foi apresentado pelo projeto na qual para o experimento tiveram resultados insatisfatórios, além da investigação dos questionamentos levantados pelo autor durante os resultados.

Assim sendo, em razão do número de circuito de câmeras de vigilância aumentando cada vez mais como consequência da crescente preocupação com segurança (DE CARVALHO PRATES et al., 2016) o problema de procurar onde uma pessoa se movimenta antes, durante e depois do acontecimento (SALAMON et al., 2015) vai possuir campo para expandir.

REFERÊNCIAS

ABADI, M. et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Software available from tensorflow.org. Disponível em: . Acesso em: 15 out. 2018.

ALENCAR, Raphael Brito; BEZERRA, Byron Leite Dantas. Sistema de comparação de imagens de faces, em múltiplas resoluções, baseado em Redes Neurais Siamesas. **Revista de Engenharia e Pesquisa Aplicada**, v. 5, n. 1, p. 50-57, 2020.

BARROS, Edna. Gabriela Alves Rodrigues. Reconhecimento de Emoções utilizando Redes Neurais Convolucionais para Auxiliar no Tratamento de Crianças com Autismo, 2020.

BEDAGKAR-GALA, Apurva; SHAH, Shishir K. A survey of approaches and trends in person re-identification. **Image and vision computing**, v. 32, n. 4, p. 270-286, 2014.

BIESSECK, Bernardo Janko Gonçalves; ZACARKIM, Valber Lemes. Avaliação da CNN FaceNet para reconhecimento facial de estudantes em sala de aula. **Brazilian Journal of Development**, v. 7, n. 3, p. 27558-27563, 2021.

BROMLEY, Jane et al. Signature verification using a “siamese” time delay neural network. **International Journal of Pattern Recognition and Artificial Intelligence**, v. 7, n. 04, p. 669-688, 1993.

CHOLLET, F. et al. Keras. 2015. Disponível em: . Acesso em: 15 out. 2018.

DE CARVALHO PRATES, Raphael Felipe et al. Correspondência entre pessoas em uma rede de câmeras de vigilância. 2019.

ENEMBRECK, Fábia Isabella Pires et al. **Re-Identificação de pessoas em imagens digitais utilizando redes neurais siamesas e triplet baseadas em uma rede neural convolucional e um autoencoder**. 2020. Dissertação de Mestrado. Universidade Tecnológica Federal do Paraná.

HADSELL, Raia; CHOPRA, Sumit; LECUN, Yann. Dimensionality reduction by learning an invariant mapping. In: **2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)**. IEEE, 2006. p. 1735-1742.

HE, Kaiming et al. Deep residual learning for image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. 2016. p. 770-778

HOWARD, Andrew G. et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. **arXiv preprint arXiv:1704.04861**, 2017.

HUANG, Gary B. et al. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: **Workshop on faces in'Real-Life'Images: detection, alignment, and recognition**. 2008.

KOCH, Gregory et al. Siamese neural networks for one-shot image recognition. In: **ICML deep learning workshop**. 2015.

LECUN, Yann et al. Convolutional networks for images, speech, and time series. **The handbook of brain theory and neural networks**, v. 3361, n. 10, p. 1995, 1995.

NAZARE JR, Antonio C.; SCHWARTZ, William Robson. A scalable and flexible framework for smart video surveillance. **Computer Vision and Image Understanding**, v. 144, p. 258-275, 2016.

MARROQUIN, Roberto; DUBOIS, Julien; NICOLLE, Christophe. WiseNET: An indoor multi-camera multi-space dataset with contextual information and annotations for people detection and tracking. **Data in brief**, v. 27, p. 104654, 2019.

MTCNN face detection implementation for TensorFlow, as a PIP package. (n.d.). GitHub. https://github.com/ipazc/mtcnn. 2019

SANDBERG, David. Facenet: Face recognition using tensorflow. **cit. on**, p. 30, 2017.

SALAMON, Nestor Ziliotto et al. Re-identificação de pessoas em imagens através de características descritivas de cores e grupos. 2015.

SANDLER, Mark et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. 2018. p. 4510-4520.

SCHROFF, Florian; KALENICHENKO, Dmitry; PHILBIN, James. Facenet: A unified embedding for face recognition and clustering. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. 2015. p. 815-823.

SIMONYAN, Karen et al. Fisher vector faces in the wild. In: **BMVC**. 2013. p. 4.

SZEGEDY, Christian et al. Rethinking the inception architecture for computer vision. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. 2016. p. 2818-2826.

TAIGMAN, Yaniv et al. Deepface: Closing the gap to human-level performance in face verification. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. 2014. p. 1701-1708.

VEZZANI, Roberto; BALTIERI, Davide; CUCCHIARA, Rita. People reidentification in surveillance and forensics: A survey. **ACM Computing Surveys (CSUR)**, v. 46, n. 2, p. 1-37, 2013.

UL HASSAN, Muneeb. VGG16-Convolutional Network for Classification and Detection. **en línea].[consulta: 10 abril 2019]. Disponible en: https://neurohive. io/en/popular-networks/vgg16**, 2018.