

# Um comparativo entre arquiteturas de redes neurais profundas aplicadas no reconhecimento de imagens

Leandro R. L. Pavão, Tatiana F. M. dos Santos, Carlos O. Rolim

Universidade Regional Integrada do Alto Uruguai e das Missões (URI)  
Departamento de Engenharias e Ciência da Computação – Santo Ângelo, RS - Brasil.  
leo.pav@hotmail.com, tatiana@san.uri.br ober@san.uri.br

**Abstract.** *Computational solutions are being used to perform the recognition of people, vehicles, and objects through image analysis. In this context, deep neural networks have been shown to be a valuable tool since they can learn about input data and find the solution to a problem almost autonomously. However, due to the diversity of existing types and architectures, it is difficult for developers to choose the most suitable for their application. Thus, this article has the purpose of analyzing two deep neural network architectures, MobilenetV1 and InceptionV3, to demonstrate which has the better accuracy to recognize images of different categories. After performing experiments, we can conclude that the Inceptionv3 network has better performance than MobiliNetV1.*

**Keywords:** *Deep Learning, Neural networks, Machine Learning.*

**Resumo.** *Cada vez mais estão sendo empregadas soluções computacionais para efetuar o reconhecimento de pessoas, veículos e objetos através da análise de imagens. Nesse contexto, as redes neurais profundas têm se demonstrado como valiosa ferramenta uma vez que elas são capazes de aprender a respeito dos dados de entrada e encontrar de forma quase autônoma a melhor solução para um problema. Entretanto, devido a diversidade de tipos e arquiteturas existentes, torna-se difícil para desenvolvedores escolherem qual a mais adequada para sua aplicação. Dessa forma, esse artigo tem por finalidade efetuar a análise de duas arquiteturas de redes neurais profundas, MobilenetV1 e a InceptionV3, com vistas a demonstrar qual possui maior acurácia no reconhecimento de imagens para diferentes categorias. Após a execução de experimentos, pode-se concluir que a rede Inceptionv3 possui melhor performance no reconhecimento de imagens que a MobiliNetV1.*

**Palavras-chave:** *Aprendizagem Profunda, Redes Neurais, Aprendizagem de Máquina.*

## 1. Introdução

Uma das grandes intenções do homem é conseguir criar uma máquina capaz de realizar tarefas sem a necessidade do controle humano, compreendendo os dados disponíveis com a finalidade de convertê-los em informações úteis. Neste contexto, as redes neurais profundas vêm sendo estudadas pela comunidade científica como uma ferramenta

importante para reconhecer padrões em dados com grandes proporções em aplicações científicas bem como na resolução de problemas cotidianos.

Um exemplo para solução de problemas cotidianos encontra-se no trabalho de (Einloft, 2016), cujo objetivo é o desenvolvimento de um modelo, baseado em visão computacional, para facilitar a navegação de deficientes visuais em ambientes internos. Este trabalho utilizou uma rede neural InceptionV3. Os resultados mostraram que esta rede neural obteve uma acurácia de 98,4% na existência ou não de portas em uma imagem. Já na pesquisa de (Wahed, 2016), foi utilizada a rede Inception3 para detectar a expressão facial em imagens com baixa resolução, onde esta arquitetura se mostrou promissora em situações que necessitam detectar imagens com uma resolução de qualidade baixa.

Uma significativa particularidade das redes neurais profundas é a capacidade de obter o conhecimento dos dados que estão sendo fornecidos, assim aperfeiçoando seu desempenho, adquirindo um novo aprendizado de diferentes tarefas, explorando novas descobertas, podendo se auto organizar quando expostas a problemas simples ou complexos, criando suas próprias soluções (Ferneda, 2006). Com a crescente disponibilidade de recursos computacionais e o aparecimento de bases de dados de imagens com milhões de amostras, as redes neurais profundas estão sendo utilizadas cada vez mais para reconhecer e localizar pessoas, veículos e uma grande diversidade de objetos urbanos através de imagens (Marques, 2016). Entretanto, existe uma dificuldade intrínseca relacionada a escolha da melhor arquitetura de rede neural profunda na identificação de imagens.

Assim, esse artigo apresenta um comparativo entre duas arquiteturas de redes neurais profundas, MobilenetV1 e a InceptionV3, com vistas a demonstrar qual possui maior acurácia no reconhecimento de imagens de diferentes categorias.

Este artigo está organizado da seguinte forma: a seção 2 apresenta informações sobre Aprendizagem Profunda e Redes Convolucionais; na seção 3 é descrita a metodologia empregada; os experimentos realizados são descritos na seção 4 e finalmente as conclusões são apresentadas na seção 5.

## **2. Aprendizagem Profunda**

Com a crescente dificuldade dos problemas a serem tratados computacionalmente, e da grande quantidade de dados produzidos por setores diversos torna-se evidente a necessidade de se ter ferramentas computacionais aprimoradas, mais independentes, assim diminuindo a interferência humana e a dependência de especialistas. Para tal fim essas técnicas devem ser aptas para criar com base na experiência passada, uma hipótese, ou função, e por si só desenvolver a capacidade de solucionar o problema que se foi proposto (Faceli et al, 2011).

A aprendizagem profunda (deep learning) é uma subdivisão de aprendizado de máquina que aventura-se a aprender abstrações de elevado nível em dados, empregando arquiteturas hierárquicas com estruturas que simulam o comportamento do cérebro humano para extrair características dos dados de entrada. É uma abordagem que tem sido largamente aplicada em cenários de uso da inteligência artificial tradicional, como análise semântica, transferência de aprendizagem, processamento de linguagem natural, visão computacional entre outros (Guo et al, 2016).

A aprendizagem profunda emprega estruturas computacionais compostas de muitas camadas de processamento com finalidade de aprender com múltiplos níveis de abstração. Com isso, obtém-se melhorias no processamento de objetos, vídeos, voz e áudio (Lecun et al, 2015). Entre as diversas arquiteturas de aprendizagem profunda, as redes neurais convolucionais vêm alcançando extraordinários resultados em tarefas de processamento de imagens, vídeos, voz e áudio.

## 2.1 Redes Convolucionais

Redes neurais convolucionais (CNN) são redes neurais propostas por Yann LeCun e Yoshua Bengio em 1995 e que recentemente, com a crescente disponibilidade de recursos computacionais, demonstraram efetividade para diferentes tipos de problemas (Bezerra, 2016). As CNNs são redes de aprendizagem supervisionada que extraem propriedades topológicas a partir dos dados de entrada. De certa forma, pode-se considerar que as CNNs são um tipo especial de Multi Layer Perceptron (MLP), porém com muitas camadas e treinadas com o uso de uma variação do algoritmo de back-propagation.

Basicamente, uma CNN consiste em três conjuntos de camadas: camadas convolutivas, camadas de pooling e camadas totalmente conectadas (Figura 1). Segundo Parker (2011) a convolução é uma operação matemática entre duas funções, produzindo uma terceira função, que pode ser interpretada como uma função modificada. No processamento de imagens, onde a imagem é definida como uma função bidimensional, a convolução é útil para detecção de bordas, suavização de imagem, extração de atributos, entre outras aplicações. A convolução de uma imagem pode ser interpretada como o somatório da multiplicação de cada elemento da imagem, junto com seus vizinhos locais, pelos elementos da matriz que representa o filtro de convolução.

As CNNs consistem em um ou mais pares de camadas de convolução e max-pooling. As camadas convolucionais utilizam filtros que acionam pequenos locais de uma imagem, e são repetidos por toda imagem. As camadas de maxpooling geram uma versão com menor resolução das camadas de convolução aplicando a ativação máxima do filtro em diversas posições dentro de uma janela. Assim, é adicionado mais tolerância para regiões específicas de um determinado objeto na imagem (Abdel Hamid et al. 2012). Após a convolução é comum utilizar uma camada que emprega a função ReLU (Retified Linear Unit) como função de ativação e busca inserir não-lineariedade aos dados, transformando todos os valores negativos da saída da convolução em zeros. A camada de pooling a convolução invariante a translação, rotação e shifting (janelamento) e com isso reduz drasticamente a dimensão espacial da entrada. A camada totalmente conectada (fully connected) atua como um classificador. Ela recebe uma entrada da camada anterior (Conv ou ReLU ou Pooling) e fornece um vetor  $N$  dimensional onde  $N$  é o número de classes de prováveis saídas.

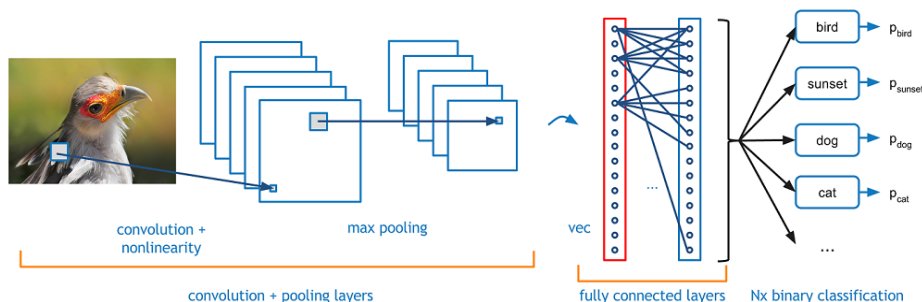


Figura 1. Arquitetura CNN

### 3. Metodologia empregada

Neste trabalho, foram analisadas duas arquiteturas de redes neurais profundas: a MobilenetV1 e a InceptionV3. Segundo Howard et al (2017) a MobileNetV1 (Figura 2a) é uma arquitetura pequena, com baixa latência e que pode ser usada em vários tipos de cenários. Dependendo do seu uso, o tamanho da camada de entrada pode ser maior do que o de saída, pois usa convolução separada em profundidade para construir redes neurais profundas leves, permitindo a escolha do tamanho do modelo onde vai ser aplicado, tendo em vista as limitações do problema. Nessa arquitetura a convolução é aplicada em um único filtro nas entradas objetivando filtrar e combinar os dados em novas saídas em uma única etapa. Esse processo reduz radicalmente o tamanho do modelo.

A arquitetura Inception (Figura 2b) foi desenvolvida com base na GoogLeNet e na abordagem baseada no córtex visual em primatas, podendo suportar várias escalas. A adaptação de rede é um critério importante dessa arquitetura pois aumenta a eficiência da rede ao mesmo tempo que reduz o uso de recursos computacionais. Sua estrutura possui uma camada de agrupamento de filtro com tamanho de 5x5x3, uma camada com 128 filtros para redução de dimensão e ativação linear retificada, possui camada totalmente conectada com 2048 unidades e ativação linear retificada e uma camada dropout com proporção de 70% das saídas para evitar a sobre adaptação aos dados (*overfitting*). A camada totalmente conectada, está completamente ligada com a camada anterior e funciona da mesma maneira que as redes neurais tradicionais (LeCun, 2015).

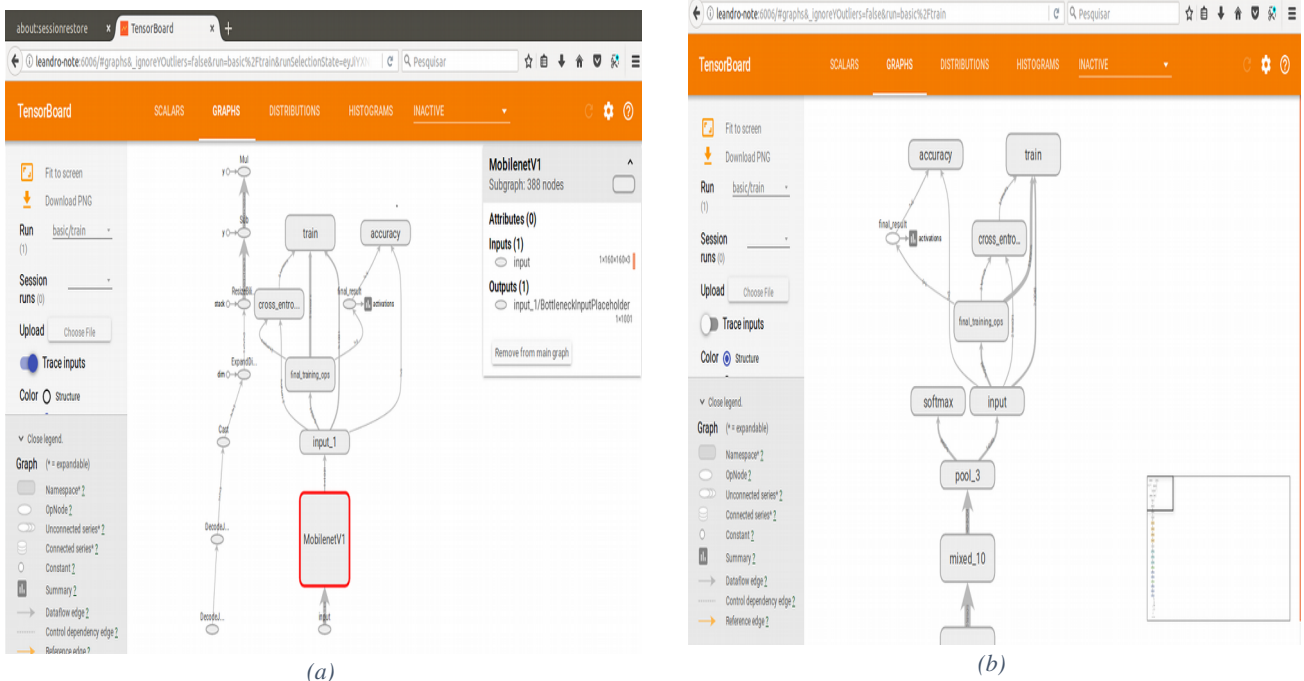


Figura 2. Arquitetura das rede: (a) MobiliNetV1 (b). Arquitetura InceptionV3

Para treinamento e testes das redes, criou-se um banco de com imagens classificadas nas categorias de flores, separadas em rosas, margaridas, dente-de-leão, girassóis e tulipas, imagens de comidas como hambúrguer, batata frita e pizza. Também foram acrescentadas imagens de resíduos sólidos. Para o treinamento da rede, foram geradas 5.437 imagens, sendo 600 como batata frita, 401 de pizza, 719 de hambúrguer, 898 de dente de leão, 699 de girassóis, 633 de margaridas, 641 de rosas, 799 de tulipas e 36 de resíduos sólidos. Além disso, gerou-se mais 50 imagens diferentes para compor o banco de teste. As imagens utilizadas para realizar o treinamento e testes das redes neurais classificou-se e se obteve via internet, sendo que as mesmas encontram-se nos formatos jpg e jpeg com um tamanho de 280 x 180 pixels em média.

Nos experimentos, a taxa de aprendizagem foi definida em 0.01, a etapa de treinamento deteve-se na faixa de 500 interações, a porcentagem de imagens a serem usadas como um conjunto de teste foi definido em 10%, e a porcentagem de imagens usadas como um conjunto de validação em 10%. A frequência para avaliar os resultados do treinamento foi de 10% e a quantidade de imagens para treinar o conjunto de imagens de cada vez em 100 vezes.

A métrica de avaliação do modelo nos dados de treinamento foi a acurácia que é a proximidade entre o valor obtido experimentalmente e o valor verdadeiro na medição de uma grandeza física. Para avaliar a performance do modelo nos dados de teste, a métrica utilizada foi a taxa de erro e o modelo mais eficiente demonstrou-se aquele com a melhor capacidade de generalização em novos dados, ou seja, aquele que possui a menor taxa de erro.

Os experimentos foram realizados usando as arquiteturas InceptionV3 e MobiliNetV1 implementados com o TensorFlow versão 1.4. Utilizou-se para realizar os testes e treinamento da rede neural uma CPU com processador Intel Core i3 CPU M 370 2,4 GHz, memória de 4 GB 1600 MHz DDR3 e placa de vídeo Intel HD Gráficos.

## **4. Experimentos realizados**

### **4.1 Etapa de treinamento**

A primeira etapa foi o treinamento das redes. A Figura 3a apresenta os resultados de acurácia referentes ao treinamento da rede neural InceptionV3. Foi realizado o treinamento com 500 interações usando 5.437 imagens do banco de dados, com taxa de aprendizado em 0.01. A linha laranja demonstra a etapa de treinamento da rede, enquanto a linha azul representa a validação desse treinamento. Nas primeiras interações a taxa de aprendizado da rede foi baixo, melhorando a partir das cinquenta primeiras interações, chegando em momentos que obteve oscilações na etapa de treinamento e validação devido ao tipo de dado que fora passado para rede, pois as imagens não possuíam o mesmo tamanho em pixel, mesmo assim a rede obteve uma precisão no treinamento em torno de 95% e a sua validação em torno de 90%. A Figura 3b apresenta o erro da rede. Como percebe-se no início do treinamento a margem de erro é muito alto nas primeiras cinquenta interações. Com o decorrer do treinamento, a taxa de erro fica próxima de

zero indicando que a rede convergiu, isso é, conseguiu aprender sobre os dados da entrada.<sup>1</sup>

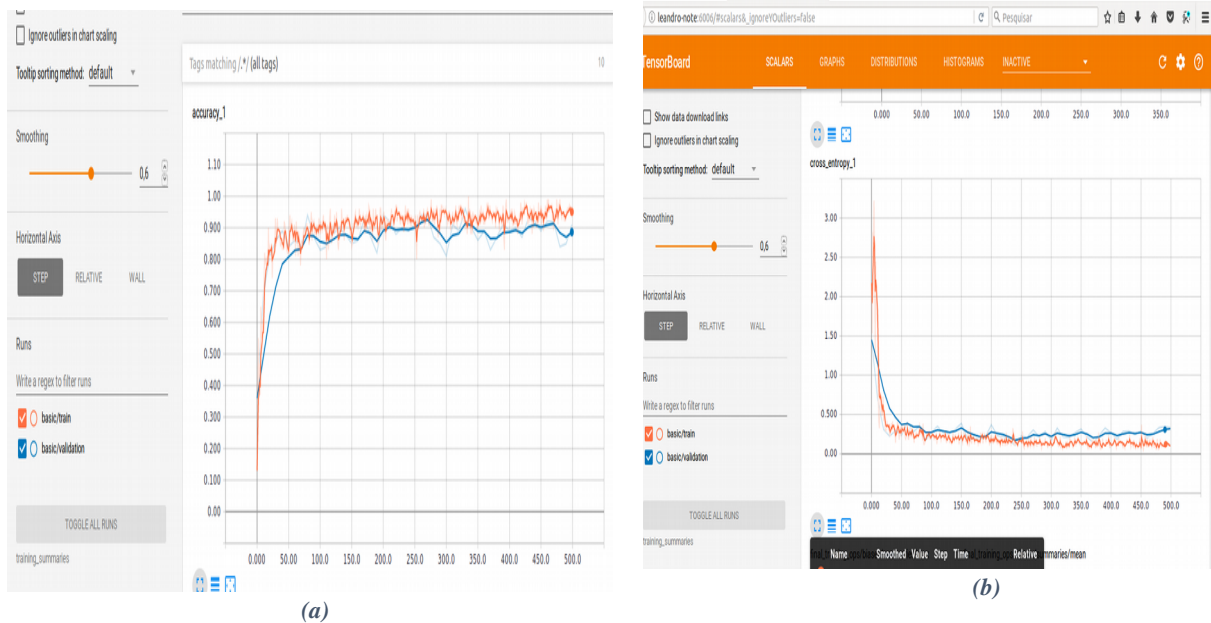


Figura 3. Treinamento da InceptionV3: (a) acurácia (b) erro

A arquitetura MobiliNetV1 foi submetida ao mesmo processo de treinamento e os resultados obtidos são apresentados na Figura 4.

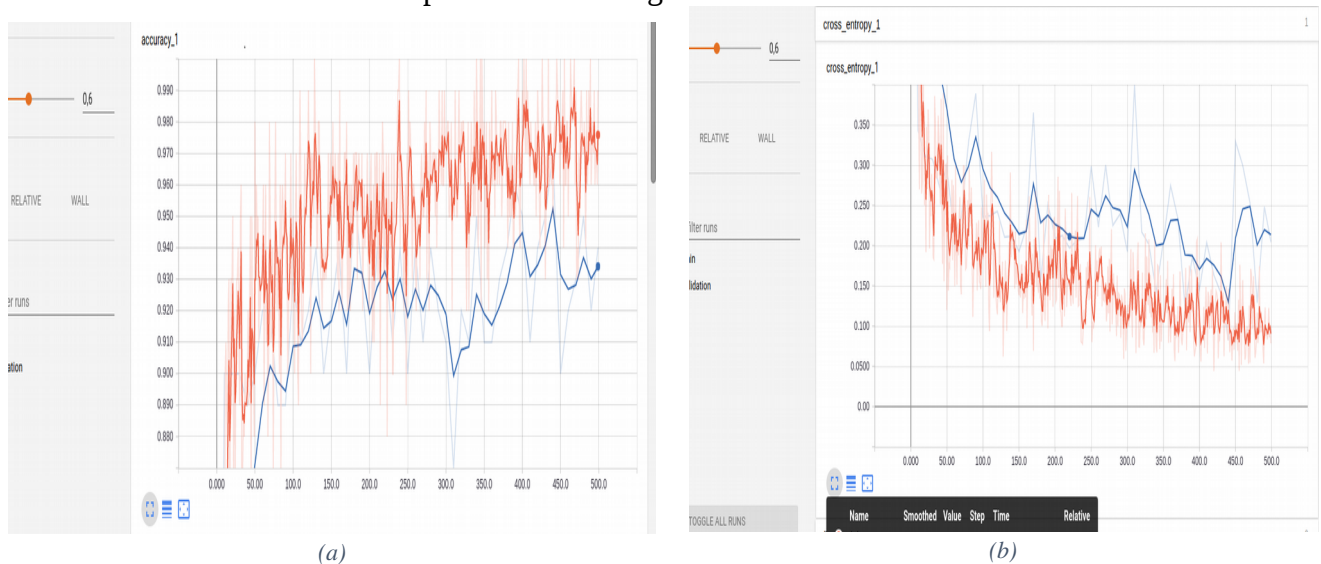


Figura 4. Treinamento da MobiliNetV1: (a) acurácia (b) erro

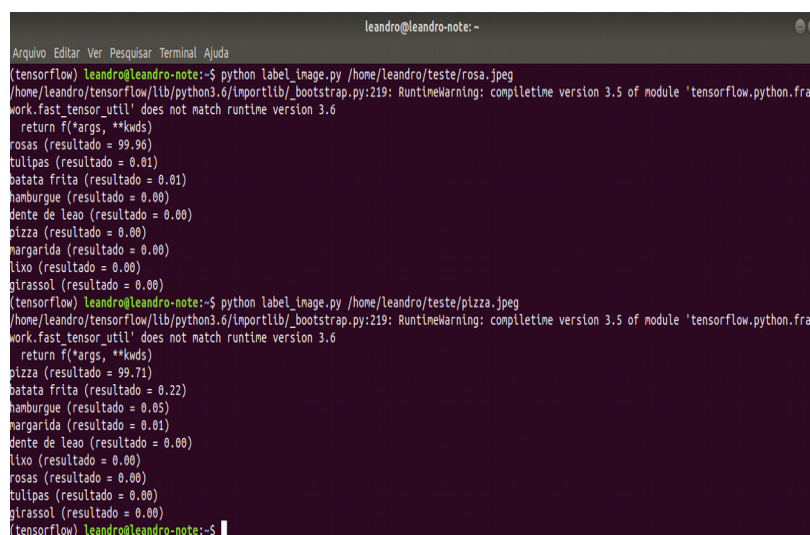
<sup>1</sup> O TensorFlow é uma biblioteca de aprendizado de máquina. Sendo utilizada como uma ferramenta para escrever e executar algoritmos de aprendizado de máquina.

Com base nos gráficos de acurácia (Figura 4a) e de erro (Figura 4b), pode-se notar a instabilidade do aprendizado da rede, ficando em torno de 97% na precisão do treinamento ao final das interações. A validação do treinamento começou a ocorrer após as primeiras cinquenta interações de aprendizado, devido à instabilidade de treinamento dessa rede, chegando em momentos que obteve oscilações na etapa de treinamento e validação devido ao tipo de dado que fora passado para rede, pois as imagens não possuíam o mesmo tamanho em pixel, ficando aproximadamente em 93% no final das 500 interações de treinamento. Nota-se também que a taxa de erro no treinamento na arquitetura MobilenetV1 foi maior que na arquitetura apresentada anteriormente, ficando em torno de 11,9%. Isso deve-se a arquitetura possuir um menor número de camadas e possuir dificuldade para lidar com a variabilidade das imagens fornecidas na entrada.

## 4.2 Validação

Após o treinamento realizado nas arquiteturas de redes neurais profunda aplicou-se testes de validação para verificar se o aprendizado foi realizado com êxito. Como a InceptionV3 obteve melhores resultados no treinamento ela foi escolhida para efetuar os testes de validação. Realizou-se testes de reconhecimento de imagens utilizando fotografias de flores, comida e resíduos sólidos.

Para certificar-se que o aprendizado teve êxito criou-se um banco de dados de testes com imagens que não estavam inseridas no banco de dados de treinamento das arquiteturas da rede neural profunda. Nos testes informou-se para a rede neural imagens que sugeria lixo, flores e comida, a rede neural devolveu o resultado que é demonstrado na Figura 5.



```
leandro@leandro-note: ~  
Arquivo Editar Ver Pesquisar Terminal Ajuda  
(tensorflow) leandro@leandro-note:~$ python label_image.py /home/leandro/teste/rosa.jpeg  
/home/leandro/tensorflow/lib/python3.6/importlib/_bootstrap.py:219: RuntimeWarning: compiletime version 3.5 of module 'tensorflow.python.framework.fast_tensor_util' does not match runtime version 3.6  
  return f(*args, **kwargs)  
rosas (resultado = 99.96)  
tulipas (resultado = 0.01)  
patata frita (resultado = 0.01)  
hamburgue (resultado = 0.00)  
dente de leao (resultado = 0.00)  
pizza (resultado = 0.00)  
margarida (resultado = 0.00)  
lixo (resultado = 0.00)  
girassol (resultado = 0.00)  
(tensorflow) leandro@leandro-note:~$ python label_image.py /home/leandro/teste/pizza.jpeg  
/home/leandro/tensorflow/lib/python3.6/importlib/_bootstrap.py:219: RuntimeWarning: compiletime version 3.5 of module 'tensorflow.python.framework.fast_tensor_util' does not match runtime version 3.6  
  return f(*args, **kwargs)  
pizza (resultado = 99.71)  
patata frita (resultado = 0.22)  
hamburgue (resultado = 0.05)  
margarida (resultado = 0.01)  
dente de leao (resultado = 0.00)  
lixo (resultado = 0.00)  
rosas (resultado = 0.00)  
tulipas (resultado = 0.00)  
girassol (resultado = 0.00)  
(tensorflow) leandro@leandro-note:~$
```

Figura 5. Resultado da validação

Como pode-se notar, em um dos experimentos foi fornecida uma imagem de uma rosa e a rede identificou corretamente com 99.96% de certeza. Em seguida foi fornecida uma imagem de uma pizza e novamente a rede identificou corretamente com 99.71% de certeza. Isto demonstra que a rede neural adquiriu conhecimento com os

dados de treinamento, tendo capacidade de generalizar o conhecimento para reconhecer imagens não vistas anteriormente com baixa taxa de erro.

## 5. Conclusão

Esse artigo teve como objetivo efetuar um comparativo entre duas arquiteturas de redes neurais profundas, MobilenetV1 e a InceptionV3, com vistas a demonstrar qual possui maior acurácia no reconhecimento de imagens de diferentes categorias.

Após os experimentos, pode-se concluir que a rede Inceptionv3 possui melhor performance no reconhecimento de imagens que a MobiliNetV1. Considerando o menor erro médio e o melhor percentual de precisão, nota-se que a rede torna-se mais generalista a medida que mais neurônios e camadas são utilizadas. Outro fato que foi observado é que quanto maior a base de imagens usada no treinamento da rede, maior é a sua assertividade. Entretanto, não devem ser fornecidas imagens muito parecidas pois podem levar a rede a sofrer sobre ajuste (overfitting) e consequentemente diminuir sua generalidade.

No desenvolver do trabalho as arquiteturas de redes neurais analisadas demonstram-se com um bom desempenho quando foram empregadas imagens de baixa resolução com um tamanho de 280 x 180 pixels em média. Além disso, pode-se constatar que para a base dados de imagens utilizadas, a melhor taxa de aprendizagem foi de 0.05. Esse valor foi obtido após diversas execuções das redes, demandando tempo e diversas análises dos resultados.

Por fim, pode-se apontar que a grande desvantagem de uma estrutura convolucional é a sua complexidade e consequentemente uso excessivo de recursos computacionais. Para poder se obter um treinamento eficiente de uma rede desse tipo é recomendado fazer o uso de GPUs.

Como trabalhos futuros aponta-se a análise comparativa de outras arquiteturas de redes neurais profundas como ResNet e SENet.

## Referências Bibliográficas

- Abdel Hamid, Ossama et al. Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE, p. 4277–4280, 2012
- Einloft, Daniel Centeno; Manssour, Isabel Harb. Detecção de Portas para Auxiliar o Deslocamento de Deficientes Visuais em Ambientes Internos. Seminário Interno de Avaliação da Iniciação Científica – PUCRS, Junho de 2017.
- Faceli, Katti; Inteligência Artificial: Uma abordagem de Aprendizagem de Máquina / Katti Faceli. [et al.]. – Rio de Janeiro: LCT, 2011.
- Ferneda, Edberto, “Redes neurais e sua aplicação em sistemas de recuperação de informação”, *Ci. Inf.* [online] vol.35, n.1, pp.25-30, 2006
- Guo, Yanming et al, “Deep learning for visual understanding: A review”, *Neurocomputing*, Volume 187, Pages 27-48, 2016



- G. Howard, Andrew & Zhu, Menglong & Chen, Bo & Kalenichenko, Dmitry & Wang, Weijun & Weyand, Tobias & Andreetto, Marco & Adam, Hartwig. “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications”, 2017
- LeCun, Yann & Bengio, Y & Hinton, Geoffrey. “Deep Learning”. Nature. 521. 436-44. 10.1038/nature14539, 2015
- Marques, Eduarda Almeida Leão, “Estudo sobre redes neurais de aprendizado Profundo com Aplicações em Classificação de Imagens”, UnB, Monografia, 2016
- Parker, J. R., “Algorithms for image processing and computer vision”, Willey, 504 pp, 2<sup>nd</sup> edição, 2011
- Wahed, Rayed Bin. “Comparative Analysis between Inception-v3 and Other Learning Systems using Facial Expressions Detection”. Department of Computer Science & Engineering BRAC UNIVERSITY. Bangladesh, 2016.