Instituto Universitário de Lisboa

Departamento de Ciências e Tecnologias da Informação (DCTI)

Classificação de imagens de interior/exterior de imóveis e a sua qualidade representativa

Maria Quintela Cruz

Dissertação submetida como requisito parcial para obtenção do grau de Mestre em Engenharia Informática

Orientador
Professor Doutor Tomás Gomes da Silva Serpa Brandão
ISCTE-IUL
Coorientador
Professor Doutor Luís Miguel Martins Nunes
ISCTE-IUL

Outubro, 2019

Agradecimentos

A realização da presente dissertação de mestrado não seria possível sem o apoio de diversas pessoas às quais gostaria de prestar os meus agradecimentos.

Em primeiro lugar, gostaria de agradecer ao Professor Doutor Tomás Brandão, orientador desta dissertação, bem como ao Professor Doutor Luís Nunes, coorientador da mesma, pela ajuda incansável na orientação deste trabalho; sempre se mostraram disponíveis e sempre procuraram obter o melhor de cada etapa do trabalho.

Quero agradecer também à minha família por me ter incentivado e motivado nesta fase final do mestrado.

Por fim, agradeço também aos meus amigos mais próximos que me acompanharam nesta fase de trabalho mais intenso.

Resumo

A aprendizagem automática tem vindo a progredir cada vez mais rapidamente, especialmente na área do reconhecimento de imagens. Apesar das redes neuronais convolucionais estarem presentes na classificação de imagens há muito tempo, apenas nos últimos anos têm sido reconhecidas pelo seu bom desempenho. A principal vantagem destas redes neuronais é a flexibilidade em transpor conhecimento entre problemas, i.e. usar uma rede treinada para um determinado tipo de problema num outro problema com pequenas adaptações.

Por outro lado, as redes neuronais convolucionais têm uma grande exigência computacional, bem como uma exigência de grande volume de dados de treino. Este nível de exigência tem vindo a ser ultrapassado ao longo dos anos e, consequentemente, têm-se vindo a obter cada vez melhores resultados.

Neste trabalho é utilizada uma rede neuronal convolucional pré-treinada (metodologia denominada *transfer learning*). Esta metodologia foi aplicada no presente trabalho com dois objetivos distintos: para a classificação de imagens de interior e de exterior de imóveis e para a classificação da qualidade representativa destas imagens.

Uma vez que para o treino e validação de uma rede neuronal convolucional é necessário ter previamente os dados classificados, de modo à rede poder extrair características e aprender, foi necessário identificar um conjunto de dados para o treino e validação da rede. Para tal, foram recolhidas e editadas imagens, de modo a permitir treinar a rede neuronal para os dois objetivos referidos anteriormente.

A aplicação deste tipo de classificações pode ser útil na área das imobiliárias, pois as imagens submetidas em plataformas de compra e venda de casas exige uma qualidade mínima. Para além da exigência no que diz respeito às imagens introduzidas, este tipo de classificações pode ser útil na filtragem de imagens dentro deste tipo de aplicações.

Os resultados obtidos nas experiências realizadas não foram completamente satisfatórios, pois a rede sofreu de *overfitting*, como se pode verificar no capítulo 4.

Palavras-Chave: classificação de imagens, interior, exterior, qualidade representativa, redes neuronais convolucionais, *transfer learning*

Abstract

Machine learning algorithms have grown very fast during the past few years, especially in the area of image recognition. Recently, convolutional neuronal networks have proven their efficiency in image classification. One of the main advantages of these neuronal networks is the facility to transfer knowledge between different problems, i.e. to use a network trained for a particular type of problem to solve another problem, only with small adaptations.

On the other hand, convolutional neuronal networks require a great computational power, as well as a large training dataset. This level of requirements have been increased during the last years and therefore accomplishing greater results.

In the present work, a pre-trained convolutional neuronal network (methodology called transfer learning) is used. This methodology was applied with two distinct objectives: image classification of real estate images (indoor and outdoor) and classification of the representative quality of these images.

Since the training and validation of a convolutional neuronal network requires a previous classified dataset, so that the network can extract characteristics and learn them, it was necessary to classify a dataset for the training and validation of the network. For this purpose, images were collected and edited in order to allow the training of the neuronal network for the two objectives mentioned above.

Keywords: image classification, interior, exterior, representative quality, convolutional neural networks, transfer learning

Índice

Lista de Figuras

Capítulo 1

1. Introdução

A capacidade visual humana é imensa e simultaneamente involuntária. Através da sua capacidade visual, um ser humano consegue reconhecer, perceber e classificar objetos, padrões e conteúdos que inconscientemente vai aprendendo ao longo da sua vida

Como esta capacidade é requerida em muitas tarefas repetitivas e monótonas, é importante realizar um esforço no sentido de automatizar processos de reconhecimento e classificação de imagens.

A visão computacional é a área científica responsável pela criação de modelos que se comportam de forma semelhante ao sistema visual humano. Existem vários campos dentro desta área, tais como o reconhecimento e a classificação de imagens.

Os resultados obtidos nas experiências de modelização do sistema visual humano mostram atualmente resultados surpreendentes, conseguindo, para algumas tarefas, ultrapassar a visão e o processamento visual puramente humano. Apesar dos resultados conseguidos até aos dias de hoje, a investigação neste âmbito e nas suas várias vertentes é extensa, deixando ainda vários problemas e dificuldades por ultrapassar.

Neste contexto, foi verificado que existem inúmeros trabalhos com a finalidade de classificar imagens quanto ao exterior e interior de cenários e ambientes [1][2][3][4].

Estes trabalhos podem servir de base para alcançar bons resultados numa temática ainda mais específica: a classificação de imagens de interior e exterior de imóveis [5]. Algoritmos como estes podem ser uma mais valia para plataformas de compra e venda de imóveis, como por exemplo para validação de número mínimo de imagens por divisão. Porém, este tipo de plataformas requer um padrão e qualidade representativa mínima das imagens nelas introduzidas.

O presente trabalho pretende contribuir para uma automatização na temática da classificação de imagens de interior e exterior de imóveis, bem como a classificação da

qualidade visual representativa das mesmas, pois esta área tem grande abertura para a exploração e contribuição.

1.1. Motivação

Os avanços na área da aprendizagem automática permitem automatizar tarefas como reconhecimento de padrões, observação de resultados e ajuste do seu funcionamento para obtenção de melhores resultados. Estes processos estão de tal modo evoluídos que, para determinados problemas e desafios, têm melhor desempenho do que o próprio ser humano.

Na classificação de imagens, as redes neuronais convolucionais têm vindo a evidenciar-se cada vez mais com um excelente desempenho em diversos problemas de classificação. Estas redes são baseadas em modelos matemáticos que replicam algumas das características do cérebro humano, em particular do córtex visual.

É imediatamente dedutível a vantagem, em termos de poupança de tempo de processamento e de recursos humanos para tarefas deste âmbito, evitando a necessidade de verificação, validação e classificação manual das imagens após a sua fase de treino. No entanto, para fazer o treino de uma CNN é necessário a construção de um *dataset* robusto em termos de tamanho e diversidade de imagens. Para isso é preciso obter um grande número de imagens, bem como a sua classificação, e dispor, para o treino, de recursos computacionais significativos.

Dentro do universo da aprendizagem automática existem as redes neuronais convolucionais que já provaram o seu poder e a sua eficiência no reconhecimento e classificação de imagens de um modo geral, tendo sido inspiradas na organização da rede neuronal humana. O treino e aprendizagem das CNN é feito através de exemplos, razão pela qual é necessário ter um conjunto de imagens pré classificadas. Tipicamente, quanto maior o conjunto de dados de treino, melhor o desempenho das CNN.

Assim, para utilização de uma rede neuronal convolucional, e para uma classificação eficaz e precisa, será necessário escolher qual a melhor abordagem a aplicar ao problema e aperfeiçoa-la de modo a obter o melhor resultado possível.

Existem diversos trabalhos quanto à classificação de imagens. As categorias de imagens poderão corresponder desde dígitos[6][7], a marcas de carros[8], plantas[9], etc.

Os melhores resultados são fruto do uso de redes neuronais convolucionais (CNN) juntamente com um *dataset* robusto no que diz respeito ao tamanho do conjunto de imagens. *ImageNet* [10] é uma bom exemplo disso. Esta base de dados contém mais de 14 milhões de imagens de diversas categorias e classes; por exemplo dentro da categoria animais tem conjuntos de imagens de cães, gatos, etc. No entanto, para a classificação de imagens de interior e de exterior de imóveis, este *dataset* não dispõe de um conjunto de imagens especificas, não podendo ser aproveitado para o fim em vista.

Assim sendo, a disponibilização de um vasto conjunto de dados é essencial, mas nem sempre facilmente alcançável. A falta de conjuntos de imagens para tarefas específicas de reconhecimento e identificação de imagens é um problema geral nesta área.

Quando a obtenção de um *dataset* robusto é dificilmente alcançável para um determinado desafio, pode ser necessário introduzir metodologias que permitam a utilização de soluções já existentes: a metodologia *transfer learning* permite, de um modo geral, reutilizar um modelo já existente, aplicando-o a um objetivo diferente. Este conceito existe para combater a grande necessidade de recursos computacionais, de tempo e do largo conjunto de dados que é necessário para desenvolver e treinar redes neuronais.

Para alcançar o objetivo do presente trabalho, é necessária a construção de uma coleção de imagens que reúna imagens de exterior e imagens de interior de imóveis, bem como a classificação prévia da qualidade representativa destas imagens. Este conjunto será criado a partir de um conjunto já existente, através do processamento do mesmo.

A classificação visual de imagens no que diz respeito à sua qualidade visual/estética é uma temática na qual se têm verificado progressos. Ainda assim, não foi possível, até à data, aferir quais os parâmetros que definem a qualidade visual representativa.

A interpretação que os humanos têm de certas imagens e a classificação destas quanto à sua representação visual é subjetiva, variando de pessoa para pessoa. A necessidade de classificar imagens quando à sua qualidade representativa é um dos desafios presentes neste trabalho. É necessário obter diversas classificações de um conjunto de imagens e perceber qual a melhor maneira de interpretar estes resultados.

A determinação do valor qualitativo de várias imagens quanto á sua representação visual é a base para a automatização deste mesmo processo.

Uma vez tendo imagens classificadas, é possível o desenvolvimento de uma ferramenta com o objetivo de classificar novas imagens.

1.2. Objetivos

O presente trabalho tem como finalidade o desenvolvimento de uma ferramenta capaz de classificar imagens de imóveis. Numa primeira fase, as imagens dos imóveis são classificadas segundo interior/exterior. Posteriormente as mesmas são também classificadas quanto à qualidade da sua representatividade.

Para alcançar este objetivo, é necessária uma coleção de imagens de imóveis com diferentes níveis de qualidade representativa, bem como a classificação prévia dessas mesmas imagens. Assim, um dos desafios deste trabalho é a classificação prévia de cada imagem no conjunto de dados. A classificação prévia é resultante do pré-processamento de um conjunto de imagens, como referido posteriormente.

Como referido anteriormente, para melhor desempenho de uma rede neuronal convolucional, é necessário um vasto conjunto de imagens classificadas para o treino desta, o que torna a composição deste conjunto um dos principais objetivos deste trabalho.

O presente trabalho pretende contribuir cientificamente da seguinte forma:

- Analisar o comportamento de CNNs para classificação de imagens de imóveis quando ao exterior e interior deste, usando a metodologia *transfer learning*, bem como classificar a qualidade representativa destas imagens.
- Perceber qual o grau de adequação destas metodologias comparativamente ao treino de raiz de uma CNN.
- Criação de um *dataset* próprio para a classificação da qualidade representativa de imagens de interior e exterior de imóveis.

1.3. Estrutura do documento

Após esta introdução, o capítulo seguinte tem como finalidade esclarecer alguns conceitos relacionados com as soluções apresentadas para a classificação de imagens.

No Capítulo 2 são apresentadas algumas das soluções encontradas até os dias de hoje no que diz respeito à classificação de imagens: de espaços/cenários e de divisões de imóveis. A exploração da definição relacionada com a estética de uma imagem, os *datasets* com mais relevância adotados nesta área, bem como as principais medidas usadas para análise e comparação do desempenho dos modelos são também explorados neste capítulo.

No Capítulo 3 é apresentada uma revisão literária dos desenvolvimentos realizados até à data, passando de um modo geral por classificação de imagens, redes neuronais e trabalhos relacionados.

No Capítulo 5 são apresentadas as experiências realizadas, que consistem na classificação única de imagens do interior e exterior de um imóvel, bem como da respetiva qualidade representativa, seguindo-se um capítulo relativo a conclusões e outro relativo a possíveis futuros desenvolvimentos.

Capítulo 2

2. Redes Neuronais Convolucionais (CNNs)

As redes neuronais convolucionais têm vindo a demonstrar uma elevada taxa de sucesso em diversos problemas que envolvem a classificação de imagens. Para uma melhor compreensão dos métodos utilizados, que serão detalhados no Capítulo 4, o presente capítulo tem como objetivo explicar e detalhar alguns dos conceitos utilizados e das diferentes camadas que compõem uma rede neuronal convolucional.

2.1 Arquitetura de uma CNN

As redes neuronais convolucionais têm vindo a marcar a sua posição através do excelente desempenho alcançado em diferentes problemas que envolvem a classificação automática de imagens.

Como referido anteriormente, estas redes neuronais são inspiradas na estrutura do cérebro humano e na disposição do córtex visual.

Em 2012, com o aparecimento da rede AlexNet [11], cujos resultados ultrapassaram qualquer método conhecido até à altura para classificação de imagens, as CNN começaram a ganhar reconhecimento nesta área.

As CNN são a estrutura de dados que suporta um algoritmo de *deep learning* que consegue, a partir de uma imagem, determinar características desta e, assim, diferenciá-la de outras imagens. Esta classificação é conseguida através da determinação dos pesos a dar a diferentes características de uma imagem. Ou seja, para cada propósito de uma classificação, certas características têm maior ou menor peso, influenciando assim de diferente forma a classificação final.

Outra característica inerente às redes neuronais convolucionais é a capacidade de aprenderem por si mesmas as características e filtros aplicados com vista a obtê-las. Esta característica distingue as CNN de outros algoritmos de classificação mais primitivos

onde os filtros e as características das imagens eram determinados e obtidos manualmente.

A arquitetura das redes neuronais convolucionais foi inspirada no sistema de visão do ser humano, capaz de receber uma imagem e, através da extração de características e do processamento, classificar essa imagem. Nas CNNs não existe um número fixo de camadas, mas sim um conjunto de camadas diferentes que é transversal a todas as arquiteturas.A sua arquitetura compõe-se principalmente por três tipos de camadas: camada convolucional, camada de *pooling* e camada *fully-connected*.

De um modo geral, uma CNN recebe na entrada imagens de dimensão fixa, passando por camadas convolucionais (ver *Secção* 2.1.1), e *pooling* (ver *Secção* 2.1.2), no qual o objetivo é obter uma representação condensada das imagens, uma espécie de resumo da imagem, através das características nestas presentes. Por fim, existe uma camada *fully-connected* (ver *Secção 22.1.3*) responsável por fazer a classificação da imagem nas diferentes classes que possam existir.

2.1.1. *Convolution Layer*

Numa CNN existem tipicamente mais do que uma camada convolucional, contudo uma imagem que passe por uma CNN só é diretamente aplicada à primeira camada convolucional. Nesta camada, a imagem passa por um ou mais filtros e é a primeira camada a extrair características de alto nível, como por exemplo cantos, cores, etc. O facto de uma CNN poder ter mais que uma camada convolucional permite obter características a outros níveis também.

Um dos seus propósitos da utilização destas camadas é reduzir o espaço necessário para os dados das imagens, isto é, reduzir a dimensão das matrizes de entrada nestas camadas sem perder características, para não influenciar o desempenho da classificação das imagens. Esta redução permite um processamento mais fácil, rápido e escalável para *datasets* muito grandes.

Para a operação de redimensionamento das imagens é utilizado um filtro, também conhecido como *kernel*, e aplicada uma operação de convolução. Na Figura 1 é ilustrada uma imagem de dimensão 5×5×1 do lado esquerdo, à qual é aplicado um filtro 3×3×1 e o output da operação de convolução é uma matriz 3×3×1 representante das características de uma imagem.

O valor do deslocamento do filtro dentro da imagem é denominado por *stride*. No exemplo da Figura 1 aplica-se um *stride* com valor = 1.

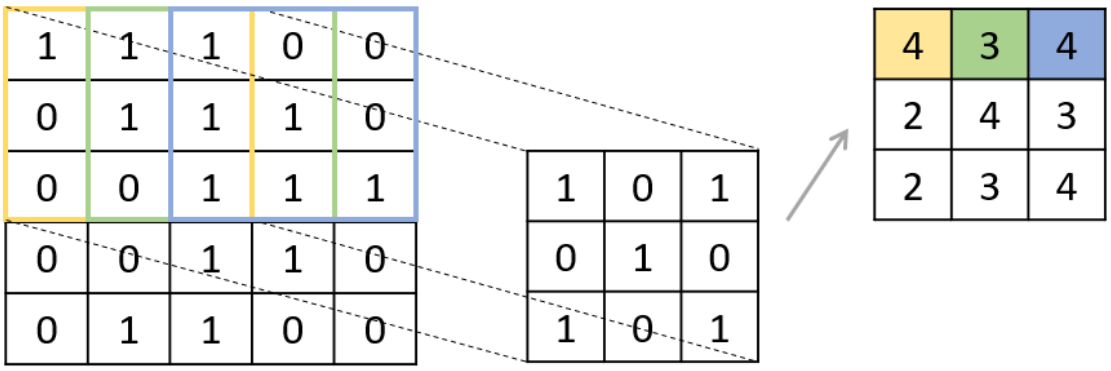

Figura 1 - Camada convolucional - Aplicação de um filtro numa imagem de dimensão 5×5×1.

Ao aplicar filtros à imagem para reduzir a sua dimensão pode estar a perder-se informações dos cantos das imagens. Na Figura 1, por exemplo, o filtro apenas tem em consideração os pixéis do canto uma vez. Por outro lado, os pixéis centrais são considerados pelo filtro mais vezes.

A aplicação de preenchimento / *padding* na imagem original permite neutralizar o efeito de redução da dimensão de saída e, também, considerar os pixéis dos cantos mais do que uma vez. Em redes cujo número de camadas é elevado, esta solução permite prevenir que os valores à saída sejam demasiado reduzidos pelo redimensionamento. Quando aplicado um filtro de *padding* nas camadas convolucionais, não é feita uma redução. A redução é feita sobretudo nas camadas de pooling.

Na Figura 2 ilustra-se a adição de pixéis de *padding* (pixéis brancos a contornar os pixéis azuis) numa imagem de dimensão 4×4×1 (pixéis azuis). A aplicação de um filtro 3×3×1 e *stride* = 1 permitiria cobrir os pixéis dos cantos mais que uma vez.

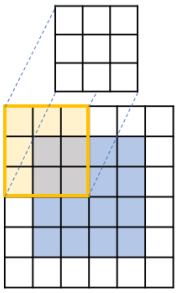

Figura 2 - Exemplo de *padding* numa imagem.

2.1.2. *Pooling Layer*

A camada designada por *Pooling layer* consiste em *sub-amostrar*/reduzir o tamanho espacial dos dados associados às características anteriormente obtidos numa camada convolucional, de modo a baixar a complexidade nas camadas seguintes.

O método de *pooling* mais conhecido é o *max-pooling*. Este método consiste em dividir uma imagem em sub-blocos e, para cada sub-bloco, devolver apenas o valor máximo dentro desse sub-bloco. A Figura 3 apresenta do lado esquerdo o resultado da aplicação do método *max-pooling* numa imagem 4×4, com filtro 2×2 e *stride* = 2, resultando numa sub-amostragem de dimensão 2×2.

Outro método de *pooling* é o *average-pooling* que calcula a média dos valores dentro de cada bloco nos quais a imagem foi dividida. Também na Figura 3, do lado direito, se pode observar um exemplo de aplicação do método *average-pooling* a uma imagem 4×4, com filtro 2×2 e *stride* = 2, resultando numa sub-amostragem de dimensão 2×2.

Ambos os métodos de *pooling* condensam a informação num sub-bloco.

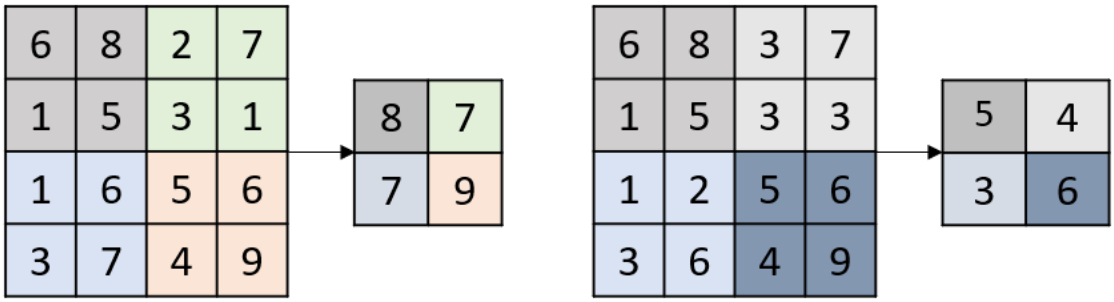

Figura 3 - Do lado esquerdo: representação do método *max-pooling*; do lado direito: representação do método *average-pooling*.

Quando aplicado o método de *pooling*, a informação da posição não é preservada, Como tal, apenas se deve aplicar quando a informação da imagem em si é mais relevante, e não a informação espacial. A eficiência pode ser melhorada utilizando "filtros" de *pooling* e *strides* desiguais, para poder manter algumas áreas sobrepostas.

Detalhes de baixo-nível podem ser alcançáveis através do aumento do número de utilizações desta camada juntamente com a camada de convolução, mas a custo de poder computacional.

2.1.3. *Fully-Connected Layer*

A organização desta camada é muito idêntica à de uma rede neuronal tradicional, ou seja, cada nó / neurónio está diretamente ligado aos nós da camada anterior e aos nós da camada seguinte.

A informação proveniente das camadas convolucionais e de *pooling* representa as características de alto-nível da imagem de entrada. A camada *fully-connected* usa esta informação e gera um vetor N dimensional, no qual N é o número de classes possíveis.

A Figura 4 representa esta camada com 4 possíveis classes e com uma probabilidade associada a cada uma destas classes. Os pesos e ligações não são aqui ilustrados.

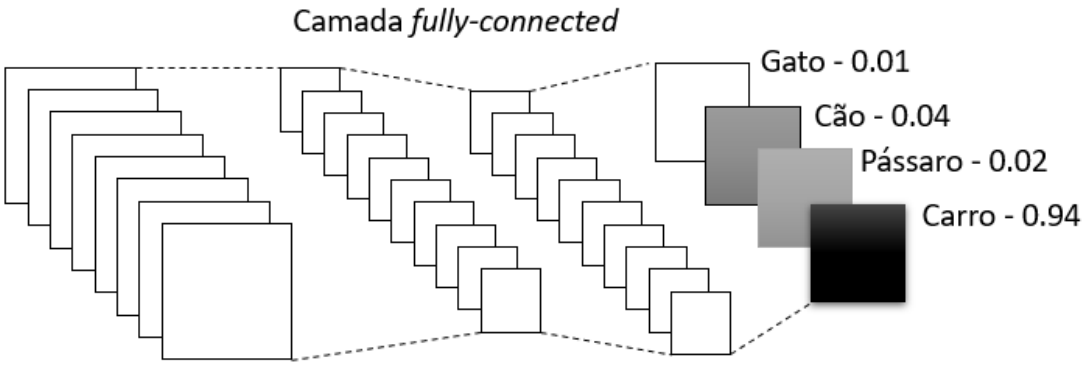

Figura 4 - Camada *fully-connected.*

Imediatamente a seguir a esta camada, existe ainda, em muitas CNNs, uma camada de *Softmax* responsável por determinar a probabilidade de uma imagem pertencer a cada classe, sendo a probabilidade um valor decimal que varia entre 0 e 1,0 e a soma das probabilidades de todas as classes igual a 1.

2.1.4. Unidade Linear Retificada (ReLU)

A camada *ReLu, Rectified Linear Unit,* calcula a função matemática expressa por *max*(0,x) [7]. Esta função converte para 0 qualquer valor negativo, enquanto os valores positivos são mantidos, como é representado na Figura 5.

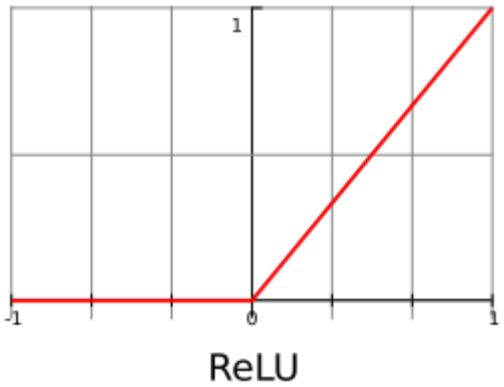

Figura 5 - Representação da Função ReLU *[7]*.

Quando se olha para uma imagem, é imediatamente percetível que esta contém características não lineares como por exemplo, cantos, cores, etc. Tendo em conta que as imagens são naturalmente não lineares, a utilização desta função também não-linear permite considerar esta não linearidade.

As principais vantagens da utilização destas funções de ativação são a facilidade da computação, tornando o tempo de treino da rede menor. A convergência também é rápida, não sofrendo assim do problema do *vanishing gradient* (ver *Secção 2.2*), que outras funções de ativação sofrem.

2.1.5. *Dropout*

Na fase de treino de uma rede neuronal é possível ignorar aleatoriamente certos neurónios. quando um neurónio é deixado de parte, também as ligações para estes neurónios são "esquecidas", como é visível na Figura 6. Este mecanismo é denominado por *dropout*.

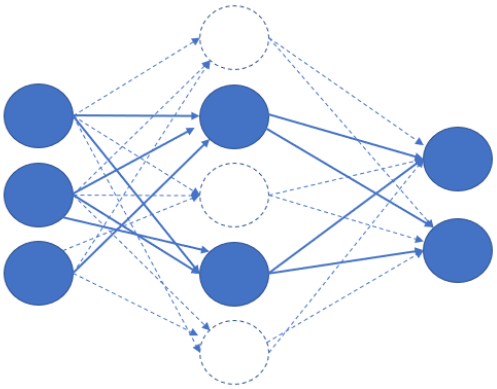

Figura 6 - Exemplo de uma rede neuronal com *dropout.*

Este mecanismo pode ser introduzido numa rede neuronal para combater *overfitting*, que pode ocorrer quando o a taxa de erro no treino é muito baixa e, ao mesmo tempo, a taxa de erro na validação é muito alta [12]. Por outras palavras, o modelo não está generalizado para dados desconhecidos, mas sim para os dados pelos quais foi treinado.

Ao introduzir a camada de *dropout* numa rede, é possível combater *overfitting*, pois, através desta medida, nenhum neurónio pode assumir a presença de outro neurónio, forçando estes a adaptarem-se à rede [13].

Por outro lado, ao ser aplicado *dropout* na fase de treino é garantido que o treinar da rede é feito numa "média" de diferentes arquiteturas de redes, pois a escolha de neurónios com *dropout* é aleatória a cada época.

2.2 Problemas no treino da rede - *Vanishing*

De um modo geral, à medida que são adicionadas mais camadas a um modelo, mais capacidade tem esse modelo de aprender informação complexa e, como tal, de prever melhor o resultado esperado [14].

Contudo, a adição de mais camadas resulta, na maioria dos casos, no uso de funções de ativação que levam o valor de perda a aproximar-se do valor 0, tornando a rede difícil de treinar [15][16]. Uma função de perda determina até que ponto os valores previstos se desviam dos valores reais.

Algumas das funções de ativação comprimem os valores de entrada para um espaço de valores entre 0 e 1. Nestes casos, quando existe uma grande alteração na entrada, o impacto na saída é pequeno. Consequentemente, a derivada é pequena, como se pode observar na Figura 7.

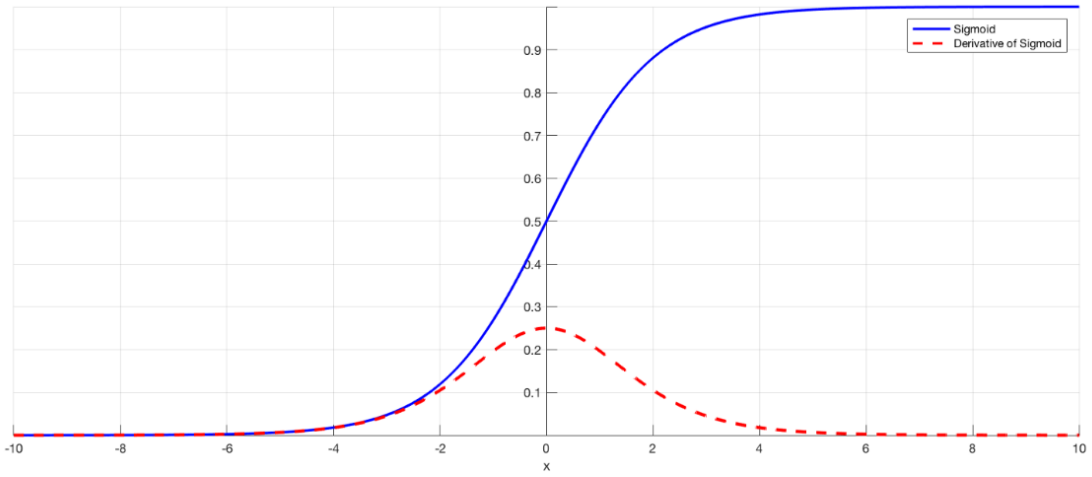

Figura 7 - Função Sigmoid e a sua derivada [15].

Na retropropagação, são aplicadas derivadas à rede, desde a camada final até à camada inicial. Assim, as derivadas de cada camada são multiplicadas e propagadas pela rede toda.

Quando se tem *n* camadas a utilizar funções *sigmoid*, *n* derivadas são juntamente multiplicadas, deixando o gradiente decrescer exponencialmente enquanto é propagado para as camadas iniciais.

Como consequência, os pesos destas camadas não são atualizados de forma eficiente, podendo pôr em causa a precisão da classificação da rede inteira. Uma das soluções apresentada para evitar tal acontecimento é a utilização de funções de ativação ReLU, tal como a descrito na Secção 2.1.4.

Capítulo 3

3. Estado da Arte

A classificação automática de imagens de espaços/ambientes é uma problemática que ainda não encontrou uma solução fixa. Mais especificamente, também a classificação de imagens de espaços segundo exterior/interior não possui ainda uma técnica de classificação consensual [1].

Métodos que utilizam como base *K-Nearest Neighbor* (K-NN), o *Support-Vector Machines* (SVM) ou redes neuronais são os métodos mais comuns e que demonstraram melhores resultados para este género de problema. Não é possível fazer facilmente uma comparação direta entre as várias soluções apresentadas, dado que vários trabalhos propostos utilizam diferentes *datasets.*

O desempenho das redes neuronais convolucionais (CNN) na classificação de imagens será apresentado nos pontos seguintes, fazendo uma ponte para a verificação imagens de imóveis de interior e exterior, âmbito específico do presente trabalho. As abordagens apresentadas até à data para a problemática de classificação de imagens de interior e exterior de imóveis, bem como a verificação da qualidade representativa destas imagens neste contexto são poucas. Deste modo, a exploração destas duas vertentes é uma contribuição para o desenvolvimento deste tema.

Quanto à classificação da qualidade de uma imagem, é um problema recorrente na área da visão computacional, embora a necessidade desta classificação tenha origem na área do processamento de sinal, mais concretamente da área da codificação de imagem e de vídeo. A solução é sempre subjetiva, dependendo da problemática em questão. Os principais trabalhos desenvolvidos neste domínio têm como objetivo a elaboração de um sistema com capacidade de processar imagens tal como o sistema de visão humano, abrangendo desde reconhecimento de padrões, classificação de imagens quanto à sua natureza, etc.

Neste trabalho será abordada a vertente de classificação de imagens quanto à sua qualidade representativa, mais concretamente a avaliação de uma imagem quanto à sua representação do espaço e ambiente de um imóvel, para os diversos espaços que o constituem.

O ser humano tem a capacidade de, com pouca informação ou má qualidade de imagem, perceber o que esta representa. Esta capacidade é dificilmente traduzível para uma máquina, pois a qualidade de uma imagem não é linear.

Em primeiro lugar, é necessário perceber os diferentes conceitos envolvidos no que pode influenciar a qualidade representativa de uma imagem: a sua qualidade em termos de compressão, focagem, luminosidade, etc., a estética e a representatividade, ou seja, se esta 4.capresenta de forma percetível o que é desejado.

A *qualidade* de uma imagem está relacionada com os sinais que formam as próprias imagens e é fortemente influenciada pelo ruído de compressão, pelo contraste, pela iluminação, pela focagem, etc. O próprio sinal é tipicamente usado para definir um valor para a qualidade de uma imagem.

A *estética* de uma imagem é, por outro lado, uma medida de quão apelativa uma imagem é num determinado contexto e que tipo de emoções desperta no observador. A definição de beleza e da sua interpretação está diretamente ligada a este conceito, sendo questionável e distinta de pessoa para pessoa.

Por outro lado, a *representatividade* de uma imagem engloba o conteúdo e o contexto da mesma. Por exemplo, se, para uma imagem de uma sala de estar, apenas se apresentar um sofá, pode afirmar-se que a sua representatividade é fraca.

Estes três conceitos estão interligados, ou seja, uma imagem cuja qualidade seja muito baixa, terá influência na representatividade desta, pois o seu conteúdo não é visível. Em alternativa, uma imagem cuja representatividade seja excelente não se traduz a um nível de estética elevado, pois o contexto e conteúdo a transmitir estão presentes.

Exemplos de imagem de imóveis com fraca qualidade representativa são apresentadas na Figura 8. Do lado esquerdo da imagem, é visível uma imagem cuja qualidade é bastante boa e esteticamente também apelativa, mas a qualidade representativa é bastante fraca, pois o foco da imagem é o ramo de uma árvore e não o imóvel, que apenas é visível ao fundo. Na imagem do meio, Figura 8, a iluminação e a perspetiva do cenário revelam uma má representação da qualidade do espaço. Por último, a imagem apresentada à direita contém apenas um cesto da roupa, significando uma má representação do espaço e, assim, classificando-se com uma imagem de fraca qualidade representativa.

Figura 8 - Três exemplos de imagens sem qualidade representativa.

Na Figura 9 são apresentados dois exemplos de imagens que, no contexto de imóveis, têm qualidade representativa. Na imagem da esquerda pode-se observar uma sala vazia. Apesar da imagem ser pouco apelativa e não suscitar qualquer emoção ao observador, e por isso poder classificada como pouco "estética", é uma imagem cuja qualidade representatividade do espaço está presente. A imagem da direita constitui um exemplo de uma imagem cuja qualidade é fraca em termos de iluminação, contudo a qualidade representativa do espaço está presente.

Figura 9 - Dois exemplos de imagens com qualidade representativa.

O presente capítulo divide-se em três secções, apresentando inicialmente as soluções existentes até à data para classificação de imagens de interior e exterior, passando pelas redes neuronais com melhores resultados na classificação de imagens em geral e posteriormente os trabalhos desenvolvidos relativamente à classificação em termos de qualidade representativa/estética do espaço.

3.1. Classificação de imagens de espaços de interior / exterior

Nesta secção apresenta-se diversas abordagens utilizadas na classificação de espaços / ambientes de interior e exterior, tendo-se distinguido estas abordagens de acordo com a solução aplicada na classificação dos espaços, desde a utilização de *K-Nearest*

Neighbor, *Support-Vector Machines* às Redes Neuronais. Estes métodos são detalhados na secção correspondente, referindo também os resultados obtidos em cada um deles.

3.1.1. Classificação utilizando *K-Nearest Neighbor* (K-NN)

Em 1998, numa fase embrionária da classificação de imagens em geral em ambientes de interior / exterior, Martin Szummer e Rosalind W. Picard [3] apresentaram uma solução que possibilita a deteção de propriedades de alto-nível (interior / exterior) baseada na extração de características baixo-nível. O desempenho foi melhorado através da extração das características em sub-blocos, classificando os mesmos e posteriormente combinando os resultados em forma de "pilha".

A extração de características de cor dos sub-blocos é feita através do uso de histogramas. Para extrair características das texturas é usada a transformada discreta de coseno (DCT) e parâmetros *multiresolution simultaneous autoregressive model* (MSAR).

A representação desta divisão de sub-blocos e a classificação separada de cores e texturas é apresentada na Figura 10.

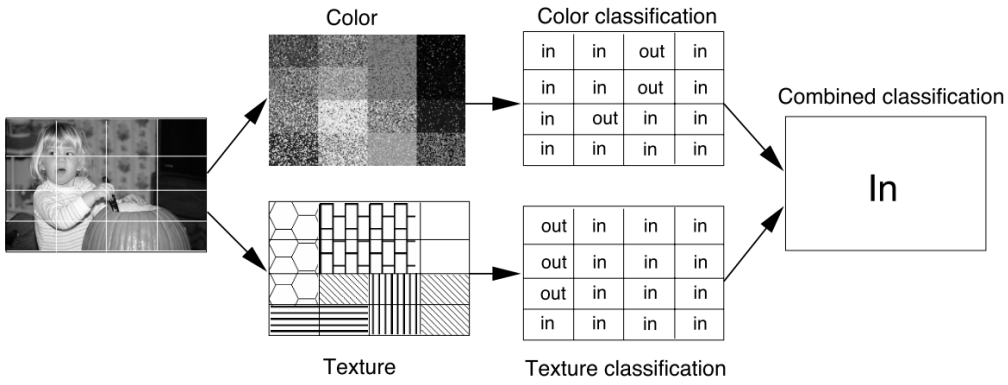

Figura 10 - Modelo apresentado por *[3]*.

Posteriormente, os vetores de características foram utilizados para classificação dos sub-blocos da imagem e, no fim, concatenados. A classificação de imagem é, aqui, feita aplicando o método K-NN, que consiste basicamente na classificação de um determinado elemento consoante as classes dos respetivos *k* vizinhos mais próximos, pertencentes a um conjunto de dados de treino.

Até à data da publicação deste trabalho, os modelos anteriormente desenvolvidos apresentavam uma taxa de classificação correta das imagens na ordem de 75-86%. Esta abordagem apresentou por isso grandes avanços na classificação de espaços interior/exterior, dado que conseguiu alcançar 90,3% de classificações corretas.

Perante a mesma questão, surge em 2001 uma outra solução [2], que segue o caminho da extração de características de baixo-nível de uma imagem, sendo estas as cores e as texturas. As técnicas dispostas são o histograma de cores, o modelo MSAR utilizado para extrair características de textura e o método K-NN aplicado para a classificação. Nesta fase, sabia-se que uma solução deste género não chegava para resolução do problema de classificação de espaços interior/exterior na sua totalidade.

Assim, a solução proposta vai mais além, sugerindo a utilização de uma rede Bayesiana [17], para a integração de características de cores e texturas, juntamente com a semântica de céu e vegetação, como é apresentado na Figura 11.

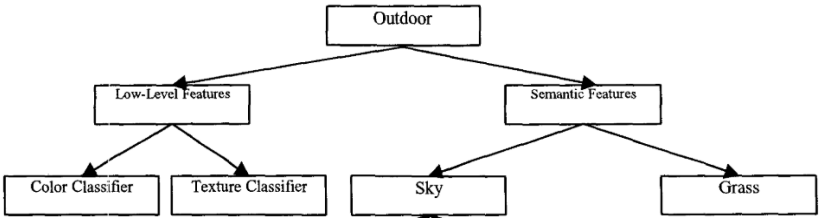

Figura 11 - Características utilizadas para classificação de imagens em *[2]*.

A aplicação da semântica dos dados para a deteção de relva e céu permite que o processo de classificação neste contexto seja otimizado. Para a extração destas características são particularizados dois procedimentos: 1) a assunção que a presença de céu e/ou vegetação é sempre correta em imagens *outdoor*, 2) a classificação baseada em elementos de cor/textura. Esta metodologia alcançou uma taxa de classificações corretas de 90,1%.

3.1.2. Classificação utilizando *Support-Vector Machines* (SVM)

Support-Vector Machine é um modelo linear de aprendizagem automática utilizado para problemas de classificação e de regressão. A principal ideia deste modelo é desenhar uma linha ou um hiperplano para dividir os dados em duas classes distintas.

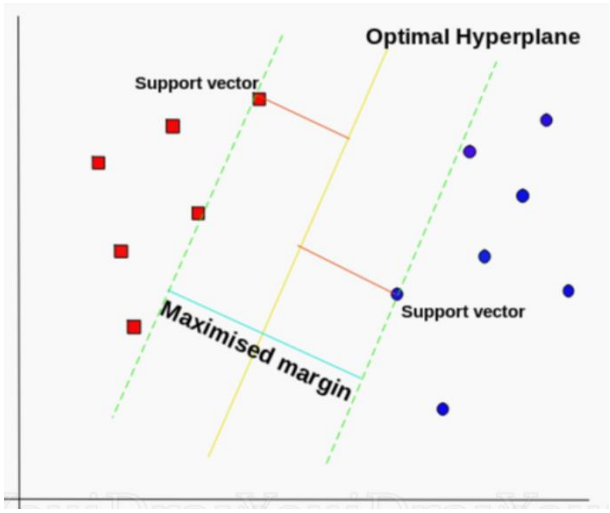

Figura 12 - Ilustração da utilização de SVM para a separação de dados em duas classes distintas [18].

Seguindo o exemplo apresentado na Figura 12, estão representadas duas classes distintas: os quadrados vermelhos e os pontos azuis. Como referido anteriormente, aplicando o modelo SVM, o objetivo será traçar uma linha que divida estas duas classes da forma mais generalista possível. Para tal, é necessário encontrar os pontos de cada classe mais próximos de uma linha; estes pontos são considerados *support vectors*. Uma vez encontrados estes pontos, é calculada a distância entre os *support vectors* e a linha. O principal objetivo é maximizar esta distância. O hiperplano para o qual a distância é máxima, é o hiperplano ideal.

No que diz respeito a soluções para a temática de classificação de interior/exterior [4] apresenta uma solução que tem como base a classificação através de SVMs, classificando as imagens em duas etapas:

- A primeira etapa corresponde ao treino das SVMs, de cor e textura correspondentemente, baseado em sub-blocos da imagem. O treino destas SVMs é feito sem qualquer dependência.
- Na segunda etapa, uma outra SVM recebe as classificações dos resultados anteriores e classifica com mais exatidão a imagem, sendo esta a classificação final. A Figura 13 apresenta uma representação visual do modelo.

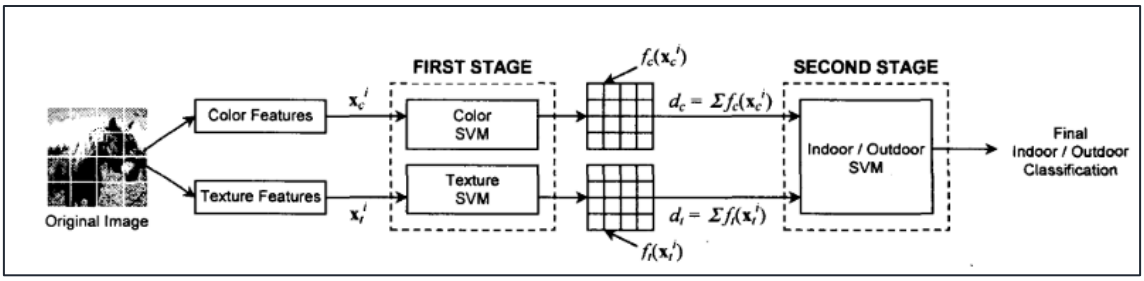

Figura 13 - Modelo de classificação de imagens com utilização de SVM *[4]*.

Muitas das técnicas utilizadas partem do princípio que uma imagem na qual as cores azul e/ou verde predominam, é uma imagem de exterior (céu e vegetação).

Para mitigar esta assunção, o método [19] foca-se na orientação das características de baixo-nível, neste caso, cores, texturas e arestas. A extração destas é feita dividindo a imagem em cinco sub-blocos, tendo cada sub-bloco diferentes pesos na classificação da imagem. A posição do sub-bloco na imagem influencia o peso que este tem.

Depois de calculados os descritores do histograma de orientação de arestas e cores, *edge and color orientation histograma* (ECOH), de cada sub-bloco, estes são agrupados para gerar um vetor que deve alimentar o classificador SVM. Os dois histogramas são calculados independentemente e só depois concatenados. Uma descrição visual deste modelo é apresentada na Figura 14.

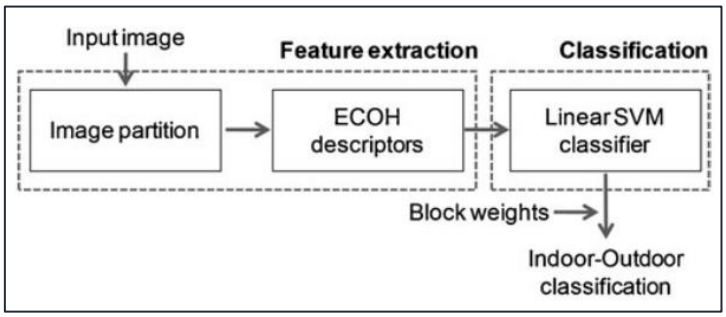

Figura 14 - Representação modelo [19].

Este modelo conduziu a um desempenho de 90,26% na classificação de imagens de interior/exterior corretas, não conseguindo destacar-se relativamente às soluções apresentadas à data (2010), contudo teve um dos melhores desempenhos nas soluções baseadas unicamente em classificadores SVM. Mostrou também ter melhores resultados em termos de tempo comparativamente à proposta apresentada em [11], quando aplicado o mesmo *dataset*, pois nesta solução foi introduzido os Wavelets em vez da utilização do

modelo MSAR para a extração de características. Numa comparação direta, para imagens com a mesma resolução, a computação de características apresentadas pelo modelo MSAR demorou 647 veze mais do que a solução através de Wavelet.

Com o passar do tempo, e ao serem apresentadas novas abordagens, tornou-se evidente que, quanto maior for a fonte de dados, melhores os resultados que podem ser obtidos. Foi então que [20] propôs utilizar dados referentes à fotografia, como por exemplo a utilização ou não de *flash*, a distância, o tempo de exposição, etc. e não apenas o cenário exposto na imagem. Para a junção das duas fontes de informação, i.e., a imagem em si mais os dados referentes à sua captura, foi utilizada uma rede Bayesiana. Esta rede permite fundir informação de fontes e domínios diferentes, para além de também se adequar à falta de informação, possibilitando receber como dados de entrada apenas imagens ou então imagens com informação extra (flash, tempo de exposição, etc). A informação a combinar pode ser visualizada na Figura 15.

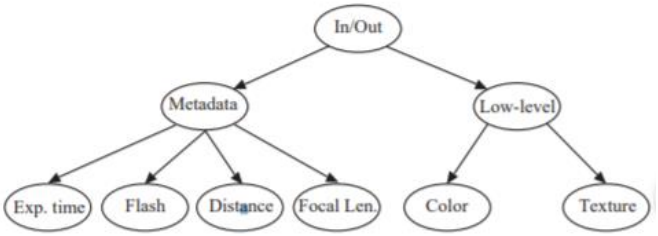

Figura 15 - Conjunto de características extraíveis dos dados de *input* [20].

Primeiramente, foi aplicado o modelo SVM para a classificação de imagens segundo as características de baixo-nível (cor e textura) e posteriormente aplicada a rede Bayesiana com o intuito de integrar a informação de baixo-nível com os dados adicionais de cada imagem, podendo assim ter uma melhor classificação.

Na melhor das hipóteses, ou seja, tendo acesso a todas as informações de uma imagem (características de baixo nível: cores e texturas, distância do objeto/sujeito e tempo de exposição) esta ferramenta conseguia um total de 94,1% de precisão. Comparativamente, a mesma ferramenta, mas apenas com os elementos extraídos a partir da imagem (cor e textura) a precisão descia para 81%. O desempenho nesta abordagem é notório, mas de certo modo ilusório, pois os dados referentes às condições de aquisição de imagens não estão geralmente disponíveis.

3.1.3. Classificação utilizando Redes Neuronais

Numa fase mais avançada das metodologias estudadas para a resolução do tema em questão, surgem as redes neuronais. Contudo, para o reconhecimento de imagens de interior e exterior houve poucos avanços e os resultados não chegam aos resultados obtidos por outros métodos.

Em 2015 surge uma nova interpretação [21] face ao desenvolvido até à altura em que, dada uma a imagem, o descritor da sua essência é calculado, resultando numa representação holística da imagem. A essência de uma imagem representa sucintamente a informação percetual que um humano compreende em cerca de 200ms ao visualizar um cenário [22]. Deste modo, a informação de orientações e escalas é resumida pelo *gist* para diferentes partes da imagem. Posteriormente, esta informação é transferida para uma rede neuronal, que é baseada na estrutura do cérebro humano, permitindo a descoberta de padrões e a correlação de dados de entrada e parâmetros de saída correspondentes. Após treinada, a rede neuronal ganha uma certa intuição de imagens de interior e exterior, conseguindo classificar a imagem com um grau de certeza. O grau de certeza desta rede após o treino atingiu os 90,8%.

O mais recente desenvolvimento para o tema específico da classificação de imagens de interior e exterior foi desenvolvido em 2017. A compra e venda de imóveis e a classificação dos espaços que compõem estes são temáticas abordados por [5]. Os espaços a identificar são seis: o quarto, a casa de banho, a cozinha, a sala de estar, o quintal e a entrada da casa. Para além da distinção dos espaços, um dos objetivos apresentados é a identificação do material usado no chão e nas bancadas da cozinha. Neste trabalho foi explorado o uso de uma CNN em comparação com uma rede *Long-Short Term Memory* (LSTM) para classificação dos diferentes cenários.

A principal diferença da rede LSTM face às outras redes é que esta funciona com blocos de memória que guardam valores num curto ou longo espaço de tempo. Estes blocos têm três portas distintas: uma de entrada, uma de esquecimento e a última é de saída. Estas portas controlam o estado e o *output* de cada bloco. A porta de entrada é responsável pela entrada de novas informações no bloco de memória. A porta de esquecimento, fica, como

o próprio nome indica, encarregue de gerir o esquecimento das informações. Por fim, a porta de saída controla quando é que as informações guardadas no bloco devem sair e prosseguir o seu caminho na rede.

O uso de LSTM mostra excelentes resultados em termos de desempenho e complexidade. Através de um conjunto sequencial de pixéis, a aplicação de LSTM nesta arquitetura permite aprender a estrutura do cenário apresentado numa imagem. A aprendizagem é feita ao correlacionar pixéis vizinhos.

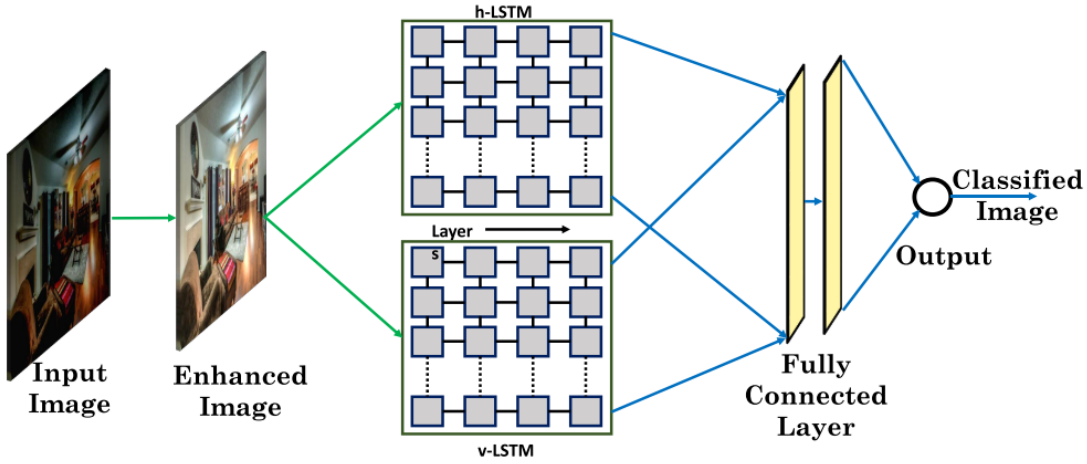

Figura 16 - Arquitetura da rede apresentada em *[5]*.

Para o treino desta rede neuronal, foi criado um *dataset* próprio. Antes de treinar a rede neuronal, as imagens são pré-processadas, através da técnica CLAHE[23]. Esta técnica calcula, para partes distintas de uma imagem, o seu histograma e junta estes dados para reajustar os valores referente à iluminação da imagem. Alguns dos resultados podem ser observados na Figura 17, onde na linha de cima são apresentadas algumas imagens antes do processamento e na linha de baixo as mesmas imagens após o processamento.

O principal modelo apresentado em [5] consiste na melhoria de uma imagem, seguida de dois componentes LSTM, um vertical e um horizontal, seguidos de uma camada *fully-connected*, como podemos observar na Figura 16.

Figura 17 - Exemplos de imagens processadas pela técnica CLAHE *[5]*.

Os resultados apresentados foram comparados diretamente com a mesma arquitetura, mas para outro *dataset*, SUN [24] (ver Seção .33.4.2). Em geral, a utilização do *dataset* REI apresenta melhor desempenho na precisão a classificar as divisões de um imóvel, com a melhor classificação de 96.92%. Esta percentagem foi conseguida através da utilização de duas camadas *fully-connected* juntamente com LSTM. Para a mesma metodologia, mas com a utilização do *dataset* SUN, foi obtida uma percentagem de 90,24%.

3.2. Classificação de Imagens utilizando CNNs

Os desenvolvimentos na classificação de imagens têm-se vindo a verificar e a mostrar resultados impressionantes ao longo dos últimos anos. A utilização de redes neuronais é uma das principais ferramentas que ajudou a alcançar estes resultados.

Dada uma imagem de entrada, a rede neuronal é capaz de atribuir um identificador previamente conhecido a esta imagem, com um determinado fator de confiança. A sua principal vantagem em comparação com outros algoritmos de classificação de imagens é a autonomia no que diz respeito à aprendizagem de características das imagens. Em vez de ser necessário introduzir manualmente o conhecimento, as CNNs aprendem a filtrar imagens de forma independente.

Desde 1998 [25] que são apresentadas soluções baseadas em redes neuronais para este tema, que vieram a servir de base para desenvolvimentos do mesmo género.

Os resultados apresentados nos primórdios da problemática de identificação e classificação de imagens [3] eram consideravelmente inferiores ao que é possível alcançar nos dias que correm. Os principais fatores que levam a este avanço são os *datasets* existentes e a capacidade de processamento.

3.2.1. AlexNet

Em 2012 surgiu um dos modelos mais influentes na área, *AlexNet* [11], que conseguiu reduzir para perto de 50% a taxa de erro apresentada no desafio *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) um enorme progresso na altura. Foi proposto o uso de uma rede neural convolucional com funções de ativação ReLU.

ILSVRC surge em 2010 [26], permitindo através do *dataset ImageNet* avaliar algoritmos para deteção de objetos e classificação de imagens. Assim, é possível detetar progressos nestas áreas.

A utilização das funções de ativação ReLU permitiu alcançar uma taxa de erro de 25% seis vezes mais rápido quando comparada com a utilização de uma outra função de ativação, *tangente hiperbólica* (*tan*), numa rede neuronal convolucional de 8 camadas, das quais cinco são convolucionais e três *fully-connected*. A taxa de erro representa o desempenho de uma rede para a fase de treino ou validação. Quanto menor o valor, melhor é o desempenho.

Na Figura 18 pode-se observar um gráfico com os valores associados a cada época e a taxa de erro a cada momento. A linha sólida representa a rede com utilização de ReLU e a linha tracejada a mesma rede com a utilização de *tan*. Este gráfico permite perceber que através do uso de ReLU a rede chega a uma taxa de erro de 25% sete vezes mais rapidamente.

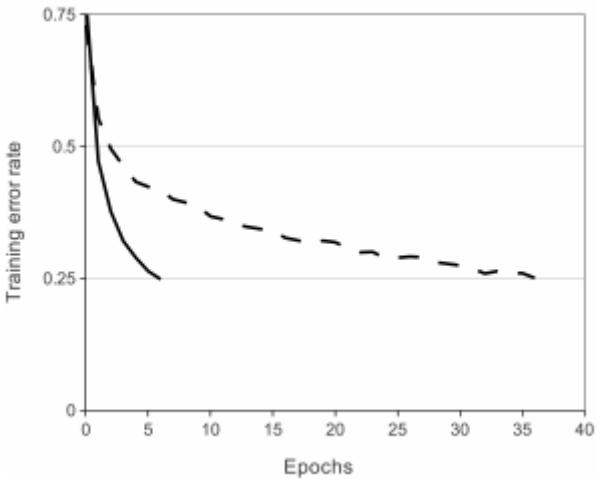

Figura 18 - Comparação de uma rede de quatro camadas com a utilização de ReLu vs tanh .[11].

Para além da utilização de funções de ativação ReLU, a rede *AlexNet* separa a rede em duas GPUs (Unidade de Processamento Gráfico) paralelas. O esquema em paralelo divide os neurónios em cada GPU, limitando a comunicação entre estes, apenas permitindo esta comunicação em determinadas camadas.

Relativamente às camadas de *pooling*, esta rede optou pelo método *max-pooling* com um filtro de dimensão 3×3 e um *stride* de valor 2, causando sobreposição do filtro, ilustrado na Figura 19, e consequentemente previne *overfitting*.

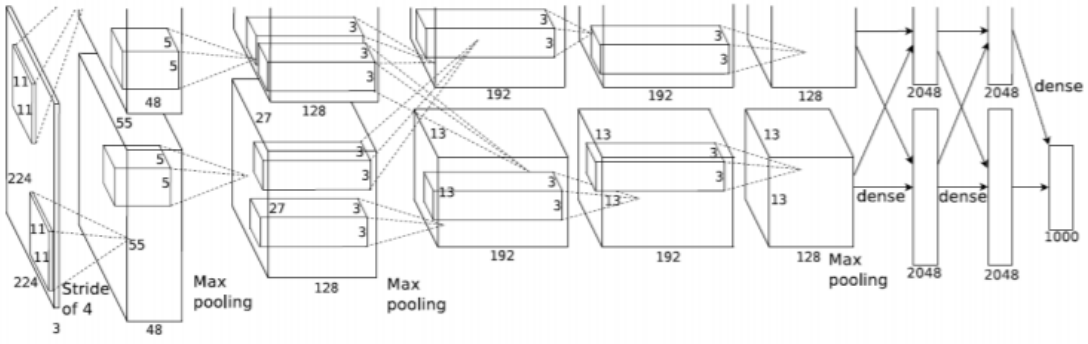

Figura 19 - Ilustração da arquitetura da rede AlexNet [11].

AlexNet é composta por oito camadas, das quais, as cinco primeiras são camadas convolucionais e as restantes são camadas *fully-connected.* O resultado da última camada *fully-connected* é dado como parâmetro de entrada a uma camada *softmax*, que faz a

distribuição de resultados para as 1000 classes possíveis. A não linearidade, *ReLU*, é aplicada nos resultados de cada camada convolucional e *fully-connected.*

Para o treino desta rede, o conjunto de imagens utilizado foi o *ImageNet* [10]. Foi através do uso deste conjunto para o treino da rede que melhores resultados foram obtidos, alcançando uma taxa de 13.5% no teste de erro top5 ILSVRC. A taxa de erro top 5 é a percentagem de vezes em que a classe correta da imagem não se encontra no top 5 das 1000 classes previstas.

O problema de *overfitting* foi combatido através de dois métodos: *data augmentation* e *dropout*.

No caso da *data augmentation* são realizadas duas operações. A primeira consiste em gerar translações e reflexões horizontais das imagens de treino. Para cada imagem de dimensão 256×256 são extraídos aleatoriamente trechos de dimensão 224×224 e reflexões horizontais destes trechos das imagens. A rede é então treinada com este subconjunto de imagens. Este método permite aumentar o tamanho do conjunto de dados por um fator de 2048. A segunda operação realizada traduz-se na alteração da intensidade dos canais RGB nas imagens de treino da rede.

O método *dropout* (já referido em 2.1.5), foi aplicado nas duas primeiras camadas *fully-connected* apresentadas na Figura 19, o que permitiu combater o fenómeno de *overfitting* substancialmente.

3.2.2. VGGNet

Ainda sobre o treino através do *dataset* ImageNet, surgiu, em 2014, uma nova arquitetura, *VGGNet* [14], onde a exploração das camadas convolucionais é aprofundada.

Esta arquitetura recebe imagens RGB de tamanho fixo de 224×224. As camadas convolucionais aplicam um filtro de dimensão 3×3. Através da aplicação de filtros de tamanho reduzido e de *stride* de valor fixo a 1, ou seja, para cada camada convolucional 3×3 é aplicado um *padding* de 1 pixel, é possível manter a resolução espacial após cada convolução.

Em termos de camadas *pooling*, são utilizadas cinco camadas de *max-pooling* após algumas das camadas convolucionais. A operação em cada camada de *max-pooling* é feita através de um filtro 2×2 e um *stride* de valor 2.

No final das várias camadas convolucionais encontram-se três camadas *fully-connected* seguidas de uma camada *softmax.* Estas camadas *fully-connected* são iguais para todas as configurações apresentadas nas várias configurações da Figura 20.

Todas as camadas apresentam funções de ativação ReLu [11].

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figura 20 - Tabela de configurações da rede. A profundidade das configurações vai aumentando da coluna esquerda para a direita, através da adição de mais colunas, apresentadas a **negrito**. [14]

Como consequência de um número maior de camadas convolucionais e da utilização de filtros de tamanho mais reduzido (3×3), esta arquitetura diminui a percentagem no teste de erro top 5, no desafio ILSVR) para 8.8%

3.2.3. ResNet

A primeira publicação do modelo *ResNet* [16] foi divulgada em 2015. Esta rede adotou alguns dos métodos utilizados nas redes *AlexNe*t [11], sendo que, primeiro, as imagens são redimensionadas aleatoriamente para um tamanho de 224×224 pixéis, pois podem ter tamanhos aleatórios e são posteriormente invertidas horizontalmente. A mesma operação de alteração da intensidade dos canais RGB nas imagens [11] é também aplicada neste modelo.

A arquitetura desta rede é inspirada na rede *VGGNet* [14]. De um modo geral as camadas convolucionais têm filtros de dimensão 3×3, com *stride* de valor 2 e é nestas mesmas camadas que a subamostragem é feita. No final da rede existe uma camada de *average-pooling* e uma camada *fully-connected* seguida de uma camada *softmax*.

Uma das principais contribuições deste modelo foi aprendizagem residual, que acabou por ser usada como referência em modelos mais recentes. A aprendizagem residual com *skip-connections* serve de atalho para as camadas mais profundas acederem diretamente às características das camadas anteriores. Através desta nova aprendizagem, redes mais profundas, isto é, com mais camadas, podem combater o problema de *vanishing/exploding* gradientes.

Os atalhos, ilustrados na Figura 21, podem ser usados diretamente quando as dimensões de entrada e saída entre as camadas são iguais. Quando as dimensões divergem, pixéis de "preenchimento" podem ser acrescentados.

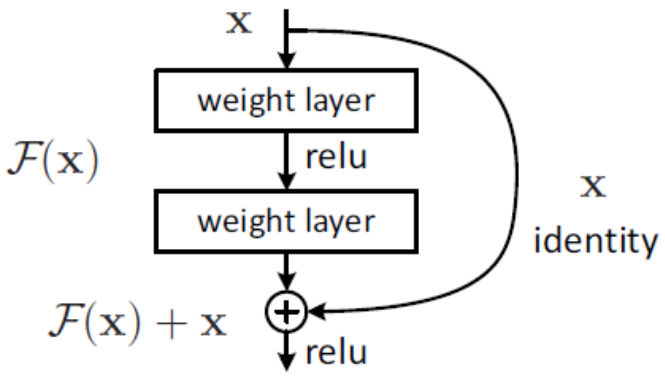

Figura 21 - Aprendizagem residual [16].

A informação das características da imagem é transmitida pela rede mais facilmente, ajudando no treino da rede neuronal. Durante o desenvolvimento e estudo da rede *ResNet*

ficou provado que a aplicação de camadas sequenciais, formando uma rede mais profunda, nem sempre facilita a classificação de imagens, podendo até ter precursões indesejadas.

No mesmo ano, esta arquitetura ganhou o primeiro prémio no desafio do ImageNet (ILSVRC 2015), baixando a taxa de erro top5 para 4,49%.

3.2.4. GoogleNet - Inception-v1

O primeiro modelo *Inception* [27] surgiu em 2014. Este modelo foi um dos algoritmos pioneiros a fundamentar a ideia de que as camadas das CNN não precisam de estar sempre organizadas sequencialmente. Conseguiu provar que é possível melhorar a performance aumentado a largura da rede e não só a profundidade.

A premissa para a construção deste tipo de rede foi a variação da localização de informação numa imagem. Quando a variação é significativa, a escolha do tamanho de um filtro é um problema, podendo observar-se exemplos na Figura 22. Para captar informações distribuídas pela imagem toda é preferível um filtro mais largo, por outro lado, para informações locais, é preferível um filtro de tamanho mais reduzido.

Figura 22 - Da esquerda para a direita: Um lavatório a ocupar um pequeno espaço na imagem, um lavatório a ocupar parte da imagem e, por fim, um lavatório a ocupar quase toda a imagem (Imagens retiradas do REI dataset *[5]*).

Deste modo foram aplicados filtros de vários tamanhos a trabalhar no mesmo nível, construindo uma rede mais larga do que "profunda", como se pode observar na Figura 23.

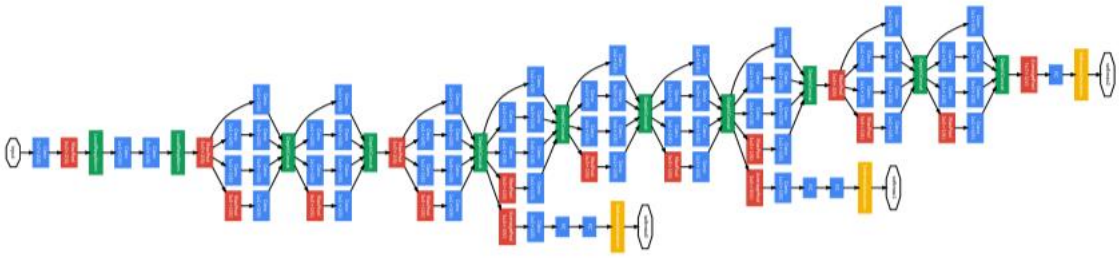

Figura 23 - Arquitetura da rede GoogleNet.

Modelos como o *Inception-v2*, *Inception-v3* [28], *Inception-v4* e *Inception-ResNet* [29], inspirados nesta rede e na *ResNet*, obtiveram o melhor desempenho demonstrado até os dias de hoje, fundindo os módulos "*Inception*" presentes na arquitetura do modelo *Inception* e os blocos residuais, presentes na arquitetura do modelo *ResNet*.

O melhor resultado apresentado por esta arquitetura é apresentada pelo modelo *Incpetion-v3* em 2015 com uma taxa de erro top5 de 3.58%.

3.3. Qualidade de imagem, estética e representatividade

A estética de uma imagem passa, essencialmente, por tudo o que não tem necessariamente a ver com o seu conteúdo. Ou seja, iluminação, contraste, ruido, *blur*, distância focal, ângulo, etc. Dentro do mesmo contexto, existem regras básicas de fotografia [30] para determinados conteúdos visuais de imagens, como por exemplo a regra dos terços. O não seguimento deste tipo de regras pode influenciar negativamente a classificação da imagem quanto à sua estética.

O julgamento humano da estética de uma fotografia advém de experiências passadas e da emoção que a imagem transmite, seja através do conteúdo, formas, orientações, do grupo de cores e objetos [31].

Para além da definição da qualidade estética ser subjetiva, não é certo em que medida os atributos definidos para a representação de estética influenciam a qualidade da imagem. O que para uma pessoa pode ser considerado estético, para outra pode não o ser. O contraste numa fotografia pode ter mais peso na avaliação estética para um determinado grupo de pessoas do que para outros.

Para tentar generalizar esta avaliação, [32] R. Datta and J. Z. Wang começa por definir quais os atributos que compõem a estética de uma imagem, adaptando depois para referências matemáticas de modo a ser possível computar automaticamente estas

características e consequentemente ter um processo automatizado. As principais desvantagens desta solução é não ser garantida a cobertura de todos os princípios da fotografia e o dispendioso custo computacional.

Em [33] X. Lu et al. defendem que a definição manual dos atributos e características que compõem a estética de uma imagem são abstratos e não contemplam todos os casos possíveis. É, em muitos casos, difícil implementar computacionalmente estas características obtidas manualmente, chegando a ser, em alguns casos, meras aproximações, o que torna a definição manual de atributos que representam a estética de uma imagem um processo pouco viável.

Com o intuito de fugir à definição manual dos atributos que influenciam a qualidade estética de uma imagem, L. Marchesottiet al. apresentam em [34] uma solução baseada na utilização dos descritores genéricos de uma imagem, mais concretamente o *Bag-of-Visual-Words* (BOV) [35] e o *Fisher Vector* (FV) [36], mostrando obter bons resultados.

Ao longo dos anos foram-se desenvolvendo diferentes técnicas para uma automatização na classificação da estética de imagens. Para fugir aos descritores genéricos de características de imagens, X. Lu et al. propõem em [33] a aprendizagem destas peculiaridades estéticas diretamente a partir das imagens, através do uso de redes neuronais convolucionais. A arquitetura apresentada pelos autores em [33] é inspirada na AlexNet [11], onde as características das imagens são extraídas diretamente destas. Contudo, é desenvolvida uma arquitetura da rede diferente, pois assumem que a estética de uma imagem é composta por elementos visuais tanto locais como globais da mesma imagem. Para tal, foi introduzida uma coluna dupla de entrada capaz de receber dois parâmetros (elementos visuais locais e globais) em paralelo, ilustrado na Figura 24.

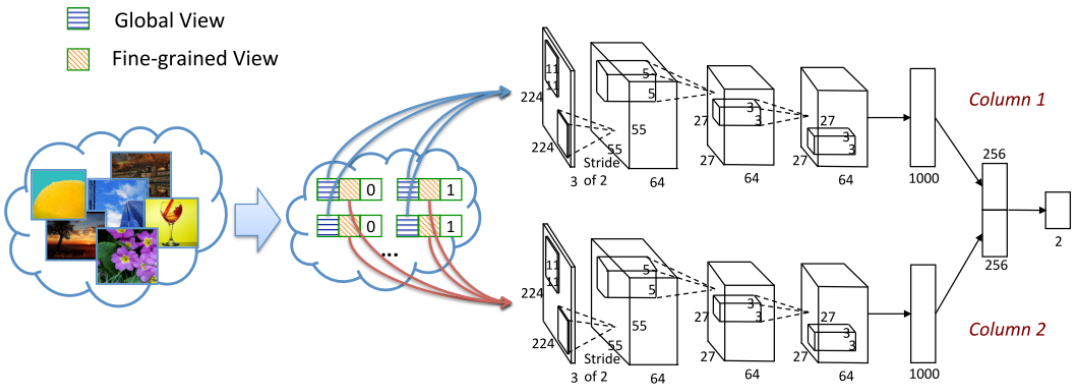

Figura 24 - Modelo RAPID [33].

Esta solução obteve os melhores resultados até à data de desenvolvimento, em comparação com outras soluções apresentadas para o mesmo conjunto de imagens [37]. A apresentação dos resultados é binária no sentido em que devolve apenas se a imagem é estética ou não. Ao contrário desta solução, [38] apresenta um trabalho onde o valor de saída corresponde a um grau de estética para cada imagem numa escala de 1 a 10. Aqui, foi aplicada a metodologia *transfer learning* tendo como base a rede *Caffe* [39], que, por sua vez, foi inspirada na rede *AlexNet* [11]. A razão pela qual o método de *transfer learning* permite alcançar resultados muito satisfatórios quando aplicado ao contexto certo, deve-se ao facto de as primeiras camadas da rede já estarem devidamente treinadas para reconhecer padrões e formas mais básicas das imagens, sendo apenas necessário treinar as camadas finais, de forma a que estas possam aprender quais destas características são relevantes na classificação de cada classe apresentada.

A solução desenvolvida teve não só os melhores resultados apresentados até à data do seu desenvolvimento, como também provou que a estética numa imagem é um atributo mais global que local. A mesma conclusão é apresentada em [40]. Nesse trabalho, é feita uma comparação direta entre a rede *AlexNet* [11] e a *VGGNet* de 16 camadas [14], usando duas bases de dados de imagens diferentes: AVA [37] e CUHKPQ [41]. Também aqui, a utilização do método *transfer learning* permitiu que a rede não tivesse de ser treinada a partir do zero e a limitação da dimensão do *dataset* fosse ultrapassada. Para além dos excelentes resultados, as principais conclusões obtidas são: a utilização da rede *VGGNet* permite obter melhores resultados, ao passo que a rede *AlexNet* permite obter bons resultados num espaço de tempo mais reduzido.

Tal como descrito anteriormente [38], comparando os resultados da utilização de perspetivas globais e locais da mesma imagem, ilustração presente na Figura 25, é possível chegar a melhores resultados apenas fazendo uso de perspetivas globais.

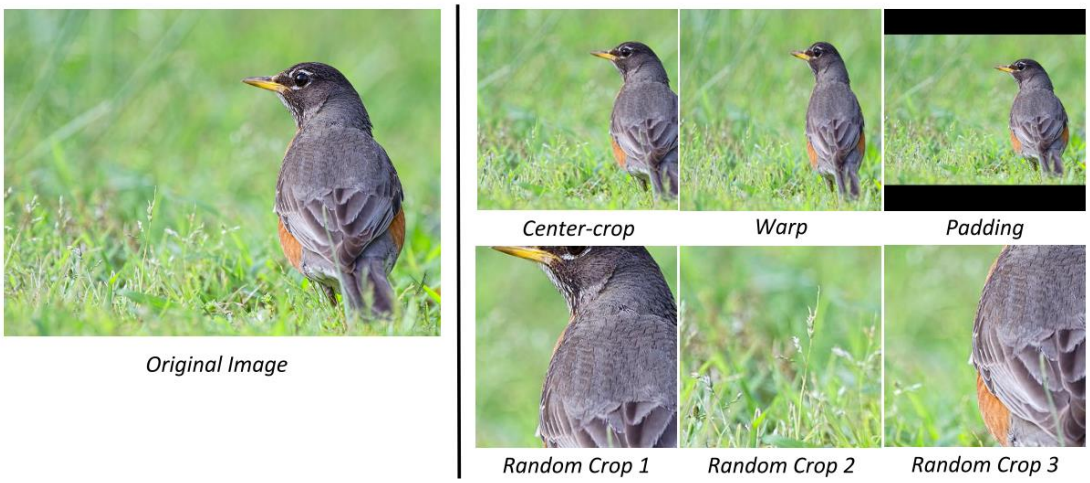

Figura 25 - Imagem de perspetivas globais (esquerda) e imagens de perspetivas locais (direita) - RAPID [33].

Atendendo ao seu excelente desempenho, a abordagem *transfer learning* foi aplicada nas experiências desenvolvidas no presente trabalho, tal como se refere no capítulo 4.

Em [42], Tian, Long e Lv. defendem que a avaliação do conteúdo visual de imagens, por comparação com outras, deve ser apenas feita em imagens sobre o mesmo tema e conteúdo visual. Uma vez que uma imagem pode transmitir sentimentos positivos, negativos ou não transmitir qualquer emoção [43], a comparação direta seria injusta e consequentemente pouco conclusiva.

A avaliação estética de uma imagem pode ser dividida segundo duas soluções diferentes: a avaliação binária ou a avaliação que prevê uma classificação dentro de um intervalo de valores.

Mean opinion score (MOS) [44] é um dos métodos mais usados para ultrapassar a dificuldade que é avaliar a qualidade e estética de uma imagem. Através de MOS pode-se determinar a qualidade subjetiva de um componente digital, seja esta uma imagem, vídeo, som, etc, pois este método consiste na média de resultados de classificações/ opiniões por parte de participantes. A classificação é feita numa escala entre as seguintes classes: excelente, bom, razoável, pobre e mau que por sua vez são mapeadas para valores entre 5 e 1.

Uma variante da utilização da escala destes cinco valores é apresentada em [45], onde os participantes da classificação de imagens tem uma barra deslizante entre os cinco níveis: excelente, bom, razoável, pobre, mau. A posição da barra é depois convertida para um valor entre 0 e 100. O valor MOS obtido para cada imagem representa a qualidade visual das mesmas.

O valor da classificação de uma imagem, dentro de uma determinada escala, pode ser calculado de diferentes formas. Pode ser descrito como um resultado único/real ou como uma distribuição de resultados [37]. O resultado único é obtido através da média de todos os valores de classificação obtidos, solução que não permite obter informação quanto ao grau de concordância e ao consenso dentro das diversas opiniões. Tal pode-se observar no exemplo ilustrado na Figura 26, onde ambas as imagens obtiveram uma classificação média de 4 valores numa escala de 1 a 5, mas o consenso entre as classificações atribuídas pelas várias pessoas é bastante diferente para as duas imagens.

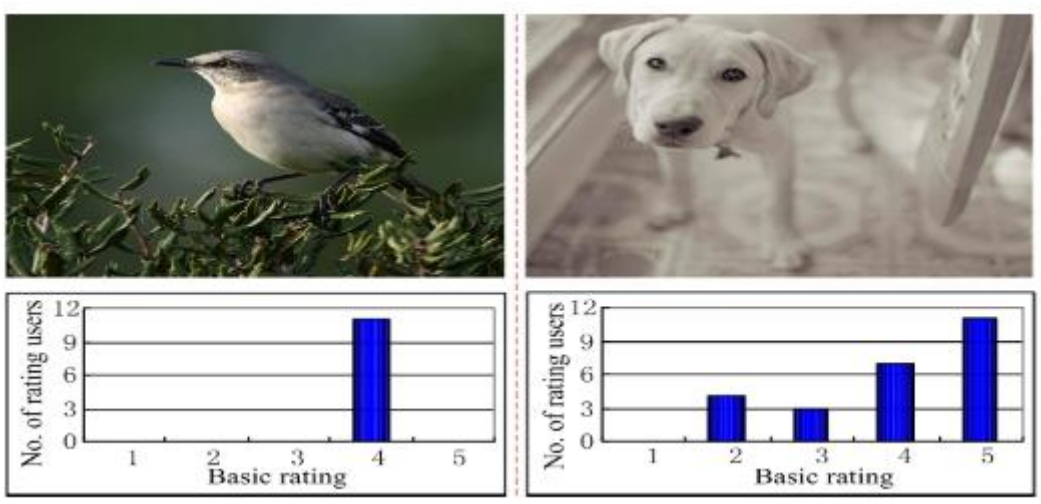

Figura 26 - Duas imagens e as classificações associadas ao número de pessoas, numa escala de 1 a 5 [46].

O trabalho apresentado em [46] introduz uma solução onde, para ter em conta a variância de opiniões, é calculada a proporção com que cada valor da escala é escolhido pelas pessoas que classificam a imagem. Assim, é associada a cada imagem uma etiqueta que descreve esta proporção, calculada através do número de pessoas que votaram num determinado valor e sobre o número total de pessoas que votaram naquela imagem. No caso da Figura 26, a imagem da direita ficaria com a seguinte distribuição associada: (0,0.16, 0.12, 0.28, 0.44).

Esta representação permite guardar mais informação original e a distribuição permite avaliar quão subjetiva é a qualidade visual de determinada imagem, sabendo que quanto maior o desvio padrão, mais subjetiva é a qualidade visual.

Em [38] faz-se uso do *dataset* AVA para treinar um modelo CNN para a classificação de estética de imagens de um modo não-binário, ou seja, através da avaliação dentro de uma determinada escala de valores. Este *dataset* é composto por variadas imagens com diferentes classificações estéticas obtidas pelo julgamento de enumeras pessoas, expresso através de uma classificação numérica entre os valores 1 e 10.

Em 2015, [47] avança com uma solução para a identificação do tipo de distorção (*blur* gaussiano, compressão, etc.), bem como para a classificação da qualidade de uma imagem. Estas duas tarefas são realizadas em simultâneo na mesma CNN. Nesta abordagem L. Kang et al. definem o tipo de distorção de uma imagem partindo-a em sub-imagens, identificando cada distorção neste sub-conjunto e o maior número de ocorrências de um determinado tipo de distorção é associado à imagem. A qualidade da imagem é definida pelo valor médio de qualidade de todas as sub-imagens.

Como se pode observar na Figura 27, a arquitetura da rede apresentada é composta por duas camadas convolucionais, cada uma com *pooling*, duas camadas *fully-connected* e uma camada de saída. Na camada de saída tem-se os resultados de ambas as tarefas a correrem simultaneamente.

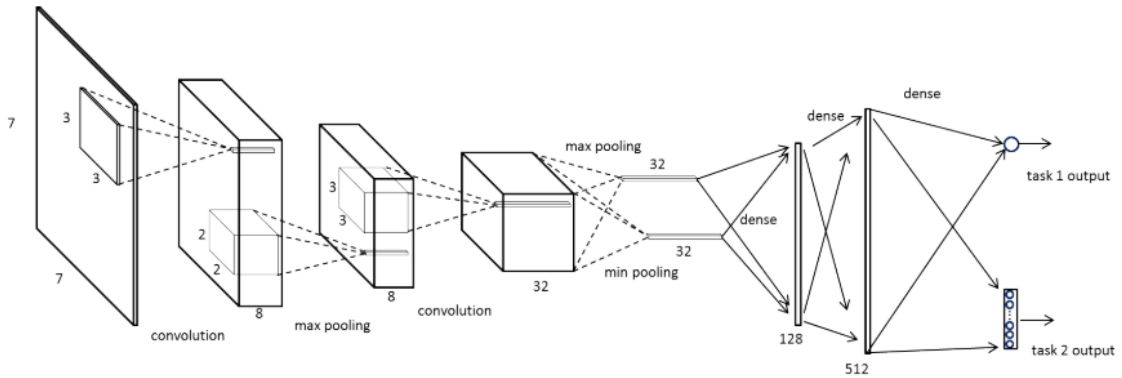

Figura 27 - Arquitetura da rede apresentada em [47].

Esta abordagem teve um desempenho muito semelhante ou melhor que as soluções apresentadas à data deste desenvolvimento.

Ainda sobre a identificação de distorção em imagens, [48] apresenta uma solução onde inicialmente o tipo de distorção de imagem é identificado e posteriormente a sua intensidade é mapeada para uma escala MOS. A classificação das imagens é apresentada através da utilização de redes neuronais convolucionais, alcançando uma *accuracy* de 94.14% e obtendo um melhor desempenho em comparação com soluções que utilizam SVMs.

3.4. Datasets

Um conjunto de dados (*dataset*) é uma coleção organizada de dados que se utiliza para o treino e para validação da rede neuronal convolucional. Nas seguintes secções apresentam-se alguns dos *dataset* mais influentes na temática de classificação de imagens, nomeadamente de imagens de interior e exterior de imóveis.

3.4.1 ImageNet

ImageNet é um dos maiores *datasets* existentes no universo da classificação de imagens. À data de apresentação da primeira "versão" [10], este conjunto de dados apresentava um total de 3,2 milhões de imagens, pertencentes a 12 sub-categorias, que por sua vez ainda se dividem em conjuntos diferentes. A estrutura deste *dataset* é hierárquica, baseada na estrutura do conjunto *WordNet* [49].

Na Figura 28 é apresentado um exemplo de 6 sub-categorias (*sub-trees*) existentes no *ImageNet*, para uma delas é possível ver o número de conjuntos que o compõem (*Synsets*) e o número a médio de imagens por conjunto.

Subtree	# Synsets	Avg. synset size	Total # image
Mammal	1170	737	862K
Vehicle	520	610	317K
GeoForm	176	436	77K
Furniture	197	797	157K
Bird	872	809	705K
MusicInstr	164	672	110K

Figura 28 - Tabela representante de seis sub-categorias do dataset ImageNet [10].

3.4.2 SUN

O conjunto de imagens SUN, *Scene Understanding,* [24] é um *dataset* composto por 899 categorias diferentes e um total de 130,519 imagens.

Ao contrário do *ImageNet*, este conjunto está orientado ao contexto dos cenários / espaço e não ao objeto representado na imagem. Exemplos de categorias encontradas neste conjunto de imagens são: esquadra, estádio, arquipélago, etc.

3.4.3 REI

O *dataset* REI apresentado em [5] foi criado para a classificação de imagens referentes à área da imobiliária. As imagens que compõem o REI foram adquiridas de *sites* de imóveis bem como motores de pesquisa *online,* com o intuito de classificar: o espaço que a imagem representa, o material utilizado em bancadas e a classificação do chão.

REI é composto por 5967 imagens distribuídas por seis classes diferentes para a classificação do espaço: casa de banho, cozinha, sala de estar, quarto, entrada e o quintal. Cada classe tem pelo menos 700 imagens.

O *dataset*, *REI,* servirá como ponto de partida para o estabelecimento do *dataset* utilizado no treino da rede neuronal do presente trabalho.

3.4.4 *Aesthetic Visual Analysis* (AVA)

O *dataset* AVA é composto por mais de 250 000 imagens com variadas anotações [37] :

- Anotações estéticas – Para casa imagem existe uma distribuição de resultados que correspondem às avaliações de cada voto individual. A média de número de votos por imagem é de 210 votos.

- Anotações semânticas – O conjunto de imagens contém 66 descritores distintos da semântica das imagens. Cada imagem pode ter um ou dois descritores associados.
- Anotações do estilo de fotografia – Cada fotografia pode ter um estilo diferente, influenciado pelo modo e configurações de câmara com que esta foi capturada. Foram distinguidos 14 estilos diferentes: *Complementary Colors, Duotones, High Dynamic Range, Image Grain, Light on White, Long Exposure, Macro, Motion Blur, Negative Image, Rule of Thirds, Shallow DOF, Silhouettes, Soft Focus, Vanishing Point.*

À data de apresentação do conjunto de imagens AVA, este apresenta ser um *dataset* bastante completo quando comparado com outros conjuntos de imagens com informação da estética associada. Na Figura 29 é apresentada a comparação de propriedades de diferentes *datasets*. Os valores Y representam a presença de determinadas propriedades, sendo o AVA o mais completo por ser um *dataset* bastante robusto em número de imagens e pela presença de informação de: distribuição de resultados/classificações, anotações ricas, descritores de semântica e descritores de estilo.

	AVA	Photo.net	CUHK	CUHKPQ	ImageCLEF
Large Scale	Y	N	N	N	N
Scores distr.	Y	Y	N	N	N
Rich annotations	Y	N	Y	Y	Y
Semantic Labels	Y	N	N	Y	Y
Style Labels	Y	N	N	N	Y

Figura 29 - Comparação das propriedades dos conjuntos de imagens com informação da qualidade estética associada [32].

3.5. Métricas de avaliação

Na avaliação de possíveis soluções para a classificação de imagens, designadamente recorrendo a um sistema de classificação automático, devem ter-se em conta algumas métricas habitualmente usadas.

Os termos *True Positive* (TP) *True Negative* (TN), *False Positive* (FP) e *False Negative* (FN) facilitam generalização da interpretação e avaliação do desempenho de algoritmos de classificação automática no contexto da classificação. Para melhor compreender estes três termos, imagine-se a criação de um modelo que recebe imagens de entrada e deve prever se estas correspondem ao interior de um imóvel.

Na classificação binária existem duas classes possíveis de resultados: positivos e negativos [50]. Neste contexto, a classe positiva contém todas as imagens relacionadas com interior de um imóvel e, consequentemente, as imagens de exterior são correspondentes à classe negativa.

- TP – *True positives*: Os verdadeiros positivos correspondem, como o nome indica, à classificação correta de um elemento tendo em conta a sua classe. No contexto do exemplo acima referido, um verdadeiro positivo ocorre quando o modelo considera que a imagem de uma cozinha corresponde ao interior de uma casa.

- FP – *False positives*: Os falsos positivos representam a classificação errada de um elemento cuja natureza foi incluída na classe positiva. Se o modelo classificar uma imagem de uma varanda como o interior de um imóvel está-se perante um FP.

- FN – *False negatives*: Um falso negativo é o oposto do falso positivo. Um elemento pertencente à classe positiva é classificado como negativo. Um exemplo prático seria um quarto de uma casa ser classificado como uma imagem de exterior.

- TN – *True negatives*: Tal como os TPs, os verdadeiros negativos correspondem a todas as classificações que classificaram corretamente as classes negativas.

Para interpretar os resultados de todas as imagens classificadas e avaliar o desempenho de qualquer modelo são apresentados três parâmetros: *precision*, *recall* e F1-*measure*. Estes parâmetros descrevem estatisticamente o total dos resultados apresentados pelas métricas anteriormente descritas [51].

A precisão (*precision*) é uma métrica que identifica a frequência com que o modelo em questão esteve correto quando classificou uma imagem como pertencente à classe positiva [50]. A precisão é dada por:

$$Precision = \frac{TP}{TP+FP} \qquad (1)$$

A sensibilidade (recall), também conhecido como true positive rate (TPR), avalia o desempenho de modelos respondendo à seguinte questão [50]: de entre o número total de imagens a classificar positivas, quantas é que o modelo identificou corretamente? Esta métrica é dada por:

$$Recall = \frac{TP}{TP+FN} \qquad (2)$$

Uma outra medida de desempenho é a F1- *measure* que calcula a combinação entre a precisão e o *recall*, mais precisamente a média harmónica entre a precisão e o *recall* [52]. Uma média considera todos os valores de igual modo, enquanto que uma média harmónica dá mais ênfase e coloca maior peso nos valores mais baixos, desequilibrando os valores mais altos.

Para melhor compreensão, esta métrica é dada por:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \qquad (3)$$

Os valores de F1 variam entre 0 e 1, com melhores resultados quando aproximado de 1, ou seja, quando tanto a precisão como a sensibilidade têm valores altos.

Existe ainda uma outra métrica de avaliação de desempenhos de um classificador, a exatidão (*accuracy*). *Accuracy* é a fração de classificações corretas face ao número total de amostras classificadas.

Para classificações binárias, a exatidão é dada por:

$$Accuracy = \frac{TP+TN}{TN+TP+FN+FP} \qquad (4)$$

Na avaliação e comparação entre modelos, utilizam-se métricas como, por exemplo, a exatidão. Apesar destas métricas serem adequadas para uma comparação direta entre modelos e arquiteturas, estas não servem para afinar os modelos diretamente no seu processo de treino.

Outros tipos de funções permitem também contribuir para a otimização do processo de treino, como as funções de perda, *loss functions*, que aplicadas diretamente no processo de treino do modelo podem ajudar a otimizá-lo. O valor calculado através de uma função

de perda é denominado por *loss*/perda. Nos algoritmos de aprendizagem automática é expectável que um baixo valor de perda seja sinónimo de um valor de exatidão alto.

As funções de perda têm como objetivo a avaliação de uma possível solução de modelo, isto é, um conjunto de pesos ideal. Ou seja, o valor de perda serve como referência para a atualização dos pesos utilizados no modelo.

Relativamente às redes neuronais desenvolvidas e treinadas para participar no desafio ILSVRC, o resultado destas é muitas vezes comparado através de duas métricas idênticas, o top-1 *error rate* e o top-5 *error rate*. Este tipo de métricas é utilizado em classificações de múltiplas classes Estas taxas de erro correspondem à percentagem de imagens de teste cuja classe correta não estava no *top* 1 ou 5 de classes com maior percentagem de correspondência [11]. Numa classificação binária, o top-1 error rate seria equivalente à *accuracy*.

Capítulo 4

4. Experiências e Resultados

O presente trabalho tem como principal motivação contribuir para a classificação de imagens de imóveis (interior / exterior), não só no que se referia à identificação desta caraterística, como também da classificação da qualidade representativa das imagens.

Para tal, foram investigados os trabalhos realizados até à data, focando a análise bibliográfica não só no uso de redes neuronais convolucionais, como também no problema em questão. Após a revisão bibliográfica, foi identificado um trabalho cujo problema é muito idêntico, designadamente a classificação de imagens referentes às diferentes divisões de um imóvel. Para o conjunto de imagens apresentado, apenas serão considerados dois espaços: o interior e o exterior de imóveis, limitando-se assim a diversidade do contexto visual.

Uma vez identificado o problema a analisar, procurou-se solucioná-lo através da utilização de redes neuronais convolucionais, alargando o âmbito do trabalho à classificação da qualidade representativa das imagens analisadas. Para tentar evitar inconsistências a nível de classificação, isto é, ter vários graus de classificação para a mesma imagem, apenas se considera a imagem como muito ou pouco representativa, tendo assim uma classificação binária. O principal foco serão aspetos como a focagem, iluminação, bem como o enquadramento.

Para tal, houve uma fase inicial de recolha e tratamento de dados de teste, designadamente de imagens, à qual se seguiu a fase de treino de uma rede existente (VGG16) e a validação das classificações obtidas.

4.1. Dataset InOut

As redes neuronais convolucionais têm ganho grande dimensão na área da computação visual. Os algoritmos e metodologias com maior taxa de sucesso na classificação e reconhecimento de imagem têm como base o uso destas redes. Contudo, para que as CNN tenham um bom desempenho, um dos principais requisitos é a disponibilidade de um vasto repositório de imagens previamente classificadas para o treino da rede.

Assim, o trabalho desenvolvido no âmbito da presente tese iniciou com a criação de um *dataset* para classificação de imagens de interior e de exterior, constituído por um conjunto de imagens obtidas a partir do *dataset* REI, complementado por imagens retiradas do *Flickr* e do *Google Images*.

As imagens foram separadas em duas classes distintas: imagens de interior e exterior.

Para a construção do *dataset* InOut, foram aglutinadas as imagens das classes de casa-de-banho, cozinha, sala de estar e quarto do *dataset* REI e agregadas em apenas uma nova classe, denominada por interior. Para a classe de exterior, juntaram-se as imagens das classes de entrada e quintal. Foram utilizadas 1865 imagens de interior e 1629 imagens de exterior do *dataset* REI.

Na Figura 30 pode-se observar um exemplo da separação das imagens pelas duas classes: imagens de interior na primeira linha e imagens de exterior na segunda linha.

Figura 30 - Exemplo de imagens encontradas no *dataset* REI *[5]*.

A este conjunto inicial de imagens retiradas do *dataset* REI foram ainda adicionadas 3508 imagens de imóveis extraídas a partir do *Flickr* e do *Google Images*.

Estas imagens são etiquetadas, possibilitando uma pesquisa mais específica e facilitada. O conjunto de dados foi igualmente dividido entre as duas classes, contando com 1754 imagens de interior e 1754 imagens de exterior.

As imagens originais possuíam variados tamanhos, pelo que foi necessário redimensionar todas para uma dimensão fixa de 224 × 224.

Assim, o número total de imagens é de 3501, tanto para as imagens do espaço interior, como do espaço exterior, perfazendo um total de 7002 imagens.

As imagens representativas de cada classe foram aleatoriamente divididas em dois conjuntos:

- Conjunto de treino, correspondente a 6 420 imagens.
- Conjunto de validação, correspondente a 582 imagens.

4.2 Dataset – RQI (representative quality of an image)

Para uma melhor comparação entre este trabalho e os trabalhos realizados até à data, foi realizada uma segunda experiência onde, para a mesma arquitetura de rede neuronal, o treino foi feito a partir do conjunto de imagens InOut referido anteriormente, mas também com alguma pré-edição, que se descreve nos parágrafos seguintes.

Para o treino da rede neuronal na classificação representativa de imagens foi necessário editar algumas das imagens de modo a comprometer a qualidade destas e consecutivamente a qualidade representativa das mesmas, de modo a criar o *dataset* RQI.

Após a edição foram geradas duas classes: uma com boa qualidade representativa, com base nas imagens não processadas e uma classe relativa às imagens com má qualidade representativa.

O processamento aplicado às imagens consistiu em:

- Alterar a exposição: Na área de fotografia, a exposição (*exposure*) é dada como a quantidade de luz que chega ao sensor de uma câmara. Quando entra demasiada luz, falamos de sobre-exposição (*overexposure*).

  Cada imagem do conjunto foi editada de modo a obter uma versão *overexposed* da mesma, como podemos observar na Figura 31 - Comparação da imagem original(direita) e da imagem após edição(esquerda) - *overexposure*.

Figura 31 - Comparação da imagem original(direita) e da imagem após edição(esquerda) - *overexposure*.

- Inserir *blur*: *Blur*, também conhecido como "suavização", consiste em remover pixels que sejam discrepantes que aplicam ruído na imagem. Como em tudo, quando aplicado em demasia, a qualidade da imagem pode ser condicionada. Deste modo, será aplicado *blur* nas imagens, provocando o efeito de desfocagem das mesmas.

  Existem várias formas do efeito *blur* ser aplicado e calculado numa imagem. Optou-se pelo *blur gaussiano*. Este tipo de suavização é feito a partir da função gaussiana, substituindo os pixels por uma média ponderada dos pixels vizinhos. O peso que cada pixel tem na média ponderada deriva da função gaussiana, dando assim mais peso aos pixels mais próximos. Na Figura 32 podemos observar a imagem original após a aplicação do efeito *blur*.

Figura 32 - Comparação da imagem após edição (*blur*) com a imagem original.

- Reduzir a luminosidade: A redução da luminosidade foi aplicada às imagens, comprometendo a qualidade representativa destas, pois o nível de perceção desceu consideravelmente. Na Figura 33 pode-se observar a diferença entre a imagem original e a mesma imagem após a redução da luminosidade.

Figura 33 - Comparação da imagem após edição (redução de luminosidade) com a imagem original.

Para 2 500 imagens retiradas do *dataset* InOut foram realizadas cada uma destas edições, conseguindo assim um total de 7 500 imagens com fraca qualidade representativa.

Algumas imagens ainda foram recortadas em sub-imagens de modo a obter mais imagens com fraca qualidade representativa, como se pode observar no exemplo da Figura 34.

Figura 34 - Exemplo de uma imagem recortada e um sub-bloco.

Estas imagens que foram editadas tiveram de ser analisadas manualmente para garantir a fraca qualidade representativa em todas elas. Nem todas puderam ser consideradas no *dataset*.

Foram também consideradas imagens originalmente com fraca qualidade representativa, que não foram previamente trabalhadas, das quais se apresentam exemplos representativos na Figura 35.

Figura 35 - Três imagens sem qualidade representativa do interior do imóvel.

A junção de imagens originais com a as imagens editadas permitiu aumentar o tamanho do conjunto de imagens para um total de 13 578 imagens. Foram aleatoriamente escolhidas um número de imagens de cada classe para os dois conjuntos:

- Conjunto de treino, correspondente a 10 906 imagens.
- Conjunto de validação, correspondente a 2 672 imagens.

O conjunto de treino subdivide-se em:

- 5453 imagens com qualidade representativa
- 5453 imagens sem qualidade representativa

O conjunto de validação subdivide-se em:

- 1363 imagens com qualidade representativa
- 1363 imagens sem qualidade representativa

4.3 Classificação de imagens de interior e exterior de imóveis

4.3.1 – Utilização da rede neuronal VGG16

Neste contexto, pretendia-se estudar qual o algoritmo melhor para classificar imagens captadas a partir do interior ou do exterior de um imóvel.

Decidiu-se utilizar o conceito *transfer learning* com esta rede neuronal pré-treinada pelo *dataset* ImageNet. Com efeito, é cada vez menos comum treinar uma rede neuronal convolucional de raiz, pois é difícil encontrar um *dataset* robusto em termos de número de imagens. *Transfer learning* é um método de *machine learning* que consiste na reutilização de um modelo pré-treinado, utilizando-o para um objetivo idêntico. Este conceito existe para evitar a necessidade do poder computacional e do tempo para desenvolver e treinar redes neuronais.

No caso específico de reconhecimento de imagens de interior / exterior de imóveis não é possível usar um *dataset* existente na sua totalidade, tendo-se identificado na seção 4.1 as alterações e complementos que foi necessário realizar à base de dados REI, para proceder ao desenvolvimento da primeira experiência realizada.

Deste modo, é possível criar redes neuronais com bom desempenho na classificação de imagens, sem necessitar de um conjunto de dados tão elevado como quando uma rede neuronal é desenhada e treinada de raiz. As primeiras camadas da rede já estão devidamente treinadas para reconhecer padrões e formas mais básicas das imagens, sendo apenas necessário treinar as camadas finais, de forma a que estas possam aprender quais destas características são relevantes na classificação de cada classe apresentada.

Foi utilizado o modelo VGG16 como modelo de partida para a classificação de imagens de interior e exterior de imóveis.

Como referido no Capítulo 3.2.2, o modelo VGG16 foi construído para conseguir classificar as imagens do conjunto *ImageNet*, onde o número de possíveis classes a classificar é bastante elevado.

O conjunto de imagens construído para esta classificação necessitou de reajustes para que estas fossem aceites pela arquitetura do VGG16. Para além do redimensionamento de todas as imagens para 224×224 pixels, como referido anteriormente, foi também necessário redimensionar os coeficientes RGB passando estes de valores entre 0 e 255, para valores entre 0 e 1, através da normalização, tornando assim o processamento das imagens possível.

Posteriormente foram removidas as últimas três camadas *fully-connected* desta arquitetura, responsáveis estas pela classificação final das imagens, reconhecendo características específicas das classes. As camadas convolucionais não foram alteradas.

Posto isto, as últimas camadas são treinadas de raiz com o novo conjunto de imagens. O treino destas camadas é feito correndo várias épocas das imagens. Inicialmente optou-se por correr apenas 20 épocas com *batch size* de 15, pois os gráficos de exatidão e *loss* mostraram não estar a progredir com o aumento do número de épocas.

Os resultados obtidos nesta experiência podem ser observados na Figura 36, em que as abcissas representam o número de épocas e as ordenadas representam a exatidão e a perda, respetivamente.

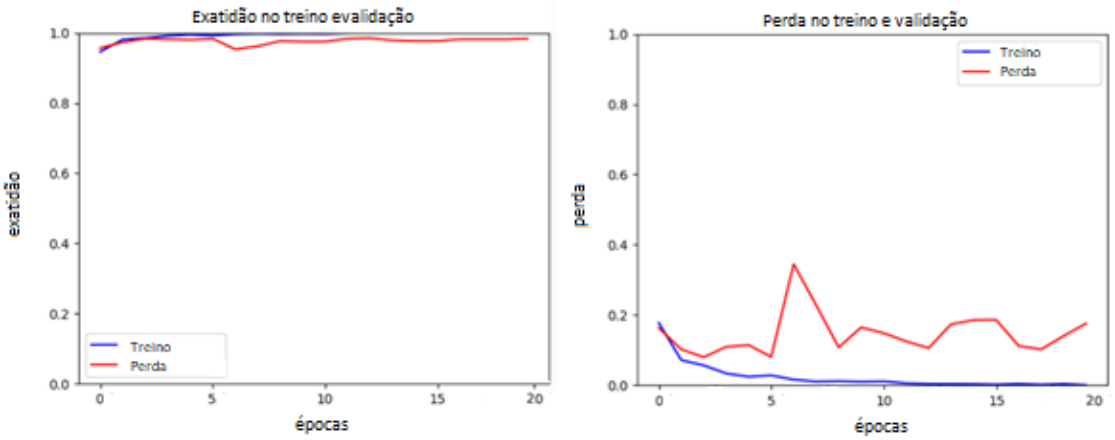

Figura 36 - Gráficos com evolução da exatidão (esquerda) e da perda (direita), utilizando a técnica de *transfer learning* aplicada ao modelo VGG16 – 20 épocas.

De seguida foram feitas mais duas experiências, correndo a rede em 50 e 100 épocas, resultando nos gráficos apresentados na Figura 37 e na Figura 38 respetivamente, em que os eixos têm o mesmo significado. Os resultados não mostram melhorias comparativamente à primeira experiência com 20 épocas.

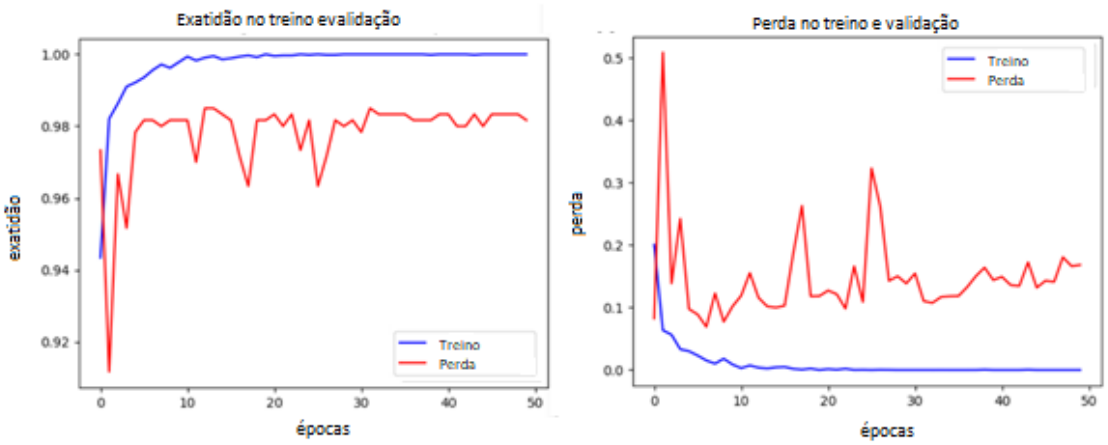

Figura 37 - Gráficos com evolução da exatidão (esquerda) e da perda (direita), utilizando a técnica de *transfer learning* aplicada ao modelo VGG16 – 50 épocas.

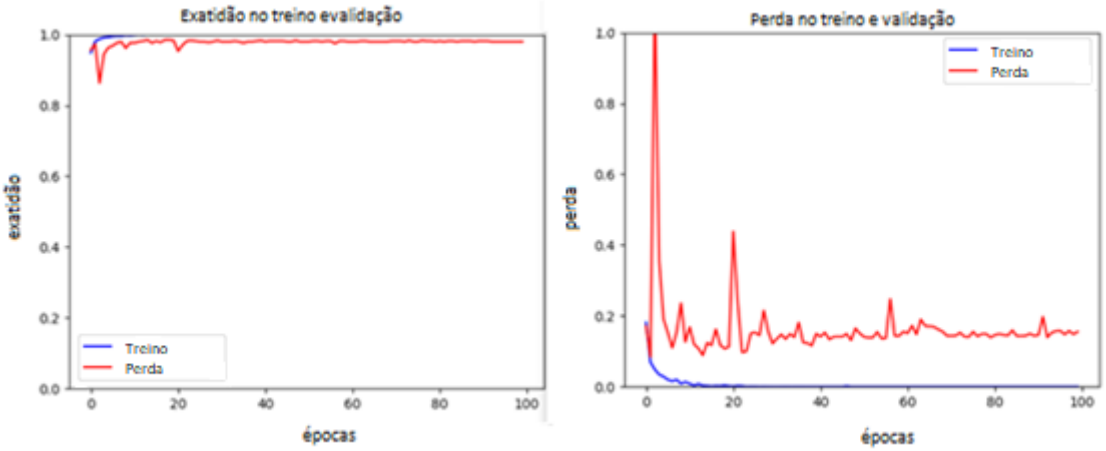

Figura 38 - Gráficos com evolução da exatidão (esquerda) e da perda (direita), utilizando a técnica de *transfer learning* aplicada ao modelo VGG16 – 100 épocas.

Na tentativa de obter resultados diferentes e poder extrair diferentes conclusões, foi alterado o número de nós da primeira camada *fully-connected*. Nas experiências anteriores esta camada contava com 512 nós e nas novas experiências foi alterado para 16, 64, 256 e 1024 nós. Os resultados gráficos representativos dos resultados obtidos podem ser observados nas Figuras Figura 39, Figura 40, Figura 41 e Figura 42 respetivamente.

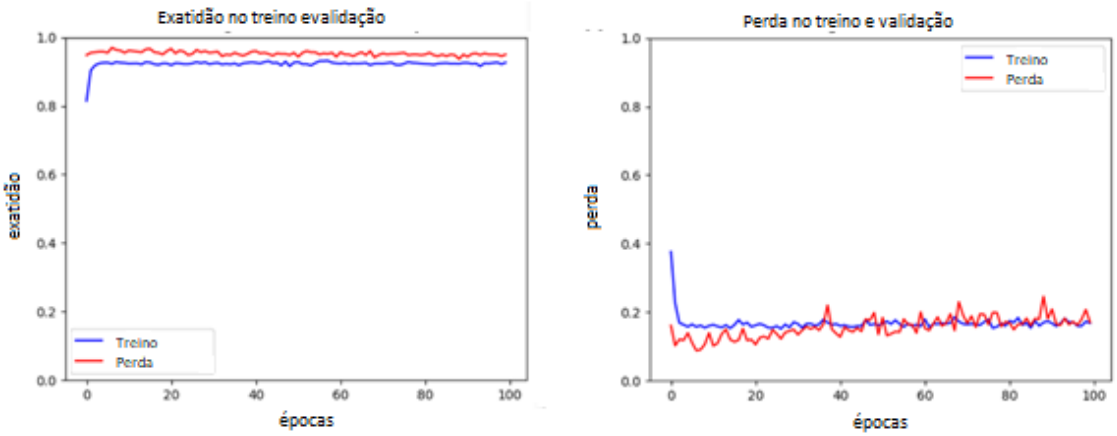

Figura 39 - Gráficos com evolução da exatidão (esquerda) e da perda (direita), utilizando a técnica de *transfer learning* aplicada ao modelo VGG16 – 100 épocas e 16 nós na 1ª camada *fully-connected*.

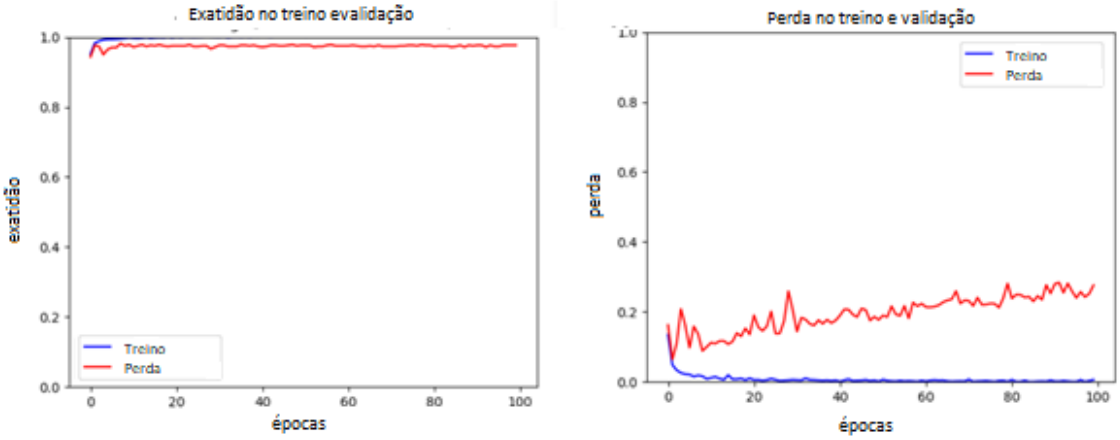

Figura 40 - Gráficos com evolução da exatidão (esquerda) e da perda (direita), utilizando a técnica de *transfer learning* aplicada ao modelo VGG16 – 100 épocas e 64 nós na 1ª camada *fully-connected*.

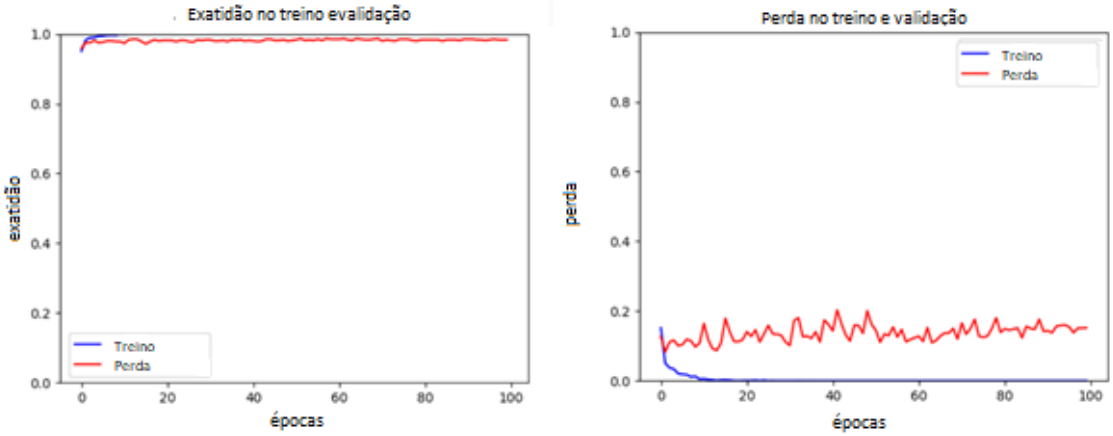

Figura 41 - Gráficos com evolução da exatidão (esquerda) e da perda (direita), utilizando a técnica de *transfer learning* aplicada ao modelo VGG16 – 100 épocas e 256 nós na 1ª camada *fully-connected*.

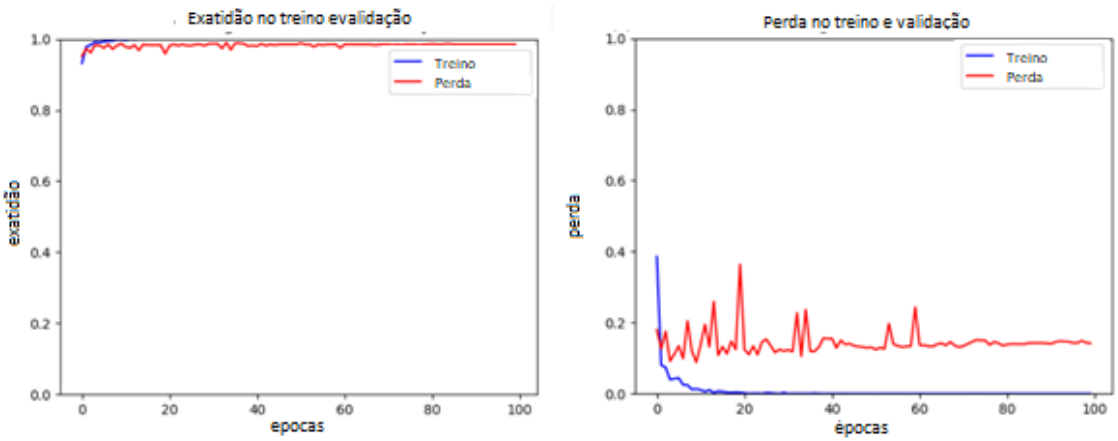

Figura 42 - Gráficos com evolução da exatidão (esquerda) e da perda (direita), utilizando a técnica de *transfer learning* aplicada ao modelo VGG16 – 100 épocas e 1024 nós na 1ª camada *fully-connected*.

4.3.2 – Análise de resultados

Das 582 imagens utilizadas para validação da rede para a experiência realizada com 100 épocas e 512 nós na primeira camada *fully-connected*, 13 foram classificadas incorretamente, das quais 1 corresponde a uma imagem de interior e as 12 restantes correspondem a imagens de exterior. Na Figura 43 são apresentados os resultados da validação de imagens após um treino de 100 épocas.

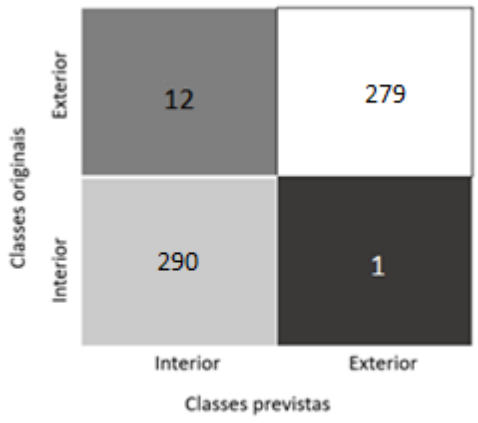

Figura 43 - Resultados obtidos na validação das imagens para a classificação de interior/exterior – 100 épocas.

Relativamente às imagens de interior mal classificadas, verifica-se que a característica mais evidente que partilham é a elevada intensidade de luz. Na Figura 44 apresentam-se

duas das imagens mal classificadas que exemplificam esta característica. A existência desta característica pode ter sido um fator preponderante na aprendizagem de classificação de imagens de exterior, já que as imagens de exterior têm maior probabilidade de ter maiores intensidades luminosas, devido ao facto de estarem mais e diretamente expostas à luz solar.

Por outro lado, verifica-se que algumas imagens de interior cujo campo visual se foque em portas ou janelas também foram mal classificadas. O facto da grande maioria das imagens de exterior utilizadas no treino da rede ser do imóvel, e como tal ter presente portas e janelas, pode ter tido influência nestas classificações erradas, bem como as tonalidades de castanho e verde que podemos observar na mesma Figura. Estas cores estão muito presentes nas imagens de exterior.

Figura 44 - Exemplo de três imagens de interior classificadas erradamente na fase de validação.

Quanto às imagens de exterior mal classificadas, existem muitas similaridades a um interior de um imóvel, sendo assim mais compreensível o motivo do erro na classificação. Na Figura 45 podem observar-se dois exemplos de imagens mal classificadas na fase de validação. Na imagem da esquerda observa-se um recanto de um pátio com paredes e janelas, o que torna mais facilmente confundível com uma divisão de interior de uma casa. Do lado direito da Figura 45 é apresentada uma imagem mal classificada de exterior, mas também nesta é justificável o erro, pois esta imagem apresenta, não só um teto como também mobília (sofá, mesa, etc..), habitualmente presentes numa sala (interior) de um imóvel.

Figura 45 - Exemplo de duas imagens de exterior de um imóvel classificadas como interior na fase de validação.

Analisando os resultados obtidos através dos gráficos da Figura 38, pode-se concluir que a exatidão e a perda, numa fase de treino, são inversamente proporcionais, tal como seria expectável uma vez que a exatidão aumenta com o aumento do número de épocas enquanto a perda é sucessivamente mais baixa. Nestes gráficos pode-se ver a azul os valores de treino e a vermelho os valores de validação.

A presença de *overfitting* é também percetível através da discrepância de valores das curvas de treino e validação, principalmente da função de perda. O valor de perda vai sempre aumentado, em vez de ir descendo à medida que as épocas passam.

No final das épocas o valor de perda na fase de treino é 0,00073734, enquanto o valor de perda na fase de validação é de 0,2144, atribuindo-se a grande oscilação dos valores da fase de validação ao reduzido conjunto de dados que utilizado para este efeito.

O facto de a perda na fase de treino estar substancialmente mais baixa que na fase de validação é mais um indício de que o modelo sofre de algum *overfitting*.

Relativamente aos dados de exatidão obteve-se 99,97% na fase de treino e 97,83% na fase de validação. Estes valores não podem ser vistos de forma absoluta, pois através dos valores obtidos na função de perda verifica-se que a rede está muito bem treinada para o conjunto de dados aplicado, mas não apresenta tão bons resultados para novos dados.

Nas experiências correndo 50 e 100 épocas os resultados não apresentaram melhorias consideráveis face à primeira experiência de 20 épocas. Através da Tabela 1 e da Figura 37 e Figura 38 pode-se observar que correndo o modelo em 50 épocas se obtém uma melhoria de 0,05% relativamente à experiência com apenas 20 épocas. Contudo, o valor de perda aumentou, passando de 0,116 para 0,1408. Na terceira experiência, foram

corridas 100 épocas, mas o desempenho desceu. O valor de exatidão alcançou os 97,83% e o valor de perda aumentou para 0,2144.

Tabela 1 - Valores de exatidão e de perda obtidos nas diferentes corridas.

Tamanho do *batch*	**Nrº. de épocas**	**Exatidão**	**Perda**
15	20	98,12%	0,116
15	50	98,17%	0,1408
15	100	97,83%	0,2144

As observações e conclusões alcançadas na experiência feita com 100 épocas aplicam-se também para as experiências com 20 e 50 épocas.

Da observação da Figura 46 percebe-se que das 582 imagens de validação, 11 foram mal classificadas, das quais 7 são de exterior e 4 de interior.

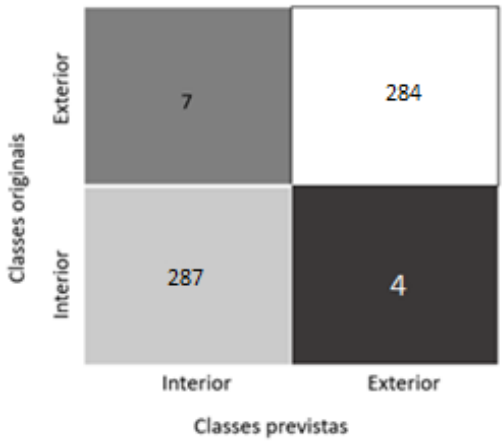

Figura 46 - Resultados obtidos na validação das imagens para a classificação de interior/exterior – 50 épocas.

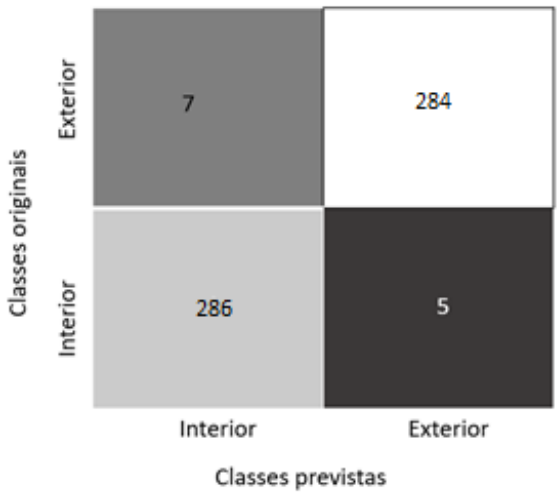

Figura 47 - Resultados obtidos na validação das imagens para a classificação de interior/exterior – 20 épocas.

Pela observação dos diferentes gráficos obtidos através da alteração do valor de nós de entrada da primeira camada *fully-connected* pode-se deduzir que, para os valores aplicados, quanto menor o número de nós, menos discrepância existe entre os valores na fase de treino e de validação. A Tabela 2 representa os valores obtidos através dos diferentes números de nós aplicados.

Tabela 2 - Valores de exatidão e de perda obtidos nas diferentes corridas.

Nrº de nós	**Exatidão**	**Perda**
16	95,00%	0,1701
64	97,67%	0,2764
256	98,33%	0,1520
1024	98,50%	0,1419

4.4 Classificação da qualidade representativa das imagens

4.4.1 – Aplicação do modelo

Para a classificação da qualidade representativa de imagens, foi também aplicado o método *transfer learning* à rede neuronal VGG16. O grupo de imagens utilizado para tal é apresentado na seção 4.2.

Um *batch* de tamanho igual a 20 significa que são utilizados 20 exemplos escolhidos aleatoriamente, neste caso imagens, a cada iteração.

Numa primeira abordagem optou-se por treinar por completo as últimas três camadas *fully-connected*, variando tanto o número de épocas como o tamanho do conjunto de imagens aplicado a cada iteração (*batch size*). Os resultados e as respetivas experiências são apresentados na Tabela 3.

Tabela 3 - Valores de exatidão e de perda obtidos nas diferentes corridas.

Nrº. de épocas	Tamanho do *batch*	Exatidão	Perda
10	15	93,00%	0,5238
10	20	91,18%	0,6636
20	15	94,25%	0,3614
20	20	94,00%	0,4054
50	15	96,44%	0,3917
100	15	96,44%	0,4473

Numa tentativa de melhorar e provar que a rede sofre de *overffiting*, foi corrida a rede com o aumento das épocas para 50 e 100 mantendo o *batch size* a 20. Os resultados desta experiência podem ser observados na Figura 48 e na Figura 49 respetivamente.

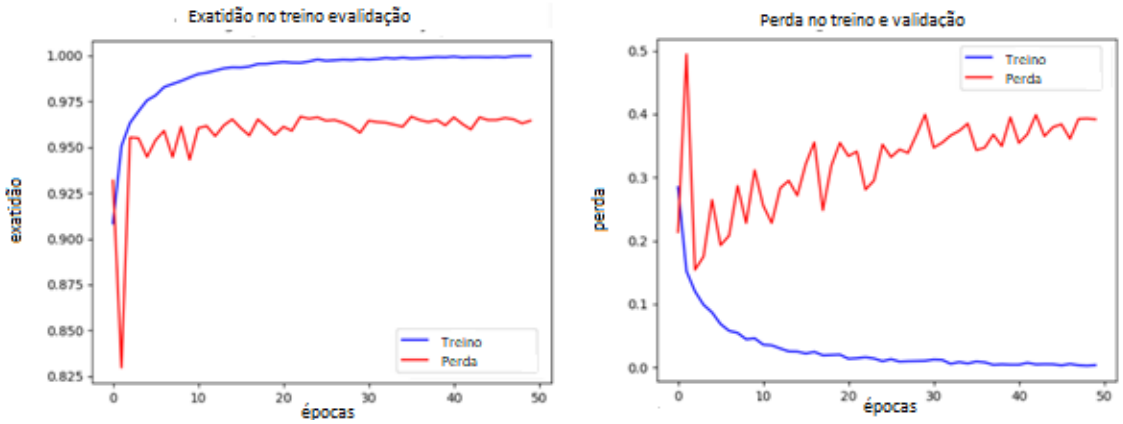

Figura 48 - Gráficos de exatidão (esquerda) e perda (direita) na fase de treino e validação do modelo - 50 épocas – *batch size*: 15.

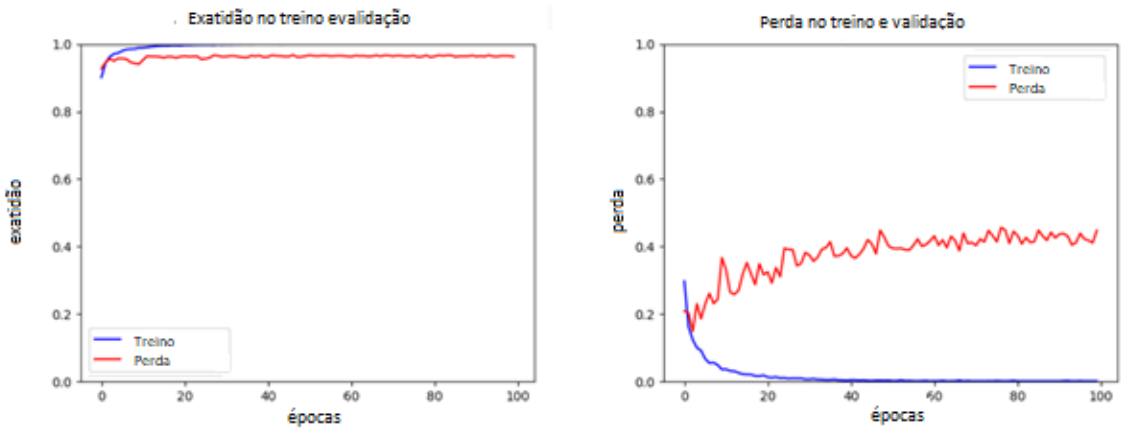

Figura 49 - Gráficos de exatidão (esquerda) e perda (direita) na fase de treino e validação do modelo - 100 épocas – *batch size*: 15.

Tal como na experiência abordada na secção 4.3, foram realizadas distintas parametrizações face ao número de nós na primeira camada *fully-connected*, adotando o mesmo conjunto de valores: 16, 64, 256 e 1024 nós. Os valores obtidos para o conjunto de valores escolhido podem ser observados nas Figuras Figura 50 - Gráficos com evolução da exatidão (esquerda) e da perda (direita), utilizando a técnica de *transfer learning* aplicada ao modelo VGG16 – 100 épocas e 16 nós na 1ª camada *fully-connected*. Figura 52 - Gráficos com evolução da exatidão (esquerda) e da perda (direita), utilizando a técnica de *transfer learning* aplicada ao modelo VGG16 – 100 épocas e 256 nós na 1ª camada *fully-connected*., 52 e 53.

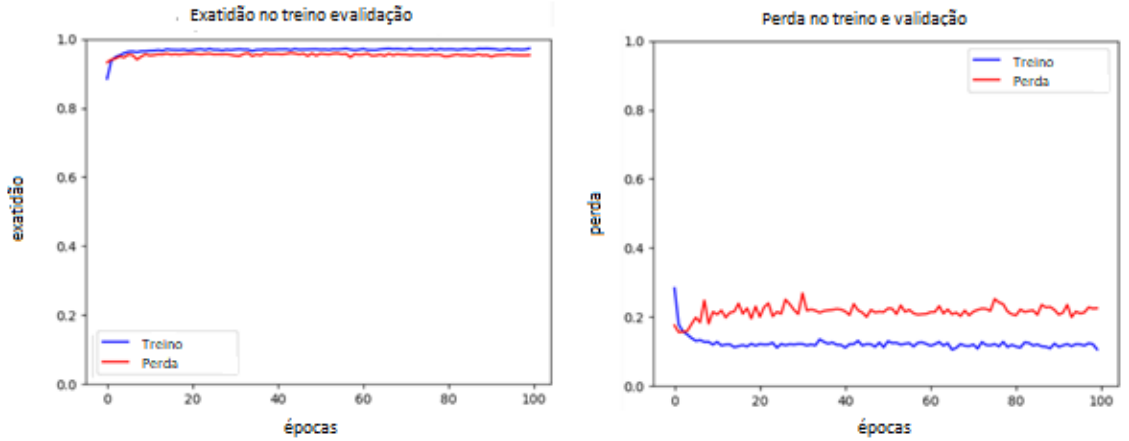

Figura 50 - Gráficos com evolução da exatidão (esquerda) e da perda (direita), utilizando a técnica de *transfer learning* aplicada ao modelo VGG16 – 100 épocas e 16 nós na 1ª camada *fully-connected*.

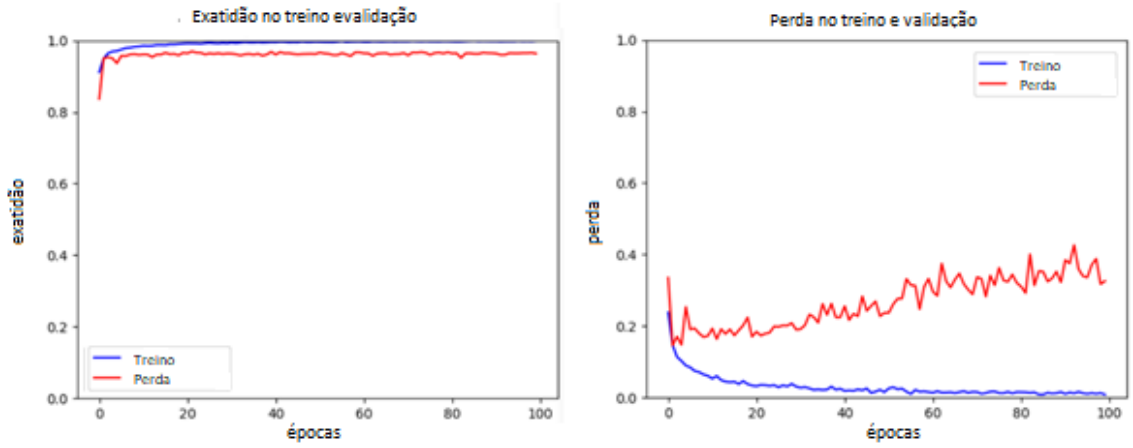

Figura 51 - Gráficos com evolução da exatidão (esquerda) e da perda (direita), utilizando a técnica de *transfer learning* aplicada ao modelo VGG16 – 100 épocas e 64 nós na 1ª camada *fully-connected*.

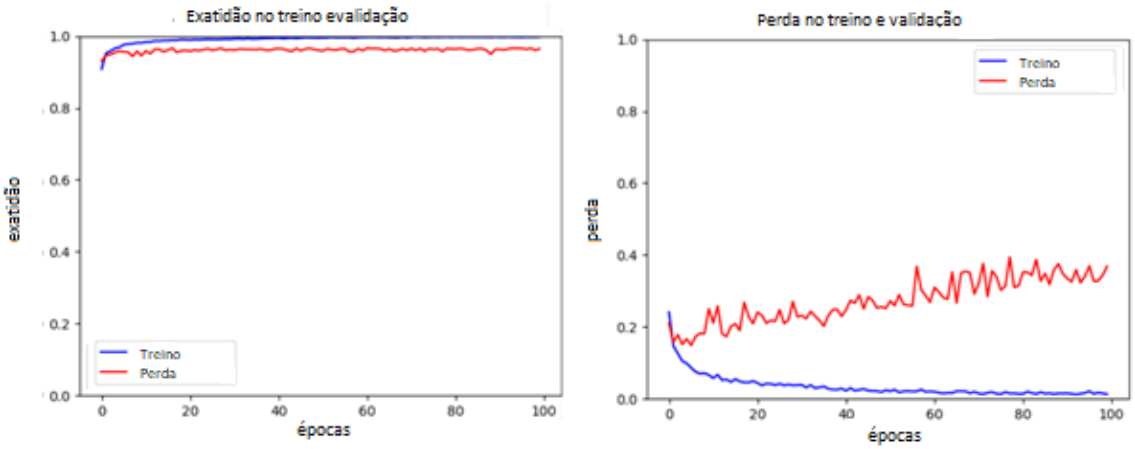

Figura 52 - Gráficos com evolução da exatidão (esquerda) e da perda (direita), utilizando a técnica de *transfer learning* aplicada ao modelo VGG16 – 100 épocas e 256 nós na 1ª camada *fully-connected*.

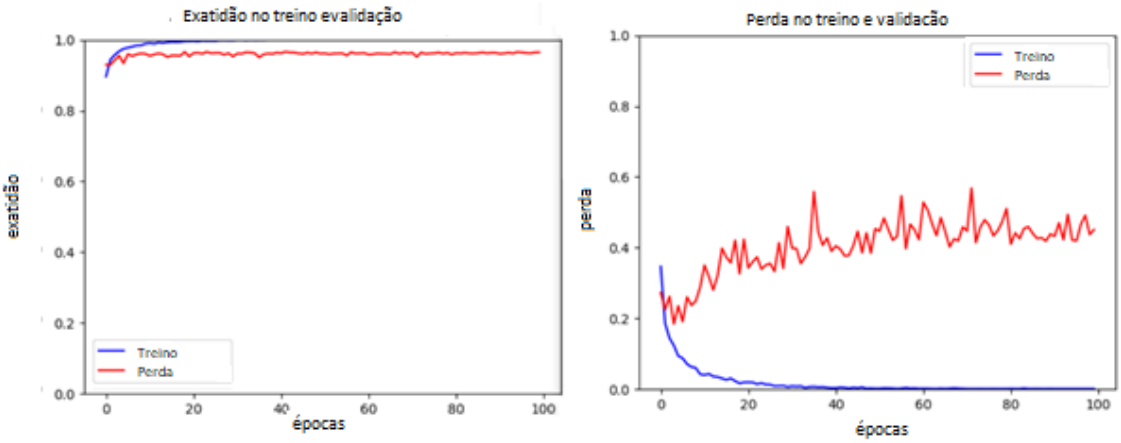

Figura 53 - Gráficos com evolução da exatidão (esquerda) e da perda (direita), utilizando a técnica de *transfer learning* aplicada ao modelo VGG16 – 100 épocas e 1024 nós na 1ª camada *fully-connected*.

4.4.2 – Análise de resultados

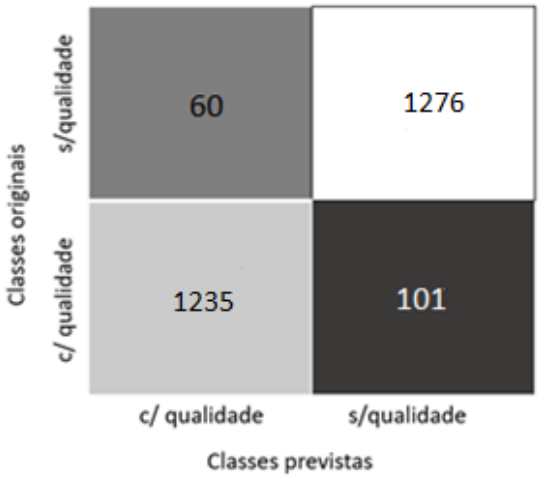

Figura 54 - Resultados obtidos na validação das imagens para a classificação de qualidade representativa -15 épocas.

Numa primeira experiência, a decisão de parar o modelo após 15 épocas de treino foi tomada uma vez que os valores de exatidão estavam a estabilizar. Foi feito um segundo treino à rede com 20 épocas, contudo mostrando resultados menos satisfatórios.

Das 2 672 imagens classificadas na fase de validação para um treino realizado com 15 épocas, 161 foram mal classificadas. Destas, 101 eram imagens com qualidade representativa e 60 eram imagens sem qualidade representativa, conforme se indica na Figura 54.

Se, para 15 épocas o modelo já apresentava indícios de *overfitting*, com 20 épocas tal foi confirmado. Das 2 672 imagens utilizadas para validação, 161 foram mal classificadas, mais 6 imagens do que na experiência anterior.

Na segunda experiência de treino do mesmo modelo com 20 épocas os resultados não melhoraram em comparação com o treino realizado com 15 épocas. A percentagem de exatidão na fase de treino aumentou para 99,55%, mas na fase de validação desceu para 91,18%. O mesmo sucedeu com os valores de perda, na fase de treino o valor de perda foi de 0,0131, contudo na fase de validação aumentou para 0,6636.

Os resultados apresentaram uma melhoria de 1,25% face ao melhor resultado obtido anteriormente, alcançando os 94,25% de exatidão, como se pode verificar na Tabela 3. O número de imagens mal classificadas desceu para 23.

Da mesma maneira, foi aumentado o número de épocas para 20, mantendo o tamanho do *batch* a 20. Os resultados obtidos foram bastante satisfatórios comparando com os resultados alcançados com o *batch* a 10. Contudo, não superou os 94,25%, mas alcançou uma exatidão de 94% com 24 imagens mal classificadas.

O valor de exatidão na fase de treino alcançou os 99,15%, mas na fase de validação apenas alcançou 93%. A mesma desarmonia é alcançada nos valores de perda, onde o valor alcançado na fase de treino é de 0,0261 e na fase de validação alcança o valor 0,5238.

Verifica-se também que as imagens mal classificadas se foram repetindo ao longo das experiências.

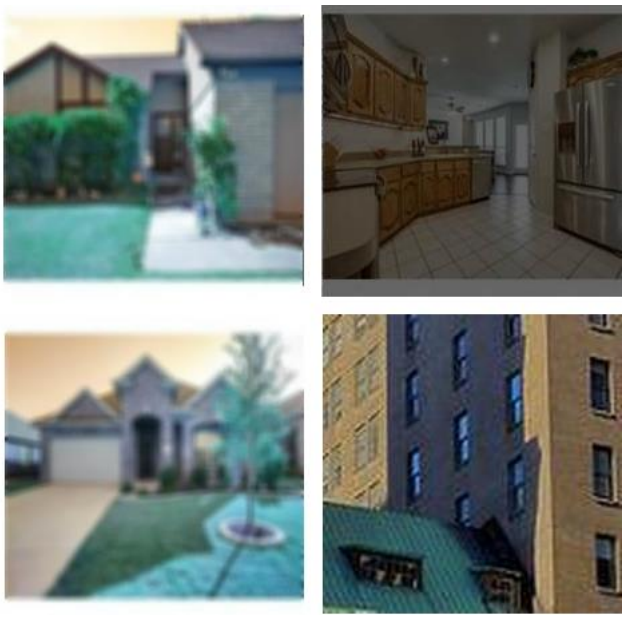

Figura 55 - Exemplo de imagens mal classificadas, consideradas como tendo qualidade representativa.

Na Figura 55 pode-se observar imagens sem qualidade representativa que foram mal classificadas na fase de validação e, na Figura 56, ilustra-se o cenário inverso, ou seja, imagens com qualidade representativa, mas classificadas como não tendo qualidade representativa.

Ao contrário do que sucedeu no conjunto de imagens apresentado na Figura 55, neste conjunto de imagens é possível observar um padrão, em que 80% são imagens de casas-de-banho. Todas estas imagens têm grande luminosidade e é possível ser este o fator que tenha influenciado a classificação errada.

Figura 56 - Exemplo de imagens mal classificadas, previstas como não tendo qualidade representativa.

Através do aumento do número das épocas corridas em cada experiência foi possível diminuir o número de imagens mal classificadas para 97, tanto com 50 épocas como com 100 épocas, tal como se pode observar na Figura 57 e na Figura 58.

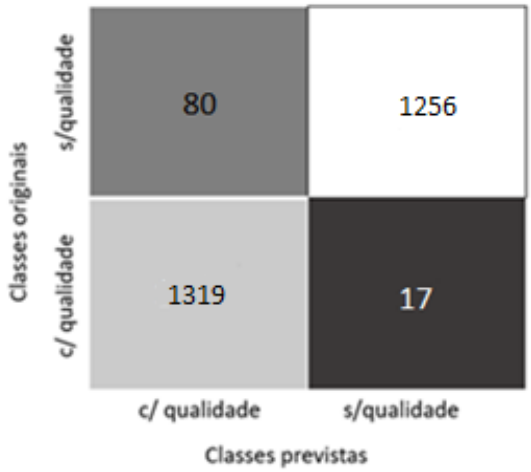

Figura 57 - Resultados obtidos na validação das imagens para a classificação de qualidade representativa – 50 épocas.

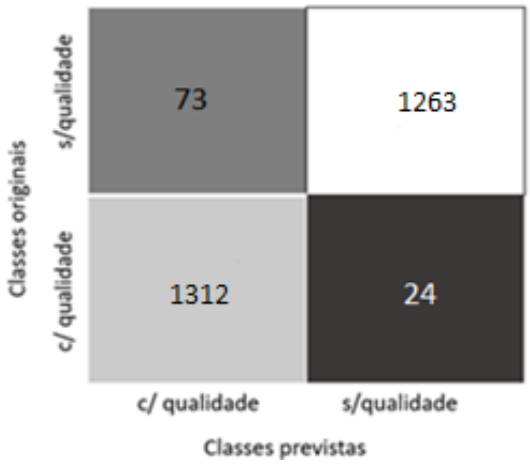

Figura 58 - Resultados obtidos na validação das imagens para a classificação de qualidade representativa – 100 épocas.

4.5 Análise crítica dos resultados

No que se refere à classificação as imagens de interior/exterior, pode concluir-se da evolução dos valores obtidos para a exatidão e para a perda ao longo das simulações, que a rede neuronal sofre de *overfitting*, principal razão pelo qual os resultados não são satisfatórios.

Por outro lado, é espectável que tanto a função de exatidão como a função de perda na fase de validação vão acompanhando os valores obtidos na fase de treino. Contudo, nas simulações efetuadas, os gráficos mostram que na fase de validação, os valores de exatidão não acompanham os valores da fase de treino; de igual modo, os valores de perda aumentam da fase de treino para a fase de validação. Por um lado, os valores de exatidão na fase de validação começam muito altos e os valores de perda muito baixos crescem com o decorrer das épocas. Pelos gráficos apresentados nas Secções 4.3.1 e 4.4.1 o treino poderia ter corrido apenas com um número muito baixo de épocas dado que não se alcançam melhores resultados à medida que o número de épocas aumenta.

Relativamente à classificação da qualidade representativa das imagens, o principal desequilíbrio entre os resultados apresentados no treino e os resultados apresentados na fase de validação deve-se também ao facto do modelo ter indícios de *overfitting*, apresentando bons resultados nos dados de treino, mas não acompanhar este resultado nos dados de validação, tal como na classificação de imagens de interior e exterior.

Verificou-se também que, quer o aumento do número de épocas quer o aumento do *batch size*, pouca ou nenhuma influência tiveram no combate ao *overfitting* e, consequentemente, no sucesso da utilização desta rede neuronal para a problemática abordada.

O *overfit* pode ser justificado pelo número de parâmetros / nós das camadas finais que são necessários estimar durante o treino em comparação com o número total de imagens utilizado para o treino. O número de parâmetros necessários treinar nas camadas *fully-connected* desta rede é 123 642 856 contudo o número total de imagens em ambas as experiências (interior / exterior e qualidade representativa) é bastante inferior, o que explica este fenómeno. Seria necessário um tamanho de *dataset* maior para que a rede pudesse generalizar melhor, tendo como base mais exemplos durante a fase de treino. Através de *data augmentation*, neste caso, *flips* horizontais, *shifts* e *crops*, poderia ser aumentado o tamanho do *dataset*.

Uma abordagem que poderia resultar, seria o treino da rede neuronal a partir de uma camada mais profunda, de modo a conseguir extrair características mais específicas presentes em cada classe.

Para tal, fizeram-se algumas tentativas de correr o mesmo software com ativação do modo que utiliza a GPU, o que exige maior poder computacional. No entanto, a instalação desse modo não foi possível, dadas as incompatibilidades existentes entre diferentes versões de software, impedindo a realização da referida experiência.

O facto de a taxa de acertos começar com um valor elevado na primeira época, poderá estar relacionado com o facto de se estar a usar uma rede pré-treinada para classificar imagens (embora para outras classes). Seria conveniente utilizar uma rede que não estivesse treinada para confirmar está hipótese.

Capítulo 5

5.1 Conclusões

O presente trabalho teve como objetivo apresentar uma arquitetura, baseada num modelo já existente, VGG16, capaz de classificar imagens de interior e exterior de imóveis e, do mesmo modo, capaz de classificar a qualidade representativa destas imagens.

Através da revisão de literatura realizada até à data foi possível concluir que a melhor abordagem para a classificação de imagens é feita através de redes neuronais convolucionais (cap. 3). Apesar das redes neuronais apresentarem a melhor taxa de sucesso na classificação de imagens, existem algumas contrapartidas para a obtenção destes resultados, nomeadamente a necessidade de usar ferramentas com grande capacidade computacional.

Uma vez feita a pesquisa e análise dos desenvolvimentos feitos até à data, foi aplicada a técnica de *transfer learning*, uma solução usada para classificação de imagens em diferentes áreas, que permite alcançar um desempenho bastante favorável sem necessidade de grande esforço computacional e com um conjunto de dados mais reduzido.

Sabendo que o *dataset* é dos fatores mais decisivos para um bom desempenho de uma rede neuronal, foi produzido um *dataset* RQI composto por um conjunto de imagens já existente, REI, adaptado ao contexto deste trabalho e incorporando um conjunto de imagens produzido manualmente.

O método de *transfer learning*, pré-treinado pelo conjunto de imagens *ImageNet*, VGG16 foi aplicado inicialmente para a classificação binária de imagens de interior e exterior de imóveis.

Até à data, os trabalhos realizados para esta temática de classificação de imagens de imóveis não são muitos. [5] apresenta pela primeira vez em 2017 a classificação de imagens das diferentes divisões de um imóvel.

Numa segunda fase foi feita uma análise aos aspetos que definem a qualidade e a qualidade representativa de uma imagem. A análise realizada permitiu concluir que a qualidade de uma imagem está ligada à sua qualidade representativa (Secção 3.3). Partindo deste pressuposto, foi produzido um segundo *dataset* para a finalidade de

classificar imagens quanto à sua qualidade representativa. Optou-se por realizar também uma classificação binária, ou seja, se tem qualidade representativa ou não.

Os resultados obtidos são satisfatórios tendo em conta o tamanho dos *datasets* utilizados e o baixo poder computacional envolvido. A utilização de uma rede neuronal pré-treinada foi fundamental e é a solução mais adequada para este tipo de problemas.

As soluções apresentadas atingem um desempenho satisfatório para as problemáticas em questão, contudo com algum problema de *overfitting.*

Algumas melhorias futuras serão descritas na próxima secção.

5.2 Trabalho Futuro

Conforme foi referido anteriormente, os resultados obtidos nas experiências efetuadas têm margem de desenvolvimento e melhoria, designadamente no que se refere à melhoria no desempenho.

Nesse sentido, a construção de um *dataset* maior seria a principal tarefa a desenvolver, uma vez que um *dataset* robusto em termos de dimensão tem como vantagem poder contornar os problemas de *overfitting*. De uma perspetiva diferente, um *dataset* maior podia permitir o treino de uma CNN de raiz.

Outra linha de desenvolvimento futuro poderia passar pela introdução de outras CNNs previamente treinadas e pela uma comparação direta dos resultados para a mesma experiência, muito embora seja expectável que os resultados obtidos sejam idênticos.

Outra possível linha de desenvolvimento poderá resultar da junção das duas experiências realizadas no âmbito deste trabalho, ou seja, a classificação direta de imagens, com qualidade representativa ou com falta desta qualidade e a classificação de imagens de interior e de exterior de imóveis.

A aplicação das redes neuronais treinadas no âmbito deste trabalho a casos práticos, tais como plataformas de promoção de imóveis poderá permitir identificar novas orientações de desenvolvimento. Numa perspetiva de futuras aplicações práticas, esta CNN poderia ser aplicada em áreas como a investigação criminal, como por exemplo a identificação do cenário.

6 Referências

[1] Z. Tong, D. Shi, B. Yan and J. Wei, “A Review of Indoor-Outdoor Scene Classificatio,” vol. 134, no. Caai, pp. 469–474, 2017.

[2] J. L. J. Luo and A. Savakis, “Indoor vs outdoor classification of consumer photographs using\nlow-level and semantic features” Int. Conf. Image Process, vol. 2, pp. 745–748, 2001.

[3] M. Szummer and R. W. Picard, “Indoor-Outdoor Image Classication”, IEEE International Workshop on Content-Based Access of Image and Video Database, 1998.

[4] N. Serrano, A. Savakis and J. Luo, “A computationally efficient approach to indoor/outdoor scene classification” Proc. - Int. Conf. Pattern Recognit., vol. 16, no. 4, pp. 146–149, 2002.

[5] J. H. Bappy, J. R. Barr, N. Srinivasan and A. K. Roy-Chowdhury, “Real estate image classification”, IEEE Winter Conf. Appl. Comput. Vision, no. March, pp. 373–381, 2017.

[6] L. Deng, “The MNIST Database of Handwritten Digit Images for Machine Learning Research” no. November, pp. 141–142, 2012.

[7] C. C. J. Kuo, “Understanding convolutional neural networks with a mathematical model” J. Vis. Commun. Image Represent., vol. 41, pp. 406–413, 2016.

[8] Q. Guo, Z. Liang and J. Hu, “Vehicle Classification with Convolutional Neural Network on Motion Blurred Images” DEStech Trans. Comput. Sci. Eng., no. aiea, pp. 40–45, 2017.

[9] J. W. Tan, S. W. Chang, S. Binti Abdul Kareem, H. J. Yap and K. T. Yong, “Deep Learning for Plant Species Classification using Leaf Vein Morphometric” IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 5963, no. c, 2018.

[10] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database” in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[11] G. H. A Krizhevsky, I Sutskever, “ImageNet Classification with Deep Convolutional Neural Networks” J. Geotech. Geoenvironmental Eng., vol. 12, p. 04015009, 2012.

[12] I. Bilbao and J. Bilbao, “Overfitting problem and the over-training in the era of data: Particularly for Artificial Neural Networks” IEEE 8th Int. Conf. Intell. Comput. Inf. Syst., pp. 173–177, 2018.

[13] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors” pp. 1–18, 2012.

[14] K. Simonyan and Z. Andrew, “Very Deep Convolutional Networks for Large-Scale Image Recognition” Am. J. Heal. Pharm., 2015.

[15] “The Vanishing Gradient Problem – Towards Data Science.” [Online]. Available:

https://towardsdatascience.com/the-vanishing-gradient-problem-69bf08b15484. [Accessed: 16-Jun-2019].

[16] H. Kaiming, Z. Xiangyu, R. Shaoqing and S. Jian, “Deep Residual Learning for Image Recognition” vol. 19, no. 2, p. 12, 2015.

[17] J. Luo, A. E. Savakis, S. P. Etz and A. Singhal, “On the application of Bayes networks to semantic understanding of consumer photographs”, International Conference on Image Processing 3:512 - 515 vol.3, 2000.

[18] “Support Vector Machines(SVM) — An Overview - Towards Data Science.” [Online]. Available: https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989. [Accessed: 15-Sep-2019].

[19] W. Kim, J. Park and C. Kim, “A novel method for efficient indoor-outdoor image classification” J. Signal Process. Syst., vol. 61, no. 3, pp. 251–258, 2010.

[20] M. Boutell and J. Luo, “Beyond pixels: Exploiting camera metadata for photo classification” Pattern Recognit., vol. 38, no. 6, pp. 935–946, 2005.

[21] W. Tahir, A. Majeed and T. Rehman, “Indoor/outdoor image classification using GIST image features and neural network classifiers”12th Int. Conf. High-Capacity Opt. Networks Enabling/Emerging Technol. HONET-ICT, 2015.

[22] A. Oliva, “Gist of the scene”, in Laurent Itti, Geraint Rees & John K. Tsotsos (eds.), Neurobiology of Attention. Academic Press. pp. 696-64, 2005.

[23] A. W. Setiawan, T. R. Mengko, O. S. Santoso and A. B. Suksmono, “Color retinal image enhancement using CLAHE” Proc. - Int. Conf. ICT Smart Soc. 2013 "Think Ecosyst. Act Converg, no. March 2015, pp. 215–217, 2013.

[24] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva and A. Torralba, “SUN database: Large-scale scene recognition from abbey to zoo” IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3485–3492, 2010.

[25] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, “Gradient-based learning applied to document recognitionre” IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1998.

[26] O. Russakovsky et al., “ImageNet Large Scale Visual Recognition Challenge” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Dec. 2015.

[27] C. Szegedy et al., “Going deeper with convolutions” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 17 Sep 201, pp. 1–9, 2014.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 2818–2826, 2016.

[29] C. Szegedy, S. Ioffe, V. Vanhoucke and A. Alemi, “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning”,Conference on Artificial Intelligence, 2016.

[30] X. Wang, J. Jia, J. Yin and L. Cai, “Interpretable Aesthetic Features for Affective Image Classification” IEEE Int. Conf. Image Process., pp. 3230–3234, 2013.

[31] Y. Deng, C. C. Loy and X. Tang, “Image Aesthetic Assessment: An experimental

survey," IEEE Signal Process. Mag., vol. 34, no. July, pp. 80–106, 2017.

[32] R. Datta and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach" ECCV, 2006.

[33] X. Lu, Z. Lin, H. Jin, J. Yang and J. Z. Wang, "RAPID: Rating Pictorial Aesthetics using Deep Learning" Proc. ACM Int. Conf. Multimed, pp. 457–466, 2014.

[34] L. Marchesotti, F. Perronnin, D. Larlus and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors" Proc. IEEE Int. Conf. Comput. Vis., pp. 1784–1791, 2011.

[35] N. Tawara, T. Ogawa, S. Watanabe, A. Nakamura and T. Kobayashi, "Visual categorization with bags of keypoints," APSIPA Trans. Signal Inf. Process., vol. 4, 2015.

[36] T. S. Jaakkola and D. Haussler, "Exploiting Generative Models in Discriminative Classifiers" NIPS, pp. 487–493, 1999.

[37] N. Murray, L. Marchesotti and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis" Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 2408–2415, 2012.

[38] S. Bianco, L. Celona, P. N. B and R. Schettini, "Predicting Image Aesthetics with Deep Learning" Advanced Concepts for Intelligent Vision Systems: 17th International Conference, 2016.

[39] Y. Jia et al., "Caffe: Convolutional Architecture for Fast Feature Embedding" In ACM Multimedia, 2, 675-678, 2014.

[40] Y. Wang, "Finetuning Convolutional Neural Networks for Visual Aesthetics", International Conference on Pattern Recognition, pp. 3554-3559, 2016.

[41] W. Luo, X. Wang and X. Tang, "Content-Based Photo Quality Assessment" IEEE Trans. Multimed., vol. 15, no. 8, pp. 1930–1943, 2013.

[42] X. Tian, Y. Long and H. Lv, "Relative Aesthetic Quality Ranking", IEEE Int. Conf. Syst. Man, Cybern. SMC 2018, pp. 2509–2516, 2019.

[43] T. Pennsylvania, "Algorithmic Inferencing of Aesthetics and Emotion in Natural Images: An Exposition", International Conference on Image Processing, pp. 8–11, 2008.

[44] R. C. Streijl, S. Winkler and D. S. Hands, "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives" Multimed. Syst., vol. 22, no. 2, pp. 213–227, 2016.

[45] Dumic, E., Sakic, K. & da Silva Cruz, L.A. "Crowdsourced subjective 3D video quality assessment. Multimedia Systems" 25, 673–694,2019.

[46] O. Wu, W. Hu and J. Gao, "Learning to Predict the Perceived Visual Quality of Photos" pp. 225–232, 2011.

[47] L. Kang, P. Ye, Y. Li and D. Doermann, "Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks" Proc. - Int. Conf. Image Process. ICIP, pp. 2791–2795, 2015.

[48] M. Buczkowski and R. Stasinski, “Convolutional Neural Network-Based Image Distortion Classification” pp. 275–279, 2019.

[49] C. Fellbaum, “WordNet: An Electronic Lexical Database” Bradford Books, 1998.

[50] “Machine Learning Glossary | Google Developers.” [Online]. Available: https://developers.google.com/machine-learning/glossary/positive_class. [Accessed: 29-Jun-2019].

[51] O. Ghorbanzadeh, T. Blaschke, K. Gholamnia, S. R. Meena, D. Tiede and J. Aryal, “Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection” Remote Sens., vol. 11, no. 2, 2019.

[52] “Supervised Machine Learning: Classification – Towards Data Science.” [Online]. Available: https://towardsdatascience.com/supervised-machine-learning-classification-5e685fe18a6d. [Accessed: 29-Jun-2019].