Avaliação da qualidade de imagens utilizando redes convolucionais
Projeto Final – Computação Natural – 2018/2

Alexandre Maros[1]

[1]Universidade Federal de Minas Gerais
Departamento de Ciência da Computação
Belo Horizonte – Minas Gerais – Brasil

alexandremaros@dcc.ufmg.br

Resumo. *Este artigo tem como propósito explorar e propor um método de avaliação de qualidade de imagens utilizando redes convolucionais e deep learning. Inicialmente é discutido o problema em se avaliar a qualidade de imagens (subjetividade, ruído), seguido da discussão do estado da arte atual, descrição do dataset utilizado, proposta do modelo e métricas utilizadas no treinamento e por fim a experimentação da proposta. Os resultados obtidos foram satisfatórios e dentro daquilo que a literura propõem. Há inúmeros outras abordagens que podem ser exploradas e as mesmas são dispostas nas conclusões.*

1. Introdução

O problema de se avaliar a estética, beleza ou qualidade de uma imagem, música, ou qualquer obra artística, vem sendo estudado por psicólogos e entusiastas da área desde muito tempo [Frost 1987, Chandrasekhar 1988]. É fácil de identificar que essa área de pesquisa há inúmeras interseções: na psicologia, têm-se o interesse de estudar e entender mais profundamente o que nós caracterizamos como "bonito"; jã na área de marketing o interesse vem da possibilidade de explorar essas características para tornar produtos mais atraentes aos consumidores; a neurociência visa o entendimento mais profundo do nosso cérebro.

Um dos problemas envolvendo pesquisas nessa área é a de que beleza é tido como algo pessoal, o que pode ser bonito para uma pessoa não necessariamente é bonito para outra, sem contar fatores externos que afetam nessa percepção: uma pintura feita por homens das cavernas, embora talvez simplista, é vista com diferentes olhos caso uma pintura semelhante seja feito na atualidade, devido aos fatores históricos relacionados a essa obra.

Embora este seja um problema complexo de se modelar (e ruidoso), há algumas características gerais dessas obras que aumentam a chance da mesma ser considerada como "bonita". Por exemplo, na fotografia, características que são levadas em conta na hora da avaliação de uma imagem são: exposição da imagem, regra dos terços, contraste, entre outros. Alguns autores, como [Birkhoff 1933], caracterizam que a beleza de uma obra vem de uma relação entre a harmonia, beleza e complexidade da mesma.

O problema que aqui será estudado é o da extração automática de característica de imagens através de modelos de redes neurais que podem ser úteis ao tentar prever se uma imagem é bonita ou não. Espera-se ser capaz de avaliar a qualidade de imagens

em um *dataset* coletado da *web* a partir de redes convolucionais, tendo como a entrada a imagem, de forma não estruturada, e extraindo um valor (nota) referente a qualidade de uma imagem. A novidade do trabalho vem de duas características: (i) avaliação das imagens através de uma nota entre 1 a 7 sobre o *dataset Photo.net* e (ii) estudo e impacto de duas funções de perda distintas *binary cross-entropy* e *earth mover*.

A Seção 2 fala sobre os trabalhos relacionados referente a este tema, na Seção 3 é disposto o *dataset* que será utilizado para a avaliação dos modelos propostos, na Seção 4 é discutido o modelo proposto, na Seção 5 é feito os experimentos dos modelos com base no *dataset* e é discutido seus resultados, e por fim, na Seção 6 é discutido a conclusão do trabalho.

2. Trabalhos Relacionados

[Birkhoff 1933] introduziu algumas das características, que define a beleza de uma obra como uma relação entra a ordem e a complexidade. Tais métricas por si só são difíceis de serem extraídas, mas [Rigau et al. 2007, Rigau et al. 2008] quantificou essas métricas utilizando Teoria da Informação. A medida de qualidade (ou *aesthetics*) de Birkhoff pode então ser definida com a relação entre a redução algorítmica da incerteza (ordem) e a incerteza inicial (complexidade). De acordo com o Birkhoff, uma boa experiência de *aesthetic* é baseado em três fases:

1. Um efeito preliminar da quantidade de esforço da atenção, que aumenta proporcionalmente conforme a complexidade (C) de uma obra
2. Um valor da qualidade estética (M)
3. Uma verificação de que a obra tenha uma certa harmonia, simetria ou ordem (O)

Em qualquer processo artístico, tem-se um repertório de elementos, como as cores disponíveis, que é transmitido ao produto final, que pode ser visto como um canal de comunicação entre o artista e sua obra final. Tendo essa ideia em mente, a quantificação da complexidade e ordem de uma obra é dada através da entropia de Shannon e da complexidade de Kolmogorv dessas obras.

[Marchesotti et al. 2011] estudou o uso de *features* extraídas junto com o treinamento de modelos de aprendizado de máquina, como o SVM. Inicialmente, os autores consideram a extração de *features* propostas por [Datta et al. 2006, Ke et al. 2006] como brilho, contraste, exposição, saturação, regra dos terços, tamanho e proporção da imagem, entre outros. No total foram 64 *features* distintas que são previamente calculadas e estruturadas. Além disso, também foram consideradas *features* de "alto nível", como o algoritmo de *Bag of Visual Words* (BOVW) [Csurka et al. 2004], *Fisher Vector* (FV) [Perronnin and Dance 2007] e GIST [Oliva and Torralba 2001]. Os algoritmos BOVW e FV funcionam através do agrupamento (utilizando K-Means ou Gradient Boosting) de vetores SIFT, que por sua vez capturam informações como "há bordas nessa região?" ou "essa parte da imagem está saturada?". GIST utiliza o histograma para capturar informações referentes a partições da imagem. O *dataset* utilizado na classificação é o *Photo.net* o mesmo aqui utilizado.

[Wang et al. 2016] apresenta o *multi-scene deep learning model* (MSDLM) para automaticamente aprender as *features* necessárias para a classificação da qualidade de imagens. Os autores afirmam que o modelo produz melhores resultados que aquelas com

features pré-definidas e também é capaz de automaticamente contornar problemas como *datasets* desbalanceados, ruído, entre outros. O modelo é proposto através de uma rede convolucional, onde há quatro camadas convolucionais, seguidas por uma camada descrita como *camada das cenas*, composta por 7 grupos separados designados para construir as cenas. Esses grupos são então ligados a duas camadas completamente conectadas de 512 neurônios. O objetivo da rede é uma classificação binária (qualidade baixa ou alta). O modelo atingiu uma acurácia de 84.88%.

Por fim, [Lu et al. 2015] também desenvolveu uma rede convolucional para avaliar a qualidade de uma foto, extraindo as *features* automaticamente. Os autores desenvolveram uma rede convolucional de dua camadas, possibilitando assim dois tipos de visões: uma global e uma local. Esse modelo tenta contornar uma limitação das redes atuais: a de que a entrada deve ser de um tamanho fixo. Esse problema é especialmente problemático neste caso já que o redimensionamento das imagens tem um grande impacto na qualidade das mesmas (o mesmo problema é visto na área de *data augmentation* dessas imagens). O uso de duas colunas, uma com a imagem completa e outra com um *crop* aleatório visa contornar esse problema. O modelo é testado no dataset AVA em 1.5 milhões de imagens. A classificação também é entre baixa e alta qualidade.

3. *Dataset*

O *dataset* utilizado neste trabalho foi retirado do *website* Photo.net[1], que é um serviço onde usuários enviam suas fotos e pessoas aleatórias que também utilizam o serviço avaliam as imagens com notas de 1 a 7. Há outros serviços como este que poderiam ser explorados mas que, por falta de tempo, foram deixados de lado, sendo estes: DPChallenge[2] e TerraGaleria[3].

Como o *website* possuí milhões de fotos para serem avaliadas, foi considerado um subconjunto dessas fotos para a avaliação. Este subconjunto foi o mesmo considerado por [Datta et al. 2006]. O primeiro passo para a utilização do *dataset* foi a criação de um *parser* de páginas HTML para salvar as imagens dado uma lista de IDs. Somente os IDs foram disponibilziados no site da autora[4] por questões de privacidade e direitos autorais, logo a necessidade do *parser* automático.

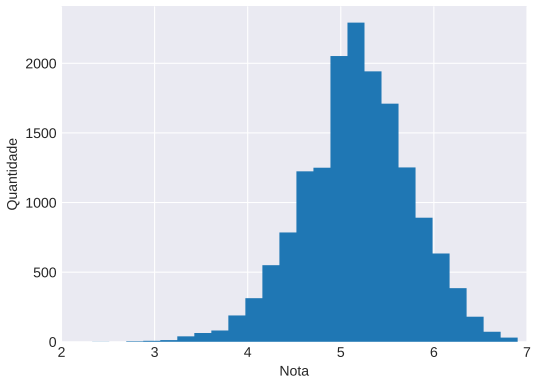
Um dos problemas enfrentados é a de que, como houve a necessidade de coletar as imagens diretamente do site, houve diversas instâncias que não estavam mais dispostas (por ter sido removida por administradores ou pelo usuário removendo-a). Das 20278 imagens dispostas no *dataset*, apenas 15936 foram obtidas (aproximadamente 79% do *dataset* original). Cada imagem acompanha a média das avaliações junto com a quantidade de avaliações de cada respectiva nota, que será utilizada como uma distribuição de probabilidade em uma das funções de perda.

A primeira análise feita sobre as imagens é a distribuição das notas que os usuários deram para as imagens. A Figura 1 mostra, respectivamente, a distribuição da média das avaliações no *dataset* completo, no subconjunto de treinamento e no subconjunto de testes. A distribuição dos conjuntos foi feita da seguinte maneira: 80% para o treinamento,

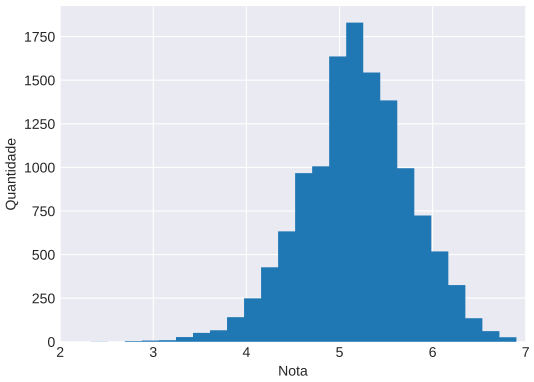
[1] `http://www.photo.net`
[2] `http://www.dpchallenge.com/`
[3] `http://www.terragalleria.com`
[4] `http://ritendra.weebly.com/aesthetics-datasets.html`

10% para a validação cruzada e 10% para o teste. É importante notar que a melhor abordagem para calcular o erro da rede seria utilizando a técnica de *K-Fold*, mas devido a limitação do poder computacional optou-se por fazer apenas um *split*. Pela distribuição das notas é possível identificar que as avaliações se concentram majoritariamente entre 4.5 e 5.5, com muito poucos exemplos a baixo de 4 e acima de 6.

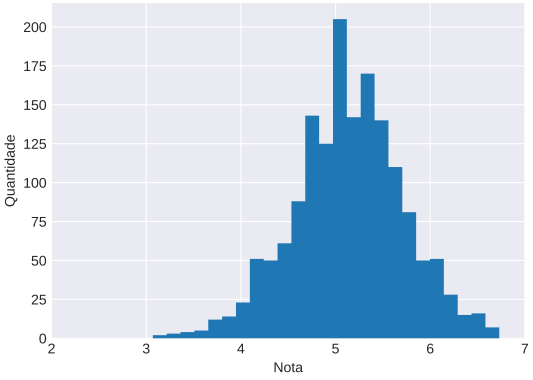
Em problemas envolvendo imagens, tipicamente é possível aplicar técnicas de *data augmentation*, onde é aplicado ruído, as imagens são rotacionadas, ou suas cores alteradas para gerar mais exemplos das classes minoritárias. Infelizmente essa técnica não pode ser utilizada aqui justamente pelo fato de essas modificações terem um impacto direto na qualidade da imagem. Por exemplo, um gato em preto e branco ou colorido ainda continua sendo um gato, entretanto, alterar as cores das imagens poderia resultar em uma avaliação completamente diferente. Para resolver o problema das classes com poucos exemplos, foi repetida essas imagens sem alteração para a rede vê-las mais vezes e também foi utilizado a técnica de atribuir pesos a cada exemplo de acordo com sua quantidade no *dataset*.



(a) Distribuição do *dataset* completo



(b) Distribuição do conjunto de treinamento



(c) Distribuição do conjunto de teste

Figura 1. Distribuição das avaliações no *dataset* completo e nos subconjuntos de treino e teste

A Figura 2 mostra alguns exemplos das imagens com as piores e melhores avaliações. As três primeiras imagens são referentes as fotos com as piores avaliações. As imagens 2(a) e 2(c) possivelmente tem avaliações ruins devido a estarem fora de foco,

enquanto que a imagem 2(b) possivelmente tem um baixa nota devido a utilização de *zoom* digital, ter um alto brilho e baixo contraste. As três últimas imagens foram as que tiveram as melhores avaliações e são notavelmente melhores que as três primeiras, tendo uma gama de cores maior, um bom contraste, brilho, foco, entre outros fatores.

(a) 2.18 (b) 2.88 (c) 2.91

(d) 6.81 (e) 6.88 (f) 6.90

Figura 2. Avaliação média das três piores e melhores imagens do *dataset* Photo.net

4. Modelo

Nesta seção é discutido os dois tópicos referentes ao modelo proposto: (i) a arquitetura que foi utilizado das redes convolucionais e; (ii) as duas funções de perda que foram utilizadas (*binary cross-entropy* e *earth mover*.

4.1. Arquitetura

A arquitetura foi baseada no artigo proposto por [Lu et al. 2015], embora seu tamanho tenha sido reduzido consideravelmente devido ao poder computacional disponível para este trabalho. A Figura 3 mostra detalhadamente a arquitetura proposta. A entrada da rede é uma imagem de 224×224 (com três canais, RGB) e possui 4 camadas de convolução e 2 camadas de *max pooling*. As três últimas camadas são completamente conectadas, a primeira com 256 neurônio, a segunda com 128 neurônios e a última, com 7 neurônios e aplicação da função *softmax* (referente a camada de saída, com 7 possíveis notas). No total a rede possui treze milhões de parâmetros (pesos) a serem treinados.

Além da rede aqui proposta, algumas redes já estabelecidas na literatura também foram testadas, sendo elas: MobileNet; MobileNetV2; VGG-16 e; VGG-19. As duas últimas redes se tornaram inviáveis a serem utilizadas devido a seu tamanho e profundidade. As duas versões da MobileNet demoraram consideravelmente mais para treinar e

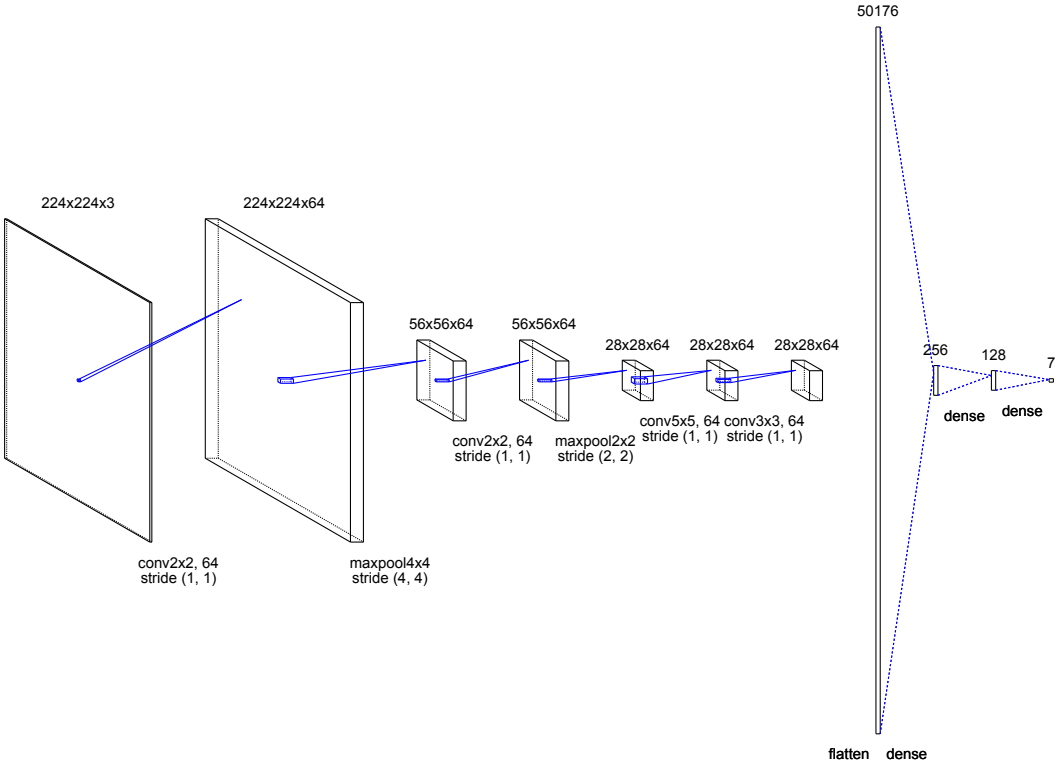


Figura 3. Arquitetura da rede proposta

não produziram melhores resultados que o modelo aqui proposto. Também foi testado essas redes com pesos pré-treinados no *dataset imagenet*, mas também não houve melhoras consideráveis.

Outro ponto que também foi considerado neste trabalho é o tamanho da imagem da entrada. Como discutido na introdução, a alteração da resolução da imagem afeta diretamente na qualidade e o principal problema dessa abordagem é a de ser necessário esse redimensionamento inicial. Uma das abordagens da arquitetura de [Lu et al. 2015] é a de usar duas colunas, onde uma é a imagem completa, redimensionada ao tamanho de entrada, e a segunda coluna é um *crop* aleatório da imagem. Essa ideia traz a questão de capturar melhor o foco, bordas e outros aspectos que são perdidos no redimensionamento. Neste trabalho, não foi considerado a versão de duas colunas devido ao problema de complexidade da rede (aumento da complexidade de treino da rede).

4.2. Funções de Perda

Duas funções de perda foram consideradas nos modelos. A primeira, denominada *cross-entropy*, disposta na Equação 1 (onde y_i e $\hat{y}_i$ são distribuição de probabilidade entre as t classes e N o número de exemplos) é a função mais utilizada quando se trata de previsão de classes e aqui serve como uma espécie de *baseline* para o modelo.

$$J = \frac{1}{N} - \sum_{i=1}^{N} y_i \times log(\hat{y}_i) \tag{1}$$

A segunda função de perda utilizada se chama *earth mover distance* (EMD). Esta função vem da teoria da informação e tenta modelar um problema de: dado dois conjuntos de montes de terra, qual é o menor esforço que deve-se fazer para igualar a distribuição desses dois conjuntos [Rubner and Tomasi 2001].

Na prática, ele tenta modelar um comportamento muito interessante para o problema de classificação de imagens. Por exemplo, se a nota de uma imagem é 7 e a rede previu que a nota é 6 a função de perda deveria ser muito menor do que se a rede atribuí-se uma nota 1 (a nota 1 está muito mais distante da realidade do que a nota 6). A função *cross-entropy* não captura esse comportamento e pode atribuir a mesma perda para a classificação da nota 1 assim como para a nota 6. Esta perda já foi utilizada em outras redes para a classificação de imagens, como a NIMA [Talebi and Milanfar 2018], e teve bons resultados. A Equação 2 descreve o cálculo da função.

$$EMD(y, \hat{y}) = \left(\frac{1}{N} \sum_{k=1}^{N} |CDF_y(k) - CDF_{\hat{y}}(k)| \right) \tag{2}$$

onde $CDF_p(k)$ é a função de distribuição acumulada (*cummulative distribution function*).

5. Análise Experimental

A análise experimental é feita em três etapas. A primeira etapa consiste em visualizar como as funções de perda se comportam no conjunto de treino e no conjunto de validação. A segunda etapa é referente a análise do erro obtido pelo modelo no conjunto de teste. E por fim, é feito uma análise mais detalhada das notas retornadas pelo modelo em imagens reais fornecidas pelo autor, comparando o impacto que diferentes filtros tem nas imagens.

Os modelos foram implementados utilizando a biblioteca Keras[5] no Python, junto com o TensorFlow[6] e biblioteca de utilidade como Pandas, Numpy, entre outros. O treinamento se deu através de 20 épocas, utilizando *mini-batch* de 12 imagens, o otimizador escolhido foi o Adam (descida do gradiente com momentum e regularização L2) com *learning rate* de 0.001. O tempo de treinamento foi em torno de 14 horas com uma placa de vídeo GTX 980M. O tempo de treinamento foi maior que o esperado devido ao tamanho do modelo, tamanho das imagens e a baixa memória da placa de vídeo utilizada.

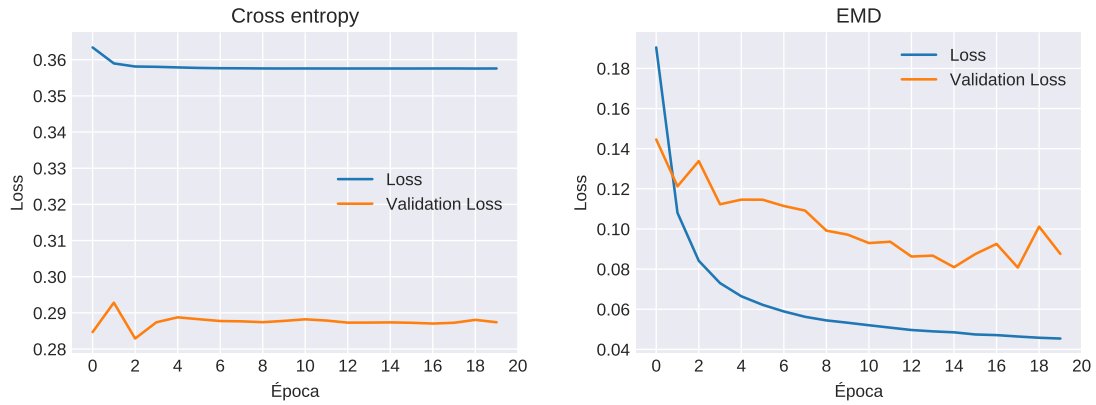
A Figura 4 mostra a curva das funções de perda de *cross entropy* e EMD. Em relação a função de *cross entropy* na Figura 4(a), é possível identificar que o erro de treinamento melhorou um pouco na primeira época e continuou praticamente estável durante todo o resto. O erro de validação teve o mesmo comportamento, tendo o melhor resultado na segunda época. Para o EMD, na Figura 4(b) é possível identificar um comportamento bem mais esperado, em que a cada época o erro decresce, tanto para a função de treino, que se aproxima de 0 nas últimas época, como para a função de validação. A partir da época 14 o *loss* de validação começou a aumentar, sugerindo que está havendo um *overfitting* dos dados.

Esse é um bom exemplo para destacar a importância da função de perda. Com a função de *cross entropy* (a mais utilizada) o modelo não parece aprender bem e tem uma

[5] `https://keras.io/`

[6] `https://www.tensorflow.org/`

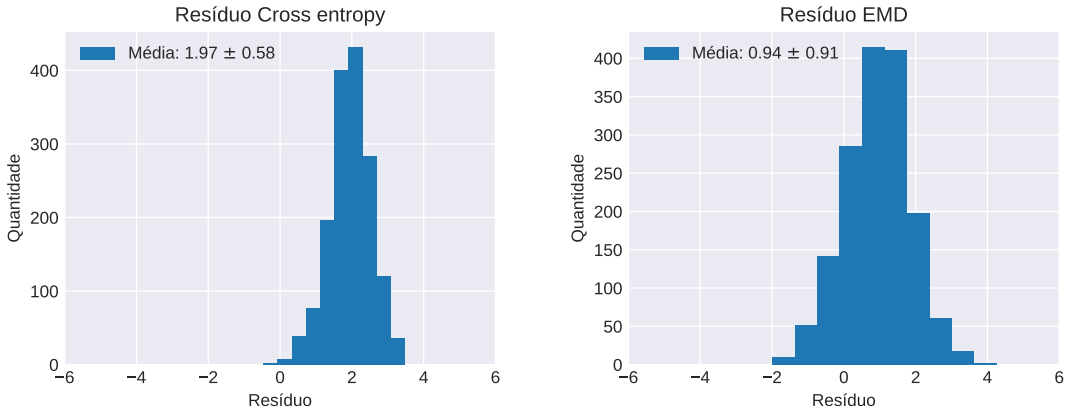
rápida convergência, enquanto que para o EMD, o comportamento esperado é atingido e parece se ajustar melhor aos dados.



(a) *Cross Entropy* (b) *Earth Mover Distance*

Figura 4. O comportamento da métrica *loss* utilizando *cross entropy* e *earth mover distance*

A Figura 5 mostra um histograma dos resíduos (o valor real, ou seja, a média da avaliação, menos a média da avaliação prevista para o modelo, com base na distribuição de probabilidade), junto com sua média e o desvio padrão. Para o modelo utilizando a *cross entropy* a média do erro de avaliação é de 1.97 pontos acima do esperado, já para o modelo utilizado o EMD, a média é de 0.94 pontos acima do esperado (um erro consideravelmente menor). Essa diferença de resíduo reflete o *loss* obtida na Figura 4, em que o EMD apresentava um comportamento mais semelhante ao esperado.



(a) *Cross Entropy* (b) *Earth Mover Distance*

Figura 5. Resíduo (real - previsao) no conjunto de teste utilizando *cross entropy* e *earth mover distance*

A Figura 6 mostra uma análise de duas figuras, uma considerada boa e outra ruim (classificação dada pelo autor). A Figura 6(a) mostra a foto de um gato, utilizando uma DSLR com uma lente 50mm, f/2.2, pós processada, obtendo assim um com um desfoque mais presente e com contraste e brilho controlados. Já a Figura 6(b) foi retirada em

(a) Imagem Boa – Nota: 4.57

(b) Imagem Ruim – Nota: 3.43

Figura 6. Fotografia considerada boa (uso de DSLR e pós processamento) *versus* imagem retirada com celular em movimento e baixa luminosidade

um Samsung Galaxy S6, com baixa luminosidade, movimentando o celular para atingir uma nitidez menor. A Figura considerada boa teve uma avaliação de 4.57, enquanto que a Figura considerada ruim obteve uma nota menor, de 3.43. Outros exemplos foram testados para verificar se a rede estava realmente retornando algo com que batesse com a percepção do autor, e foi validado essa consistência.

(a) Imagem Original – Nota: 4.57

(b) Imagem Borrada – Nota: 3.17

(c) Imagem com alta exposição – Nota: 3.74

Figura 7. Previsões das notas de uma foto normal *versus* diferentes tipos de filtros

Por fim, na Figura 7, é visto o efeito que filtros de imagem tem sobre as avaliações retornadas pela rede. A imagem original pode ser vista na Figura 7(a) e obteve uma nota de 4.57. A Figura 7(b) é referente a aplicação de um filtro de borramento agressivo (*gaussian blur* sobre a imagem, onde a a nota decaiu em 1.4 pontos. A Figura 7(c) mostra a aplicação de um filtro para aumentar a exposição da imagem. Neste caso, a nota foi de 4.57 para 3.74, decaindo apenas 0.87 pontos.

A aplicação desses filtros faz com que as imagens apresentem o comportamento esperado, ou seja, suas notas serem menores, dado que há uma claro decaimento na qualidade da mesma. O efeito de borramento teve um efeito maior do que o de exposição, o que faz sentido, dado que normalmente se preza por uma melhor nitidez e uma alta exposição não necessariamente nos dá uma má imagem.

6. Conclusão

Neste trabalho foi feito a avaliação da qualidade de imagens utilizando redes convolucionais, sendo o objetivo principal do trabalho o aprofundamento do conhecimento das técnicas de rede convolucional, a validação desta técnica para esse propósito e também verificar o efeito de duas funções de perdas distintas que capturam informações diferentes e melhor entender o comportamento de cada função.

As redes se mostraram efetivas na avaliação das imagens produzindo resultados dentro do esperado, mesmo com alguns problemas inerentes da classificação e avaliação de imagens (subjetividade, ruídos) assim como problemas do *dataset* selecionado (principalmente envolvendo o desbalanceamento do mesmo). As avaliações ficaram com uma média de resíduo de 0.94 para o modelo com a melhor função de perda para este caso (EMD).

Além disso, foram testados os efeitos que diferentes filtros de imagens tem sobre as avaliações resultantes, como o *gaussian blur* e ajustes extremos na exposição. Foi identificado um comportamento similar com o que se esperaria, ou seja, as avaliações da qualidade da imagem decaíram significativamente.

Como trabalhos futuros, sugere-se os seguintes tópicos:

- Teste dos erros utilizando *K-folds*
- Utilização de *datasets* mais extensos e menos desbalanceados, como o AVA[7]
- Avaliação com diferentes arquiteturas

Referências

[Bianco et al. 2016] Bianco, S., Celona, L., Napoletano, P., and Schettini, R. (2016). Predicting image aesthetics with deep learning. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 117–125. Springer.

[Birkhoff 1933] Birkhoff, G. D. (1933). *Aesthetic measure*, volume 38. Harvard University Press Cambridge.

[Chandrasekhar 1988] Chandrasekhar, S. (1988). Truth and beauty: Aesthetics and motivations in science. *Bibliovault OAI Repository, the University of Chicago Press*, 22.

[7] `https://research.google.com/ava/`

[Csurka et al. 2004] Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague.

[Datta et al. 2006] Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2006). Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*, pages 288–301. Springer.

[Deselaers et al. 2008] Deselaers, T., Pimenidis, L., and Ney, H. (2008). Bag-of-visual-words models for adult image classification and filtering. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE.

[Frost 1987] Frost, L. (1987). *What makes a painting good?* PhD thesis, Rhodes University Grahamstown.

[Kahneman and Egan 2011] Kahneman, D. and Egan, P. (2011). *Thinking, fast and slow*, volume 1. Farrar, Straus and Giroux New York.

[Ke et al. 2006] Ke, Y., Tang, X., and Jing, F. (2006). The design of high-level features for photo quality assessment. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 419–426. IEEE.

[Lu et al. 2015] Lu, X., Lin, Z., Jin, H., Yang, J., and Wang, J. Z. (2015). Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia*, 17(11):2021–2034.

[Marchesotti et al. 2011] Marchesotti, L., Perronnin, F., Larlus, D., and Csurka, G. (2011). Assessing the aesthetic quality of photographs using generic image descriptors. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1784–1791. IEEE.

[Oliva and Torralba 2001] Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175.

[Perronnin and Dance 2007] Perronnin, F. and Dance, C. (2007). Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE.

[Rigau et al. 2007] Rigau, J., Feixas, M., and Sbert, M. (2007). Conceptualizing birkhoff's aesthetic measure using shannon entropy and kolmogorov complexity. In *Computational Aesthetics*, pages 105–112.

[Rigau et al. 2008] Rigau, J., Feixas, M., and Sbert, M. (2008). Informational aesthetics measures. *IEEE Computer Graphics and Applications*, 28(2).

[Rubner and Tomasi 2001] Rubner, Y. and Tomasi, C. (2001). The earth mover's distance. In *Perceptual Metrics for Image Database Navigation*, pages 13–28. Springer.

[Talebi and Milanfar 2018] Talebi, H. and Milanfar, P. (2018). Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011.

[Wang et al. 2016] Wang, W., Zhao, M., Wang, L., Huang, J., Cai, C., and Xu, X. (2016). A multi-scene deep learning model for image aesthetic evaluation. *Signal Processing: Image Communication*, 47:511 – 518.