Classificação de Gênero de Vídeo usando Fusão de Redes Neurais Convolucionais

Victor Lúcio e Jurandy Almeida
Instituto de Ciência e Tecnologia, Universidade Federal de São Paulo – UNIFESP
12247-014, São José dos Campos, SP – Brazil
Email: {victor.lucio, jurandy.almeida}@unifesp.br

***Resumo*—A geração de dados de vídeo atualmente ocorre de maneira muito rápida e em grande escala, os dispositivos móveis com essa capacidade estão cada vez mais difundidos, gerando um grande conteúdo de dados, que precisam ser catalogados e recuperados de maneira eficiente, utilizando por exemplo, informações de alto nível detectadas por sistemas de recuperação de dados, o que facilita a busca para o usuário final. A classificação automática de vídeos por gênero é o primeiro passo para catalogar grandes coleções. Nesse contexto, muitos trabalhos vêm sendo feitos utilizando redes neurais convolucionais e algoritmos de classificação clássicos, ou até combinações do último, que virou uma técnica difundida pela quantidade de características que podem ser combinadas, melhorando assim o desempenho final. Neste trabalho, foram avaliadas diferentes estratégias para combinar redes neurais convolucionais com o intuito de melhorar o desempenho em tarefas de classificação de gênero de vídeo.**

***Abstract*—Video data generation currently occurs very quickly and on a large scale, mobile devices with this capability are increasingly widespread, generating a large amount of data, which need to be cataloged and retrieved efficiently, using for example, high level information detected by data recovery systems, which facilitates the search for the end user. Automatic prediction of videos by genre is the first step in cataloging large collections. In this context, many works have been done using convolutional neural networks and classical classification algorithms, or combinations of the last one, which has become a technique diffused by the amount of characteristic that can be combined, thus improving the final performance. In this work, different strategies for combining convolutional neural networks were evaluated aiming at improving the performance on video genre classification tasks.**

I. Introdução

Hoje em dia, as pessoas podem criar, compartilhar e armazenar vídeos em qualquer lugar a partir de dispositivos móveis, redes sociais e ambientes em nuvem, levando a um aumento contínuo na produção de vídeos e criando acervo crescente desse material. Nesse cenário, é fundamental ter ferramentas apropriadas para catalogar e recuperar dados de vídeo, atraindo um interesse cada vez maior em sistemas capazes identificar e selecionar informação relevante de acordo com as necessidades de um usuário [1].

A classificação automática de um vídeo de acordo com o seu gênero é um dos primeiros passos para catalogar coleções grandes e heterogêneas de material de vídeo [2]. Ao longo das últimas décadas, várias tarefas de categorização de vídeo têm sido abordadas usando técnicas de aprendizado de máquina [3]. Dentre os diversos métodos propostos na literatura, o aprendizado profundo (do inglês, *deep learning*), mais especificamente, as redes neurais convolucionais (do inglês, *convolutional neural networks* - CNNs) têm se estabelecido recentemente como o estado da arte na resolução de uma grande variedade de problemas em computação visual [2].

Em muitas aplicações envolvendo computação visual e inteligência de máquina, a fusão de características extraídas dos dados provou ser indispensável para melhorar o desempenho [4]. A principal motivação das abordagens de fusão consiste em obter uma representação mais precisa dos dados a partir da combinação de características distintas [5].

Nesse contexto, a proposta deste trabalho foi avaliar diferentes estratégias para combinar CNNs com o intuito de melhorar a eficácia na tarefa de classificação de gênero de vídeo. Mais especificamente, duas estratégias foram consideradas: (1) média, na qual as CNNs foram combinadas tomando-se a média de sua saída; e (2) concatenação, em que a saída das CNNs foram concatenadas em uma única camada.

Experimentos foram realizados em um conjunto de dados da tarefa de atribuição de gêneros (do inglês, *genre tagging task*) do MediaEval 2012 [6]. Nesses experimentos, foram avaliadas quatro CNNs distintas e todas as suas possíveis combinações usando as duas estratégias supracitadas. Os resultados obtidos foram promissores, indicando que a fusão das CNNs pode melhorar o desempenho na tarefa de classificação de gênero de vídeo, alcançando uma eficácia comparável ao estado da arte.

II. Conceitos Básicos

A. Redes Neurais Convolucionais

Neste trabalho, foram usados métodos de aprendizado profundo, mais especificamente, CNNs. Elas são redes neurais constituídas de neurônios que possuem pesos e vieses que podem ser aprendidos. Cada neurônio recebe algumas entradas, executa um produto escalar e, opcionalmente, segue-o com uma não linearidade. Toda a rede ainda expressa uma única função de pontuação diferenciável: dos pixels da imagem bruta em uma extremidade às pontuações da classe na outra. E eles ainda têm uma função de perda (por exemplo, SVM / Softmax) na última camada (totalmente conectada). A grande diferença entre uma rede neural tradicional e uma CNN é que as arquiteturas convolucionais fazem a suposição explícita de que as entradas são imagens, permitindo codificar certas propriedades na arquitetura [7].

Uma CNN é uma sequência de camadas. Cada camada de uma CNN realiza uma série de ativações na camada seguinte a partir de uma função diferenciável. São usados três tipos principais de camadas para construir arquiteturas CNN: camada convolutiva, camada de agrupamento e camada totalmente conectada. Cada camada é especificada a seguir e o diagrama de uma CNN é apresentado na Figura 1 [8].

- **Camada Convolutiva.** Os parâmetros da camada convolutiva consistem em um conjunto de filtros que podem ser aprendidos. Cada filtro é pequeno espacialmente (ao longo da largura e altura), mas se estende por toda a profundidade do volume de entrada. Por exemplo, um filtro típico em uma primeira camada de uma CNN pode ter tamanho 5x5x3 (ou seja, 5 pixels de largura e altura, e 3 porque as imagens têm profundidade 3, os canais de cor). Durante o passe para frente, cada filtro é deslizado (mais precisamente, convolvido) pela largura e altura do volume de entrada, calculando os produtos de ponto entre as entradas do filtro e a entrada em qualquer posição. À medida que o filtro é deslizado pela largura e altura do volume de entrada, um mapa de ativação bidimensional é produzido, fornecendo as respostas desse filtro em todas as posições espaciais.
- **Camada de Agrupamento.** É comum inserir periodicamente uma camada de agrupamento (do inglês, *pooling*) entre camadas convolutivas sucessivas em uma arquitetura CNN. Sua função é reduzir progressivamente o tamanho espacial da representação para reduzir a quantidade de parâmetros e computação na rede e, portanto, também controlar o sobreajuste (do inglês, *overfitting*)[1]. A camada de agrupamento opera independentemente em cada fatia de profundidade da entrada e a redimensiona espacialmente, usando a operação max (normalmente, porém a operação pode ser modificada).
- **Completamente Totalmente Conectada.** Assim como em redes neurais tradicionais, neurônios em uma camada totalmente conectada (do inglês, *fully connected*) têm conexões completas com todas as ativações na camada anterior.

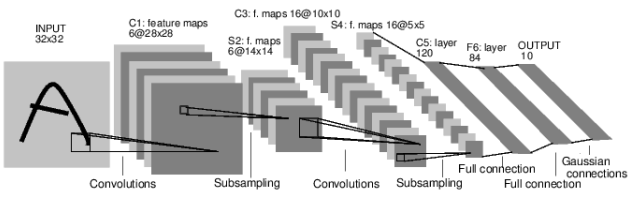

Figura 1. Diagrama que representa as camadas de uma CNN [8].

B. Utilização de Redes Pré Existentes

Existem basicamente três formas de utilizar uma arquitetura de CNN já desenvolvida [9]:

[1]Fenômeno que ocorre se a rede se sobreajusta para a base de treino, gerando resultados ruins nos testes.

- **Treino Completo.** Os pesos e vieses associados a neurônios e filtros convolutivos são inicializados com valores aleatórios e ajustados pelo algoritmo de retropropagação a partir das predições feitas com a base de treino. É mais recomentado utilizar essa abordagem quando a base é grande o suficiente para chegar a um bom ajuste de pesos, gerando uma rede com muito potencial de predição para o problema em específico.
- **Transferência do Aprendizado** (do inglês, *Transfer Learning*). Consiste na utilização de pesos pré-definidos em uma rede que já foi treinada para algum tipo de problema de classificação, modificando apenas as camadas totalmente conectadas (opcional) e o classificador, uma boa abordagem para quando se tem poucas instâncias em uma base ou quando não se tem muito tempo para o treinamento.
- **Ajuste Fino** (do inglês, *Fine Tuning*). E um conjunto das duas abordagens anteriores, em que algumas camadas são congeladas e iniciadas com pesos pré-definidos, porém, as últimas camadas convolutivas, as camadas totalmente conectadas e o classificador são retreinados.

C. Fusão de Redes

Em geral, a ideia de fundir CNNs normalmente remete a um processo de fusão tardia (i.e. *late fusion*) da última camada, a camada preditiva [10]. Porém, neste trabalho, toda a fusão foi feita antes da última camada.

III. Implementação

A implementação foi totalmente desenvolvida na linguagem de programação Python, que facilita a prototipagem e análise de resultados. A utilização de CNNS foi facilitada utilizando o framework Keras e GPUs NVIDIA.

A. A Base de Dados

Os experimentos foram realizados em um conjunto de dados da tarefa de atribuição de gêneros do MediaEval 2012 [6]. Essa base é composta de 14.838 vídeos (3.288 horas) coletados do blip.tv e é dividida entre treino (36% ou 5.288 vídeos) e teste (64% ou 9.550 vídeos). Os vídeos são distribuídos entre 26 categorias de gênero definidas pela plataforma blip.tv. O desafio principal dessa base é a grande diversidade de gêneros, assim como a grande variedade de conteúdos visuais em cada categoria de gênero. Vale apontar que a base é desbalanceada, por exemplo, há uma classe chamada "padrão" onde está grande parte dos vídeos. Todas essas características apontadas fazem com que o problema seja considerado difícil.

A base tem quadros-chave (imagens) extraídos de cada vídeo de maneira que eles podem ser utilizadas para a sua classificação, não há um padrão de tamanho ou quantidade de quadros por vídeo. O tamanho das imagens foi modificado para que se adaptasse a entrada padrão de imagens do Keras, que é 224x224 pixels (uma imagem quadrada). Porém, as imagens tinham tamanhos diferentes e também não eram quadradas. Foi então decidido retirar as bordas pretas (superior e inferior) e cortar as laterais proporcionalmente até que

a imagem ficasse quadrada (mantendo o centro, o foco do vídeo). Após isso, ela foi redimensionada para o tamanho desejado.

B. Combinação de Redes Neurais

Foram consideradas diferentes CNNs utilizadas em desafios de classificação de imagens. Um importante critério levado em consideração foi o quão diferentes são as CNNs que serão fundidas (conceito utilizado em trabalhos anteriores [11]). Quanto mais diferentes, melhor eles serão para serem fundidos. Porém, também deve haver um bom desempenho geral de cada classificador individualmente para que o resultado seja satisfatório. As CNNs selecionadas foram: VGG-19 (VGG) [12], DenseNet (DEN) [13], Inception-V3 (INC) [14] e ResNet-50 (RES) [15].

A estratégia utilizada para o aproveitamento das CNNs já desenvolvidas foi a de *transferência do aprendizado*, com todas camadas convolutivas congeladas e sem o topo (camadas totalmente conectadas). A partir daí foi adicionada a cada uma dessas redes uma camada *GlobalAveragePooling2D* para redução de parâmetros e de dimensões de saída seguida de uma totalmente conectada de 512 neurônios (padronizando a saída de todas as redes para 512 neurônios) e uma camada *softmax* com 26 neurônios.

Independente da combinação das redes, todas seguiram o mesmo padrão: uma entrada única, que é levada paralelamente a cada rede, gerando uma camada totalmente conectada com 512 neurônios para cada rede, as quais são combinadas por meio de uma camada de agrupamento (e.g., concatenação, média, máximo, dentre outros), que é totalmente conectada a uma camada de saída com 26 neurônios, na qual é gerada predição final. O diagrama de exemplo de combinação de três redes pode ser visto na Figura 2.

Todas as combinações de redes foram consideradas, levando em conta permutações com 1, 2, 3 ou 4 CNNs. O exemplo de implementação da rede combinado as 4 CNNs com estratégia de agrupamento por média pode ser visto no Algoritmo 1.

Algoritmo 1. Código Fusão das Quatro Redes

```
#creating nets
vgg19 = VGG19(include_top=False,
              weights='imagenet',
              input_shape=input_shape)
inception = InceptionV3(include_top=False,
                        weights='imagenet',
                        input_shape=input_shape)
densenet = DenseNet201(include_top=False,
                       weights='imagenet',
                       input_shape=input_shape)
rn50 = ResNet50(include_top=False,
                weights='imagenet',
                input_shape=input_shape)

commonInput = Input(shape=input_shape)
out1 = rn50(commonInput)
out1 = GlobalAveragePooling2D()(out1)
out1 = Dense(512, activation='relu')(out1)

out2 = vgg19(commonInput)
out2 = GlobalAveragePooling2D()(out2)
out2 = Dense(512, activation='relu')(out2)

out3 = densenet(commonInput)
out3 = GlobalAveragePooling2D()(out3)
out3 = Dense(512, activation='relu')(out3)

out4 = inception(commonInput)
out4 = GlobalAveragePooling2D()(out4)
out4 = Dense(512, activation='relu')(out4)

mergedout = Average()([out1, out2, out3, out4])

newModel = Model(commonInput, mergedout)
newModel = add_new_last_layer(newModel, 26)
setup_to_transfer_learn(newModel, rn50, vgg19, densenet,
    inception)
```

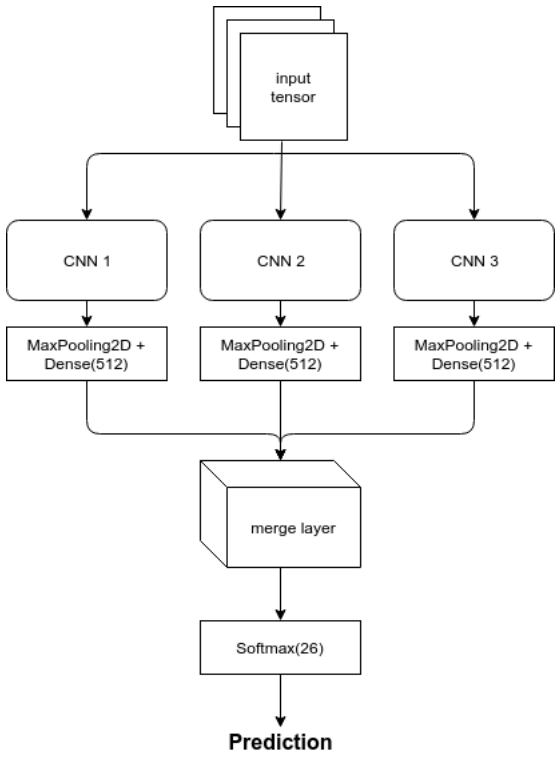

Figura 2. Diagrama que representa a fusão de CNNs abordada.

O treinamento foi feito utilizando 75% da base de treino e a validação com 25% da mesma. Para lidar com o desbalanceamento, cada amostra foi ponderada com o seguinte valor:

$$\text{peso amostra da classe } i = \frac{\text{qtd. total de amostras da base}}{\text{qtd. amostras da classe } i}$$

Os parâmetros usados foram: *batches* de tamanho 64 e *early stopping* (o treinamento pára quando o valor da perda no conjunto de validação não diminui duas épocas seguidas).

IV. Resultados

A medida utilizada para aferir a eficácia dos classificadores finais foi o MAP (do inglês, *Mean Average Precision*) após atribuir a classe predita a cada vídeo de teste. Como o classificador recebe como entrada imagens, a predição de um video foi relacionada a todas imagens que compõem o mesmo. Esse cálculo de predição para o vídeo foi feito de duas maneiras:

1) **Voto majoritário (MV):** a classe atribuída ao vídeo foi obtida a partir da maior quantidade de classes com o maior valor dentro dos vetores de probabilidades e

a probabilidade do vídeo estar associado a essa classe foi calculada utilizando a média das probabilidades da classe vencedora de cada vetor que votou na maioria.

2) **Média normalizada dos vetores de probabilidades (AVG):** a classe atribuída ao vídeo foi a de maior probabilidade após gerar a média normalizada de todos os vetores de probabilidades e a probabilidade do vídeo estar associado a essa classe é simplesmente o maior valor desse vetor gerado.

Duas estratégias foram avaliadas para implementar a camada de agrupamento que combina a camada totalmente conectada de 512 neurônios de todas as CNNs: *média* e *concatenação*. Os resultados obtidos para as diferentes combinações avaliadas são apresentados na Tabela I.

Tabela I
RESULTADOS OBTIDOS PARA MEDIDA MAP CONSIDERANDO DIFERENTES ESTRATÉGIAS DE COMBINAÇÃO DE CNNS.

CNNs	**Média**		**Concatenação**	
	MV	*AVG*	*MV*	*AVG*
DEN	18,88%	18,36%	**18,88%**	18,36%
INC	12,87%	12,61%	12,87%	12,61%
RES	9,44%	8,79%	9,44	8,79%
VGG	10,44%	9,26%	10,44%	9,26%
INC + DEN	**18,95%**	**19,58%**	18,14%	**18,40%**
RES + DEN	13,49%	14,48%	14,37%	14,31%
RES + INC	11,71%	12,04%	8,81%	9,18%
VGG + DEN	15,81%	16,59%	14,21%	14,48%
VGG + INC	14,39%	15,82%	10,85%	11,67%
VGG + RES	11,86%	11,55%	12,08%	11,44%
VGG + RES + INC	11,60%	12,07%	11,39%	11,53%
VGG + RES + DEN	15,48%	15,08%	13,09%	12,20%
VGG + DEN + INC	15,26%	16,03%	15,29%	15,60%
DEN + INC + RES	13,69%	14,86%	12,45%	13,10%
VGG + DEN + INC + RES	16,73%	16,46%	12,20%	12,95%

Em diversos resultados, fundir duas CNNs teve um desempenho melhor que usar cada uma individualmente. Aparentemente, a camada de agrupamento das CNNs procura utilizar o melhor de cada rede. O maior valor do MAP foi obtido pela fusão das redes INC e DEN, que foram as CNNs que atingiram os dois maiores MAPs considerando os desempenhos individuais de cada rede. O resultado da combinação INC + DN superou os reportados pelos trabalhos de Jan Schlüter et al. [16], que atingiu 19,41% utilizando o conteúdo visual e sonoro; e o de Almeida et al. [17], que atingiu 12,38% utilizando apenas o conteúdo visual.

V. CONCLUSÃO

A fusão de CNNs se mostrou interessante para o tarefa de classificação de gênero de vídeo, superando resultados de diferentes abordagens utilizadas no MediaEval2012. Vale a pena mencionar que os melhores resultados nessa tarefa utilizaram diferentes tipos de dados (e.g., áudio, imagens e meta-dados) e estratégias mais elaboradas de fusão de algoritmos clássicos.

Trabalhos futuros envolvem a utilização de diferentes tipos de dados, como o áudio e os meta-dados da base, além de testar outras abordagens de fusão.

AGRADECIMENTOS

Os autores agradecem o apoio financeiro da Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP (processo 2016/06441-7) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (processos 423228/2016-1, 313122/2017-2 e 123467/2017-9) e o apoio da NVIDIA Corporation com a doação da GPU Titan Xp usada para esta pesquisa.

REFERÊNCIAS

[1] W. Hu, N. Xie, L. Li, X. Zeng, and S. J. Maybank, "A survey on visual content-based video indexing and retrieval," *IEEE Trans. Systems, Man, and Cybernetics, Part C*, vol. 41, no. 6, pp. 797–819, 2011.

[2] J. Chen, H. Lu, R. Wei, C. Jin, and X. Xue, "An effective method for video genre classification," in *IEEE International Conference on Image and Video Retrieval (CIVR'10)*, 2010, pp. 97–104.

[3] I. Mironica, B. Ionescu, P. Knees, and P. Lambert, "An in-depth evaluation of multimodal video genre categorization," in *IEEE International Workshop on Content-Based Multimedia Indexing (CBMI'13)*.

[4] J. Almeida, D. C. G. Pedronette, and O. A. B. Penatti, "Unsupervised manifold learning for video genre retrieval," in *Iberoamerican Congress on Pattern Recognition (CIARP'14)*, 2014, pp. 604–612.

[5] H. K. Ekenel and T. Semela, "Multimodal genre classification of TV programs and youtube videos," *Multimedia Tools Appl.*, vol. 63, no. 2, pp. 547–567, 2013.

[6] S. Schmiedeke, C. Kofler, and I. Ferrané, "Overview of mediaeval 2012 genre tagging task," in *Working Notes Proceedings of the MediaEval 2012 Workshop (MEDIAEVAL'12)*, 2012.

[7] Y. Bengio, "Learning deep architectures for ai," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[8] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[9] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Annual Conference on Neural Information Processing Systems (NIPS'14)*, 2014, pp. 3320–3328.

[10] H. Ergun, Y. C. Akyuz, M. Sert, and J. Liu, "Early and late level fusion of deep convolutional neural networks for visual concept recognition," *Int. J. Semantic Computing*, vol. 10, no. 3, pp. 379–398, 2016.

[11] V. Lúcio, F. A. Faria, and J. Almeida, "Um sistema de fusão de classificadores aplicado à fenologia," in *Workshop of Undergraduate Works (WUW) in the Conference on Graphics, Patterns and Images (SIBGRAPI'17)*, 2017, pp. 1–4.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'17)*, 2017, pp. 2261–2269.

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'16)*, 2016, pp. 2818–2826.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'16)*, 2016, pp. 770–778.

[16] I. M. M. S. Jan Schlüter, Bogdan Ionescu, "An uninformed approach to violence detection in hollywood movies," in *Working Notes Proceedings of the MediaEval 2012 Workshop (MEDIAEVAL'12)*, 2012.

[17] J. Almeida, T. Salles, E. F. Martins, O. A. B. Penatti, , R. S. Torres, M. A. Gonçalves, and J. M. Almeida, "UNICAMP-UFMG at mediaeval 2012: Genre tagging task," in *Working Notes Proceedings of the MediaEval 2012 Workshop (MEDIAEVAL'12)*, 2012.