RELATÓRIO PARCIAL

Atenção visual para sistemas robóticos com Deep Learning

Aluno: Erik de Godoy Perillo
Orientadora: Profa. Dra. Esther Luna Colombini

Instituto de Computação
Universidade Estadual de Campinas

16 de fevereiro de 2018

(a) (b)

Figura 1. Examplo de saliência visual. b) é o mapa de saliência onde regiões mais brilhantes (cores mais quentes) representam regiões mais salientes para humanos na imagem original a).

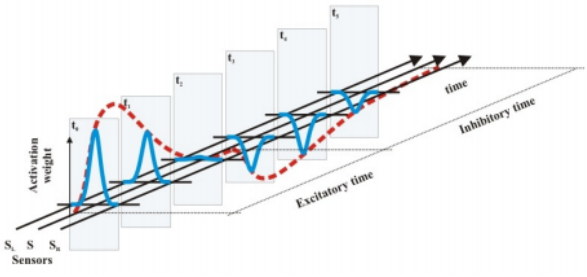

Figura 2. Exemplo de variação de estímulo atencional no tempo com IOR.

I. Introdução

A capacidade de percepção e construção de um modelo da realidade ao seu redor é fundamental para que sistemas robóticos interajam com o ambiente e executem tarefas diversas e complexas que podem ter as mais variadas utilidades para os humanos. Um componente fundamental para isso é a habilidade de dar foco apenas ao relevante, evitando assim o processamento desnecessário de enormes quantias de dados.

A atenção é um processo que faz parte do dia a dia de diversos seres vivos em diversas maneiras e é razoável inspirar-se nela para a construção de mecanismos semelhantes para a construção de sistemas de inteligência artificial em máquinas. Tal área tem sido foco de estudo há anos, resultando em diversas teorias em psicologia sobre a atenção humana que inspiraram a implementação de modelos computacionais bem sucedidos, que geram imagens semelhantes ao exemplo da figura 1.

Em trabalhos anteriores, foi desenvolvido um modelo de saliência visual com uma rede neural convolucional eficiente [4]. A arquitetura da rede permitiu que fossem atingidos resultados comparáveis ao estado da arte, com um número de parâmetros reduzido em 75%. Neste trabalho, objetivamos construir um modelo de saliência visual eficiente para vídeo.

A. Objetivos da primeira parte do projeto

Os objetivos principais para o primeiro semestre do trabalho eram:

- Revisão bibliográfica sobre atuais trabalhos em vídeo.
- Escolha das técnicas mais adequadas para o processo atencional e adaptação do das técnicas conhecidas para vídeo.
- Implementação de um modelo atencional para vídeo.

Embora haja relevantes avanços recentes na detecção de saliência visual para imagens estáticas, não há ainda na literatura um considerável número de trabalhos focando em vídeo. Ainda assim, com estudo da literatura já existente em atenção e no número reduzido de trabalhos que focam em vídeo, houve progresso.

II. Resumo das atividades

A. Revisão Bibliográfica

Um conceito importante para o entendimento da literatura do meio é o de *Bottom-up vs. Top-down*: Por componente *bottom-up* de atenção entende-se saliências instintivas percebidas por mudanças e/ou contrastes muito grandes em uma cena. O componente *top-down* é aquele que dá saliência variável às *features* de acordo com a meta do agente do momento. A maioria dos modelos computacionais baseia-se em teorias formadas na psicologia. Duas das mais famosas são a *Feature Integration Theory* (FIT) [13] e a *Guided Search* [6]. Ambas provêm contribuições importantes para o entendimento dos processos de saliência visual.

Para saliência visual estática (única imagem), relevantes avanços foram feitos nos últimos anos com o uso de *Deep Learning*. Modelos como *DeepFix* [9], *SALICON* [7] e *MLNet* [3] usam arquiteturas de redes neurais completamente convulucionais. Muitos modelos atuais usam redes pré-treinadas como a VGG-16 com uma etapa de refinamento. Tais modelos são o estado da arte atual para imagens estáticas.

Saliência estática, entretanto, não é suficiente. Em humanos, há um importante fenômeno: IOR (do inglês: *Inhibition of Return*.) Tal efeito faz com que o foco atencional mude com o passar do tempo. Isso provê uma vantagem evolutiva pois permite ao ser explorar o ambiente ao seu redor. Assim, uma região de uma imagem que um modelo estático identificou como altamente saliente pode não o ser dependendo do contexto da sequência de imagens e de focos atencionais (figura 2).

O principal trabalho estudado que aborda a saliência para vídeo [14], onde um modelo com uma rede neural convolucional *end-to-end* é usada para a geração de mapas de saliência visual. Sua arquitetura combina uma

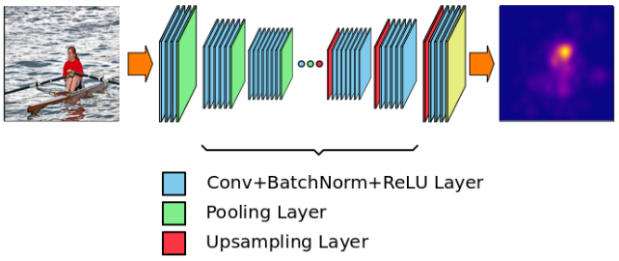

Figura 3. Arquitetura do modelo proposto para imagens estáticas.

rede convolucional treinada para imagens estáticas a uma nova rede convolucional para sequências de frames de vídeos. Tal modelo obtém resultados melhores que apenas a rede para imagens estáticas no dataset DAVIS [11].

A revisão bibliográfica feita mostrou-se fundamental para os trabalhos desenvolvidos no semestre.

B. Modelo atencional

1) Modelo estático: O primeiro modelo desenvolvido em trabalhos anteriores *DeepPeek*, obteve métricas consistentemente entre os dez primeiros colocados no *MIT Saliency Benchmark* [1]. Sua arquitetura se destaca por ser uma rede neural totalmente convolucional feita especificamente para a detecção de saliência visual e treinada do zero, com 3717841 parâmetros. Sua entrada consiste em imagens no espaço de cor LAB, mais semelhante ao sistema visual humano.

A primeira contribuição deste trabalho foi uma versão melhorada do modelo antigo para imagens estáticas (figura 3): com menos parâmetros (2323145) e mesmo desempenho, a arquitetura é baseada na *Unet* [12], e é mais flexível para futuras alterações que adaptem o modelo para vídeo. O novo modelo também usa imagens em LAB e foi treinado usando-se os *datasets SALICON* [7] e *Judd* [8].

2) Implementação de IOR: O principal objetivo de uma implementação de *IOR* para mapas de saliência contínuos é simular o curso atencional em humanos. Isso é feito aumentando-se a probabilidade de áreas com menos foco atencional no passado de serem atendidas no futuro e, de modo análogo, diminuindo-se a probabilidade de áreas com grande saliência de serem altamente avaliadas em um período de tempo próximo. Para isso, dado um fluxo de imagens F, com S_t a saliência de um ponto (x, y) no mapa no momento t, computa-se o mapa de inibição de retorno IOR_{t+1} da seguinte forma:

$$IOR_{t+1} = \frac{k_p P(t) + k_i I(t) + k_d D(t) + k_c C}{k_p + k_i + k_d + k_c}$$

Tabela I
NÚMERO DE FILTROS EM CADA CAMADA DA REDE PROPOSTA.

Bloco	Tipo	Número de filtros
input/downsample	conv	32
	conv	32
	conv	32
downsample	conv	48
	conv	48
	conv	48
downsample	conv	64
	conv	64
	conv	64
downsample	conv	96
	conv	96
	conv	96
downsample	conv	128
	conv	128
	conv	128
upsample	conv	180
	conv	180
	conv	180
upsample	conv	128
	conv	128
	upconv	128
upsample	conv	96
	conv	96
	upconv	96
upsample	conv	64
	conv	64
	upconv	64
upsample/output	conv	48
	conv	48
	upconv	48

com

$$P(t) = 1 - S_t$$

$$I(t) = \frac{1}{1 + \int_0^t S_\tau d\tau}$$

$$D(t) = (1 + S'(x)|_{x=t})/2$$

C sendo o mapa constante.

$P(t)$ diz respeito à parte proporcional, o mapa de saliência no momento imediato antes do novo mapa. $I(t)$ diz respeito à parte integrativa, isto é, o quanto esta área foi saliente no passado. Quanto mais saliente uma região já foi, maior é sua inibição. $D(t)$ diz respeito à parte derivativa. Se a saliência mudou rapidamente, esta área merece um realce. C é o mapa constante. Todos os mapas variam no intervalo $[0, 1]$. k_p, k_i, k_d, k_c são constantes que controlam a influência de cada componente no mapa final. Com base nos experimentos realizados, os melhores valores para k_p, k_i, k_d, k_c são, respectivamente, $0.2, 0.4, 0.1, 0.3$. Figura 4 ilustra um exemplo de mapa de IOR gerado com a técnica.

Figura 4. Exemplo de geração de mapa de IOR. À esquerda, um frame do vídeo. No centro, o mapa de saliência correspondente ao frame. À direita, o mapa de IOR, com valores mais escuros onde há maior inibição.

3) Modelo final: O modelo final combina a saída da rede estática com IOR. Dado R_{t+1}, o valor da saliência calculado pelo método estático, o mapa final S_{t+1} é dado por:

$$S_{t+1} = R_{t+1} * IOR_{t+1}$$

4) Implementação: Todas as etapas do modelo descritas aqui foram implementadas. A linguagem utilizada foi *Python*, usando-se *OpenCV*, *numpy* e *tensorflow*. O código está disponível de modo público em `https://github.com/erikperillo/att`.

C. Avaliação de desempenho

Foi feita uma pesquisa das métricas e *datasets* mais usados para avaliação dos modelos computacionais, pois isso é muito importante para o avanço do trabalho.

1) SAVAM: SAVAM é um dataset com 41 vídeos com informação de mapas contínuos de saliência obtida por dados de *eye-tracking* de observadores [5]. Por conter vídeos de variados contextos, o dataset foi considerado a mais completa das opções e então utilizado para avaliação do desempenho.

2) Métricas: Muitas das métricas aqui mostradas são discutidas em [2], onde a aplicabilidade, significado, vantagens e desvantagens de cada uma são discutidas mais a fundo. Aqui daremos apenas uma breve descrição das mesmas.

- `Similarity`
  Definido como:
  $$SIM(P,Q) = \frac{1}{N}\sum_{i=1}^{N} min(P_i, Q_i)$$
  Com P e Q indo de 0 a 1. Um valor de 1 define mapas idênticos e 0 mapas totalmente diferentes. Ambos os mapas são normalizados pela soma de cada um. Usado em [10].
- `Correlation Coefficient`
  `Similarity` penaliza *false negatives* mais que *false positives* [2]. A métrica `CC` trata os dois simetricamente. Dada por:
  $$CC(P,Q) = \frac{cov(P,Q)}{\sigma(P)\sigma(Q)}$$
  Onde P é normalizado no intervalo $[0, 1]$. Usado em [10].
- `Mean-square error`
  Dado por:
  $$MSE(P,Q) = \frac{1}{N}\sum_{i=1}^{N}(P_i - Q_i)^2$$
  Usado em [10].

Tabela II
RESULTADOS DO MODELO ESTÁTICO E COM IOR.

Modelo	Métrica	Valor médio
Estático	CC	0.41
Estático	SIM	0.37
Estático	MSE	0.40
Estático + IOR	CC	0.46
Estático + IOR	SIM	0.41
Estático + IOR	MSE	0.11

D. Resultados

Usando-se apenas o modelo para saliência estática, obteve-se um resultado pior do que quando o ajuste para IOR foi feito. A tabela II-D ilustra os resultados.

E. Conclusão

Neste trabalho, obteve-se um modelo de rede neural convolucional para imagens estáticas eficiente para detecção de saliência visual. Combinado a um método de simulação do fenômeno de *Inhibition of Return*, o modelo obteve resultados melhores para detecção de saliência visual em vídeos.

III. PRÓXIMOS PASSOS

O próximo passo é a concepção de um sistema que usa aprendizado de máquina de uma maneira *end-to-end*, sem a necessidade de entrada manual de IOR.

REFERÊNCIAS

[1] Zoya Bylinskii et al. *MIT Saliency Benchmark*. http://saliency.mit.edu/index.html. 2016. (Acesso em 27/09/2016).

[2] Zoya Bylinskii et al. "What do different evaluation metrics tell us about saliency models?" Em: (2016).

[3] Marcella Cornia et al. "A Deep Multi-Level Network for Saliency Prediction". Em: *arXiv preprint arXiv:1609.01064* (2016).

[4] Esther Colombini Erik Perillo. "Processos Atencionais e Aprendizado de Máquina para Sistemas Robóticos". Em: (2017).

[5] Yury Gitman et al. “Semiautomatic Visual-Attention Modeling and Its Application to Video Compression”. Em: *2014 IEEE International Conference on Image Processing (ICIP) (ICIP 2014)*. Paris, France, out. de 2014, pp. 1105–1109.

[6] Susan L. Franzel Jeremy M. Wolfe Kyle R. Cave. “Guided Search: an alternative to the feature integration model for visual search”. Em: (1989).

[7] Ming Jiang et al. “Salicon: saliency in context”. Em: *CVPR* (2015).

[8] T. et al Judd. *Learning to predict where people look*. 2016. URL: http://people.csail.mit.edu/tjudd/WherePeopleLook/index.html (acesso em 04/10/2016).

[9] Srinivas S S Kruthiventi, Kumar Ayush e R. Venkatesh Babu. “DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations”. Em: *arXiv preprint arXiv:1510.02927* (2015).

[10] *MIT saliency benchmark*. 2016. URL: http://saliency.mit.edu/index.html (acesso em 27/09/2016).

[11] Jordi Pont-Tuset et al. “The 2017 DAVIS Challenge on Video Object Segmentation”. Em: *arXiv:1704.00675* (2017).

[12] Olaf Ronneberger, Philipp Fischer e Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. Em: *CoRR* abs/1505.04597 (2015). arXiv: 1505.04597. URL: http://arxiv.org/abs/1505.04597.

[13] A M Treisman e G Gelade. “A feature-integration theory of attention”. Em: *Cognit Psychol* 12.1 (jan. de 1980), pp. 97–136.

[14] Wenguan Wang, Jianbing Shen e Ling Shao. “Deep Learning For Video Saliency Detection”. Em: *CoRR* abs/1702.00871 (2017). arXiv: 1702.00871. URL: http://arxiv.org/abs/1702.00871.