doi:10.20396/revpibic.v0i0.id

Atenção visual para sistemas robóticos com Deep Learning

Erik de Godoy Perillo*, Esther Luna Colombini

Resumo
Detecção de regiões salientes é componente fundamental da nossa visão. Neste projeto, estendemos trabalhos anteriores e introduzimos uma rede convolucional eficiente para detecção de saliência em imagens e colocamos componentes que melhoram seu desempenho em vídeo.

Palavras-chave:
Deep Learning, Robótica, Atenção Visual.

Introdução

Um componente fundamental para que formas de vida complexas interajam de forma eficiente com o ambiente é a habilidade de dar foco apenas ao relevante, evitando assim o processamento desnecessário de enormes quantias de dados[1].
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.
Em trabalhos anteriores, foi desenvolvido um modelo de saliência visual com uma rede neural convolucional eficiente. 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 com arquitetura ainda mais leve e também eficiente para vídeo.

Resultados e Discussão

Baseando-se em trabalhos anteriores, foi desenvolvida uma rede neural convolucional com encoder e decoder, no estilo de redes Unet. A entrada são imagens no espaço de cor LAB e a saída são os mapas de saliência[2]. Treinou-se em 15000 imagens do dataset SALICON. A figura 1 ilustra a rede. Para detecção de saliência em vídeos, são necessários ajustes. Seres vivos não mantêm o foco visual fixo em um componente do que veem, mas o foco muda com o tempo por um processo conhecido por Inhibition of Return (IOR), que basicamente torna mais difícil focar em áreas já muito focadas com o passar do tempo. O IOR foi aplicado por meio de um ajuste no mapa de saliência de saída produzido pela rede. É calculado um mapa de IOR e o mapa final do tempo t+1 é dado por:

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

com

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

e o cálculo de cada mapa proporcional, integrativo e derivativo ilustrado na figura 2. Comparou-se o desempenho no dataset SAVAM de saliência em vídeo.

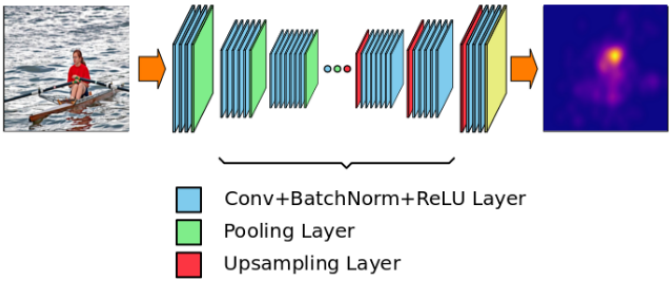

Figura 1. Ilustração da arquitetura de rede neural usada no trabalho.

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

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

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

Figura 2. Equações dos mapas intermediários de IOR.

Tabela 1. Comparação de desempenho do modelo estático e com IOR no dataset SAVAM.

Modelo	Métrica	Valor
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**

Conclusão

Neste trabalho, desenvolvemos uma rede convolucional eficiente para detecção de saliência visual e foi adicionada uma extensão para vídeos que melhorou o desempenho.

Agradecimentos

Agradecemos ao CNPq pelo fomento à pesquisa.

---

[1]Treisman, A. M.; Gelade, G. A feature-integration theory of attention. **1980.**
[3]Frintrop, S. VOCUS: a visual attention system for object detection and goal-directed search. **2005.**

(cc) BY-NC-ND