Sensitivity of Deep CNNs to noise in training Data in the problem of visual saliency prediction

Souad CHAABOUNI
LaBRI UMR 5800,
University of Bordeaux

Jenny BENOIS-PINEAU
LaBRI UMR 5800,
University of Bordeaux

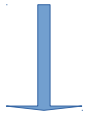
Chokri BEN AMAR
REGIM-Lab LR11ES48,
University of Sfax

Plan

1- Introduction and motivations.
2- Deep CNNs for saliency prediction.
3- Sensitivity of Deep CNNs to noise.
4- Experiment and results.
5- Conclusion and perspectives.

Introduction
Motivations

Prediction of saliency in images and video by machine learning methods is an intensive research subject.



These methods require a large amount of training data. when the data are gathered can produce the noise as well.

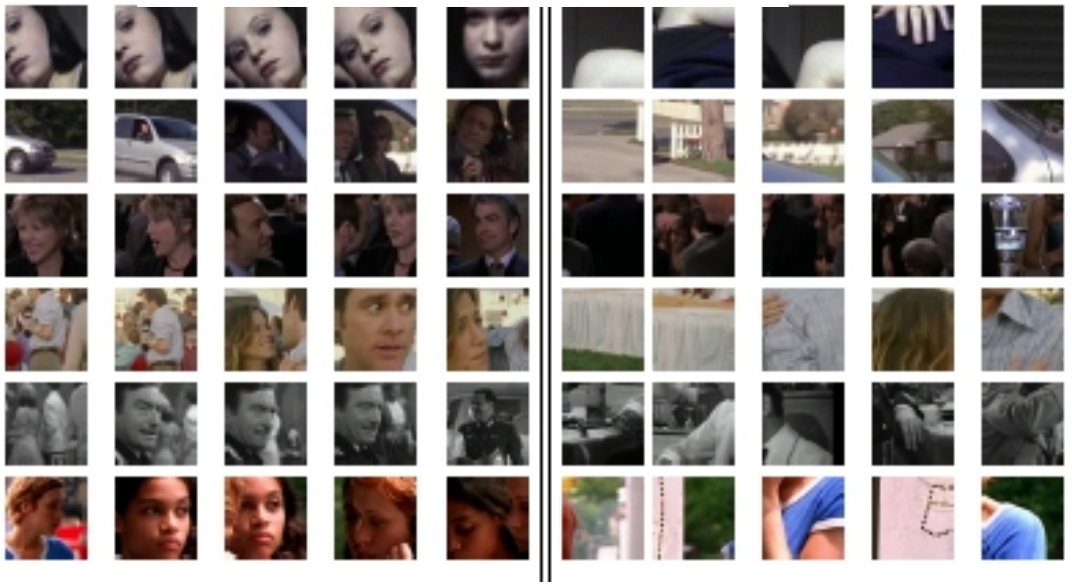
In data mining → noise has two different main source

- **Implicit errors** Different type of measurement tools induces
- **Random errors** introduced by experts or batch processes

The main contribution of this paper is to identify how noise of data impacts performance of deep networks in the problem of visual saliency prediction. we will focus on noise produced by random errors.

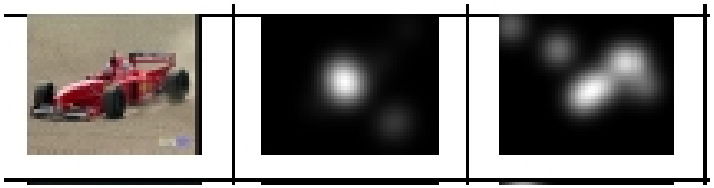
Solution: We have conducted two experiments on the large database HOLLYWOOD[6].

1. For square patches in video frames to solve a two-class classification problem: salient-non-salient



Hollywood dataset

2. Build a dense predicted saliency map from sparse decisions



Saliency map predicted with Deep CNNs #frame from IRCCyN dataset.

Going Deeper : Input layer

One patch P is a vector in $R^{t \times t \times n}$: n stands for the quantity of primary feature maps.

- n = 3 : RGB planes of a colour video sequence are used,
- n = 4 : adding th emagnitude of residual motion to RGB planes.
- n = 8 : using spatial contrasts with residual motion.
- n= 11: adding spatial contrast and residual motion to RGB/HSV planes

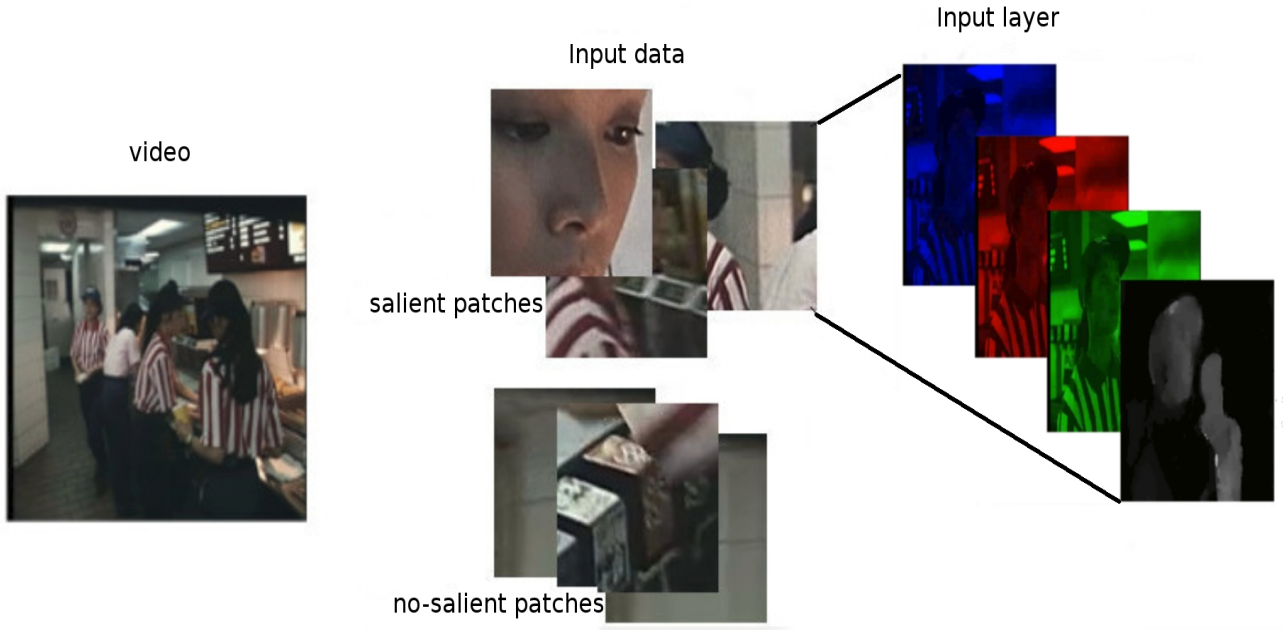
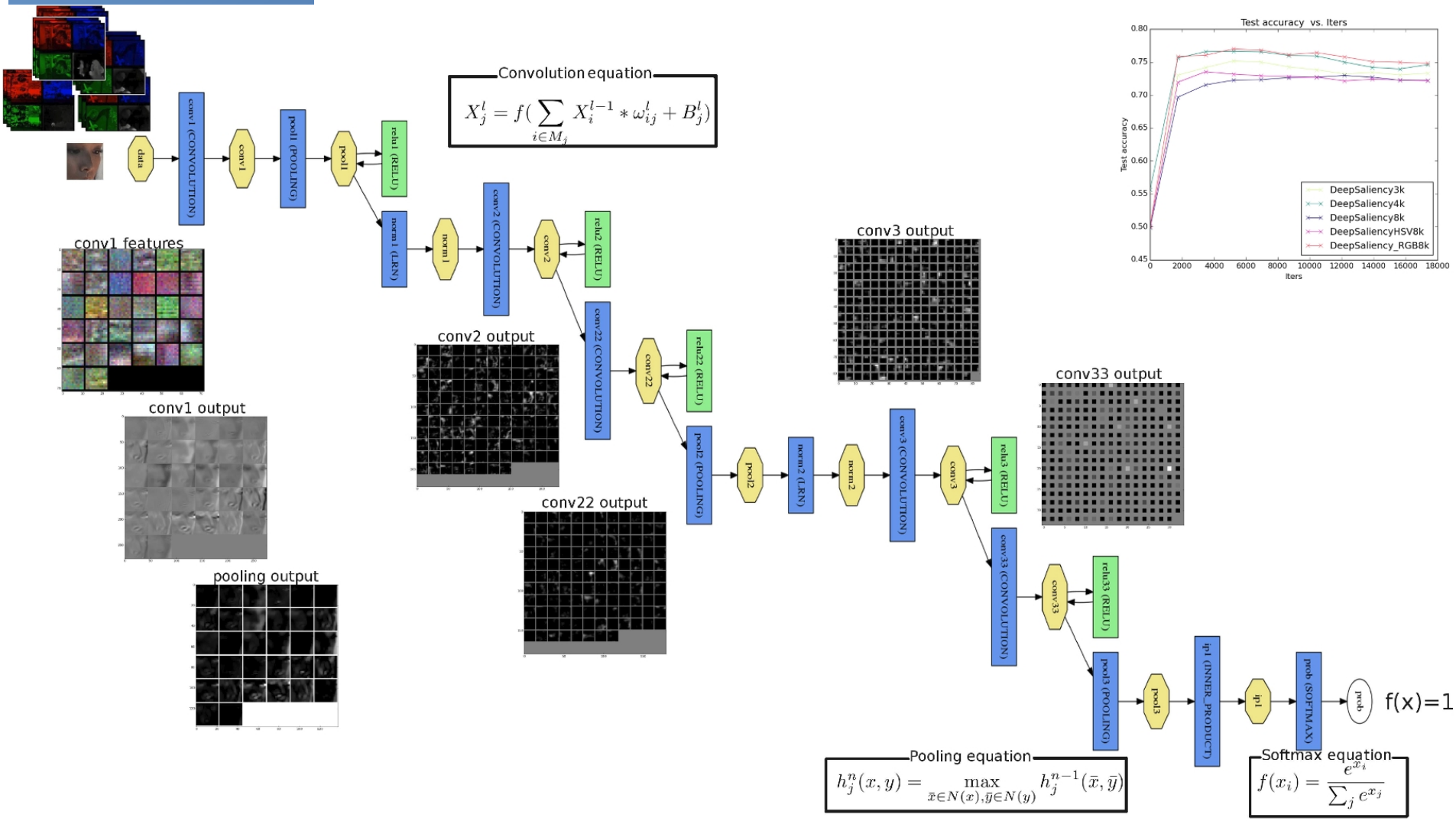


Fig 1 Input data layer : different features joined to the network.

DeepCNNs for saliency prediction

Classifier:Deep CNN

Deep CNN for two-class classification problem on patches:
-salient -non-salient



Inspired by AlexNet, Krizhevsky et al. 2012

Architecture

Three patterns in the main architecture of proposed deep CNN:

-P1: makes the operation of pooling before max-Relu accelerate the computation times.

-P2 and P3: stacking of two convolutional layers before every pool layer can develop more complex features before the destructive pooling operation.

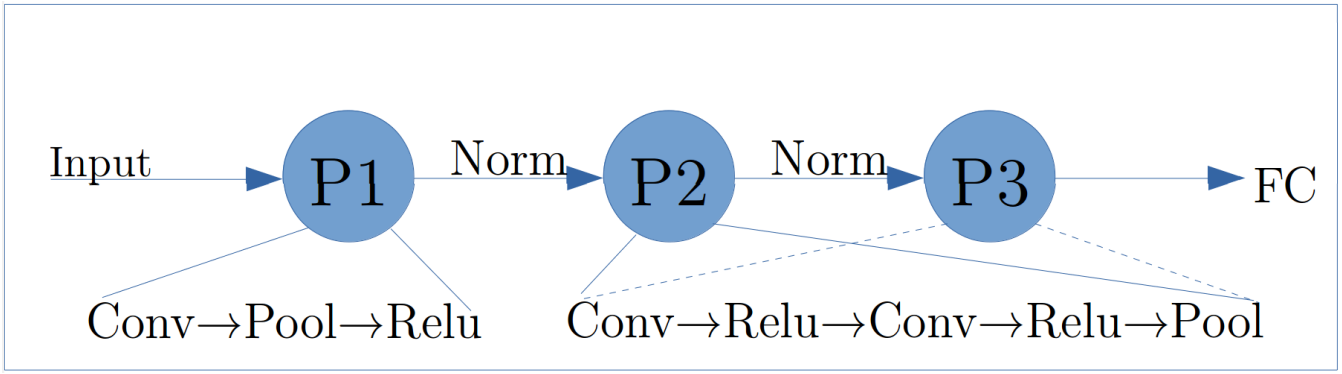


Fig 2 Architecture of video saliency convolution network 'ChaboNet'.

Sensitivity of Deep CNNs to noise

Training: selection of salient and non-salient patches(1)

SGD optimisation method is sensitive to the noise in the data!

-1. Salient Patch is selected on the basis of Wooding map
a binary label is associated with pixels X of each patch P_i: (Patch size t =100 for SD video)

$$l(X) = \begin{cases} 1 & if \quad W(x_{0,i}, y_{0,i}) \geq \tau_J \\ 0 & otherwise \end{cases}$$

with $(x_{o,i}, y_{o,i})$ the coordinates of the center of the patch.

The choice of the threshold:

$$\begin{cases} \tau_0 = \max(W(x,y), 0) \\ \tau_{(j+1)} = \tau_j - \epsilon\tau_j \end{cases}$$

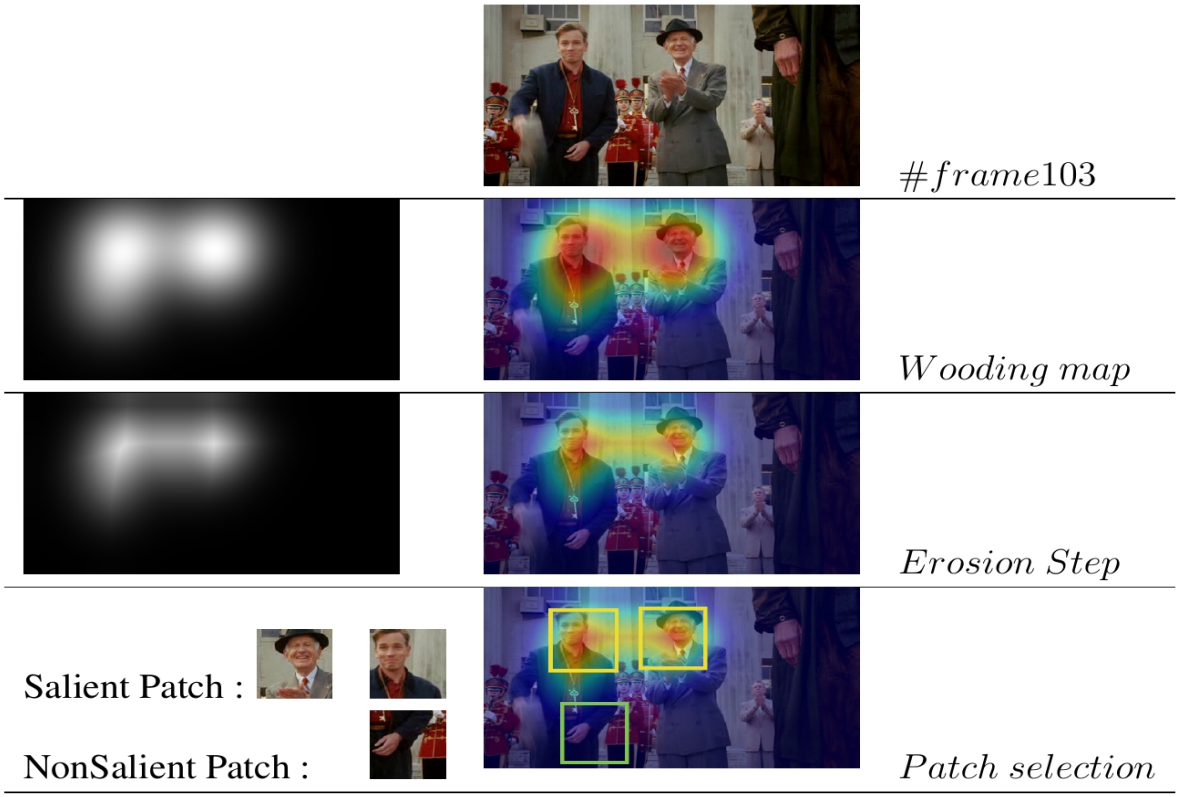


Fig 3 Policy of patch selection : example and step

Training: selection of salient and non-salient patches(2)

-2. Non-salient patch selection :

« The rule of thirds » for produced and post-produced content
let $(x_{o,i}, y_{o,i})$ *be* the coordinates of the center of the patch P_i ,

width is the width of the frame and
height is its height.
{Salient} is the set of salient positions already chosen.

$$x_{0,i} \in BorderX | x_{0,i} \notin \{Salient\}; \quad and \quad y_{0,i} \in BorderY | y_{0,i} \notin \{Salient\}$$

$$with \begin{cases} BorderX = [0, \frac{width}{5}[\cup]width - \frac{width}{5}, width] \quad \wedge \quad BorderY = [0, height] \\ or \\ BorderX = [\frac{width}{5}, width - \frac{width}{5}] \quad \wedge \quad BorderY = [0, \frac{height}{5}[\cup]height - \frac{height}{5}, height] \end{cases} \tag{3}$$

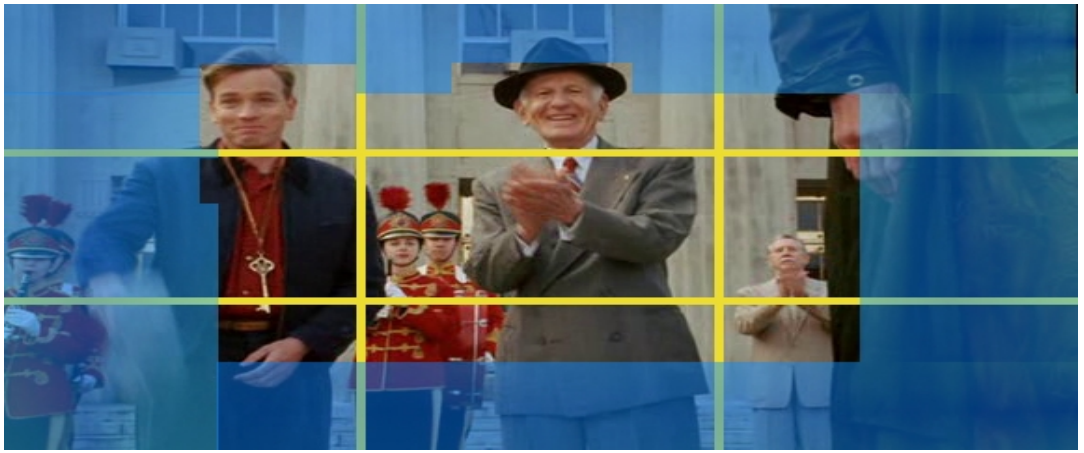


Fig 4 Space of selection of NonSalient patches.

In the first experiment : we have selected non-salient patches randomly in a standard way. This policy of selection of non-salient areas yields random errors.

In the second experiment : we used cinematographic production rule of 3/3 for non-salient patches selection, excluding the patches already defined as salient area in all the videos frames.

-using a large dataset HOLLYWOOD21* containing :
- recorded gaze fixations of up to 19 subjects for each video.
- 1707 videos ,

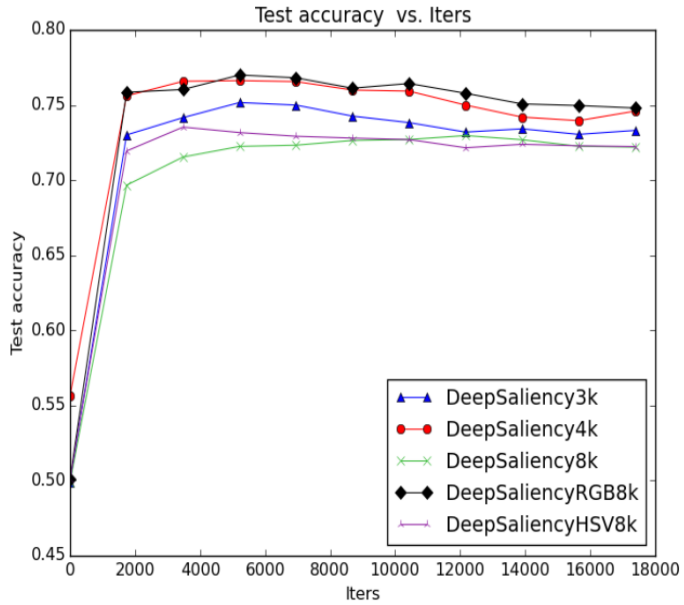
	salient	**Non-salient**
training	222863	221868
validation	251294	250169

* M. Marsza lek, I. Laptev, and C. Schmid, \Actions in Context," in IEEE Conference on Computer Vision
& Pattern Recognition , 2009.

For the HOLLYWOOD dataset,

- the interest of adding the residual motion and contrast as a new feature together with spatial colour maps.
- the essential of accuracy is obtained with purely spatial features (RGB).
- The best trained model were obtained at the iteration 5214 for the DeepSaliencyRGB8k.

Conttrastive features in average do not bring an improvement

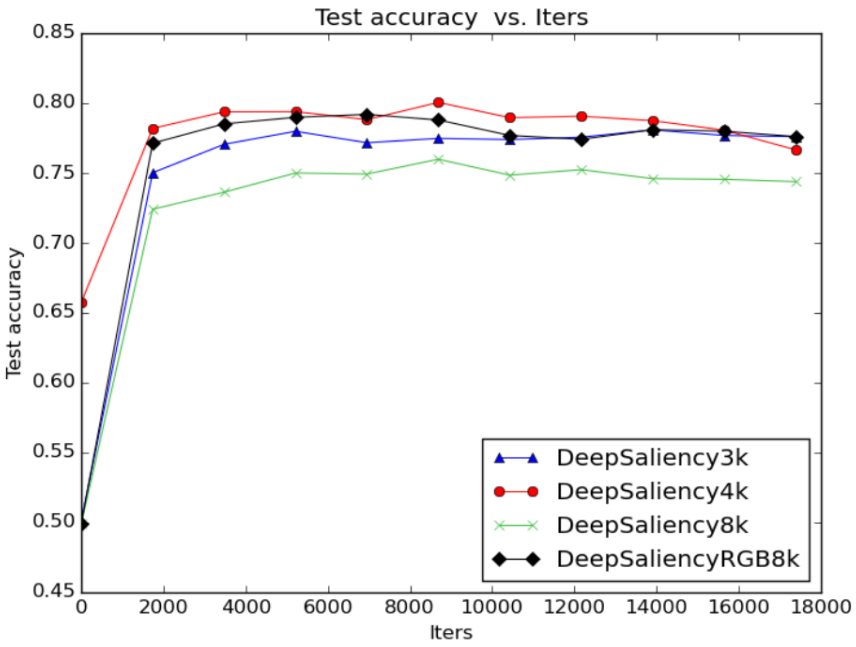


(a) Accuracy vs iterations

	3k_model	4k_model	8k_model	RGB8k_model	HSV8k_model
$min_{(\#iter)}$	49.8%(#0)	55.6%(#0)	49.8%(#0)	50.1%(#0)	50.1%(#0)
$max_{(\#iter)}$	75.1%(#5214)	76.6%(#5214)	72.9%(#12166)	76.9%(#5214)	73.5%(#3476)
$avg \pm std$	71.6% ± 0.072	73.6% ± 0.060	70.1% ± 0.067	73.5% ± 0.078	70.5% ± 0.068

(c)The accuracy results on HOLLYWOOD dataset during random selection of non-salient patches experiment

The results show the increase in accuracy in the most efficient model up to 8% .
Analyzing the results, we have noticed that purely random selection process of non-salient patches yielded errors in our training dataset for all used models.



(a) Accuracy vs iterations

	3k_model	4k_model	8k_model	RGB8k_model
$min_{(\#iter)}$	$50.11\%_{(\#0)}$	$65.73\%_{(\#0)}$	$49.88\%_{(\#0)}$	$49.92\%_{(\#0)}$
$max_{(\#iter)}$	$77.98\%_{(\#5214)}$	$80.05\%_{(\#8690)}$	$75.98\%_{(\#8690)}$	$79.19\%_{(\#6952)}$
$avg \pm std$	$77.30\% \pm 0.864$	$78.73\% \pm 0.930$	$74.55\% \pm 0.968$	$78.14\% \pm 0.703$

(c)The accuracy results on HOLLYWOOD dataset

Conclusion et perspective

- Compared the performances of prediction with Deep CNN when different kinds of features are ingested by the network: color pixel values only, color values with residual motion, color values with pre-compued contrasts.
In function of tasks - recognition of actions (dynamic content) or objects (spatial content) the saliency is better predicted with purely spatial model (RGB values) for objects and spatio-temporal features (RGB and residual motion) for actions.
- Compared to the literature we have proposed a new selection process of non-salient patches which is based on composition rules of produced content. This allowed increasing the accuracy of prediction of salient patches up to 8%.

Perspectives

- temporal continuity!

Merci pour votre attention.