IMVC, March 28th 2017

BIG DATA-SMALL DATA: IMAGE CLASSIFICATION FOR COLORECTAL CANCER DIAGNOSIS

PRESENTED BY
DORI PELEG, PHD
ALGORITHMS DIRECTOR & TECHNICAL FELLOW

Medtronic
Further, Together

INTRODUCTION TO MEDTRONIC YOKNEAM SITE AND CAPSULE ENDOSCOPY

Medtronic

Further, Together

Medtronic

MEDTRONIC YOQNEAM BY NUMBERS

300
employees
23% with advanced academic degrees

$210M
revenues

3M
capsules ingested

450
registered patents,
200 pending

16
Manufacturing lines

1,500
peer reviewed publications

- Part of a worldwide leader of medical devices (over 85000 employees, #2 in the world).

MEDTRONIC YOQNEAM

- Gastroenterologists use endoscopy to visualize the digestive system for diagnosis and treatment.
- The pioneers who brought the **game-changing innovation** that revolutionized **endoscopy**

Before

1500-2000 [mm]

After

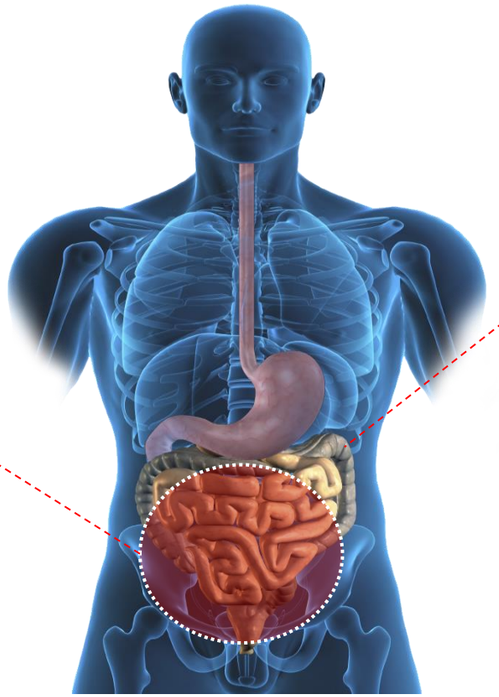
26[mm]

11[mm]

MEDTRONIC YOQNEAM PRODUCTS

PillCam® capsule endoscopy SB 2 SB

- Most widely used capsule endoscope for visualization of the entire small bowel symptoms related to bleeding, CD and IDA.



PillCam® capsule endoscopy SB 2 Colon

- Safe, minimally invasive, sedation-free, patient-friendly modality to visualize the colon.

THE CHALLENGE OF POLYP DETECTION

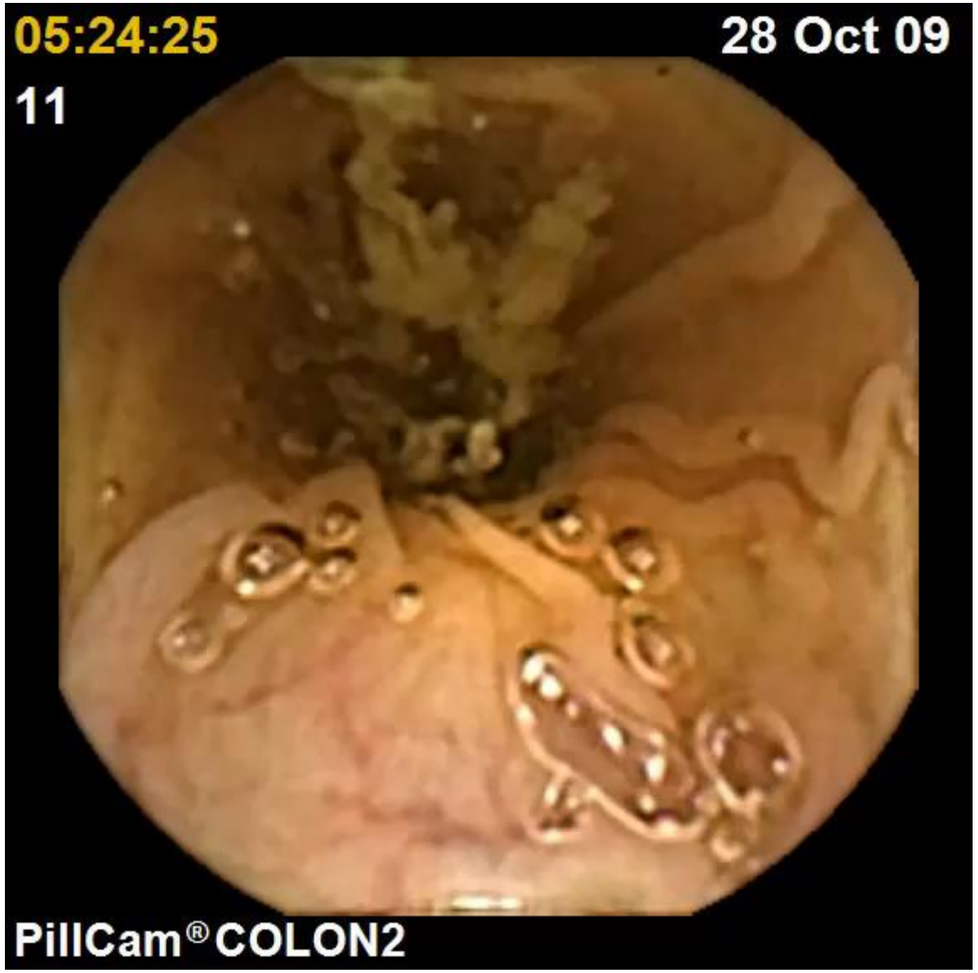
Medtronic

Further, Together

Medtronic

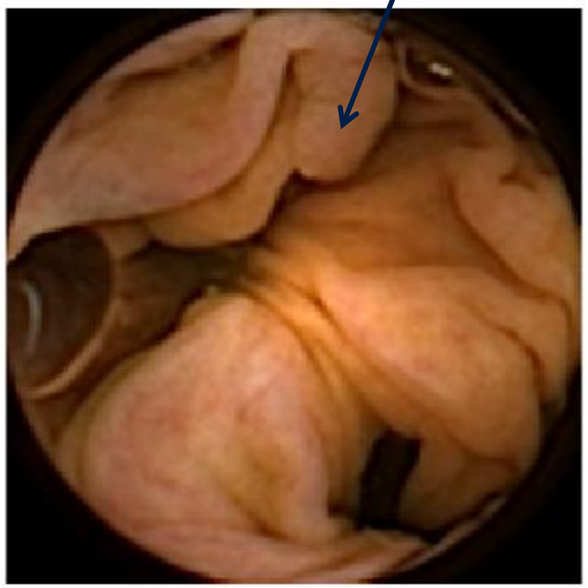
PILLCAM FOR THE COLON

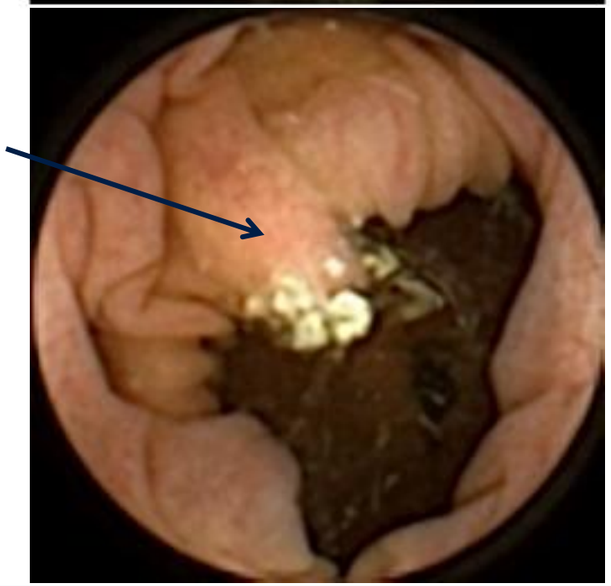
- Capsule Endoscopy for the Colon has the potential to become a screening tool for detecting colorectal cancer.
- The primary focus of the pill for the Colon is to detect polyps, which are a precursor to Colon cancer.
- If a significant polyp is detected with the capsule, the patient will be referred to Colonoscopy to have it treated.
- The main objective is image classification- is there a polyp in the image?
- It is not important to count how many polyps are in the image.

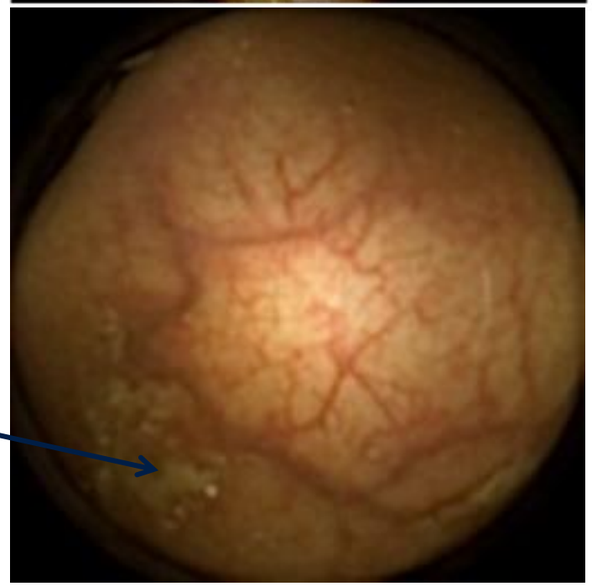


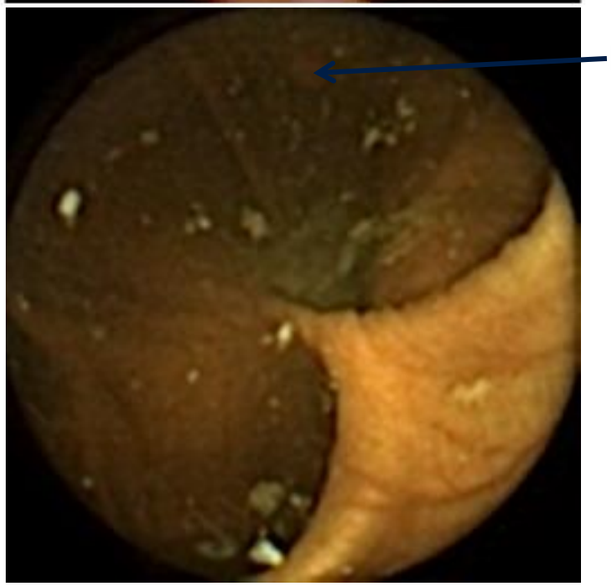
Medtronic

EXAMPLES OF POLYPS









Medtronic

OVERVIEW OF PRESENTED APPROACH

Medtronic

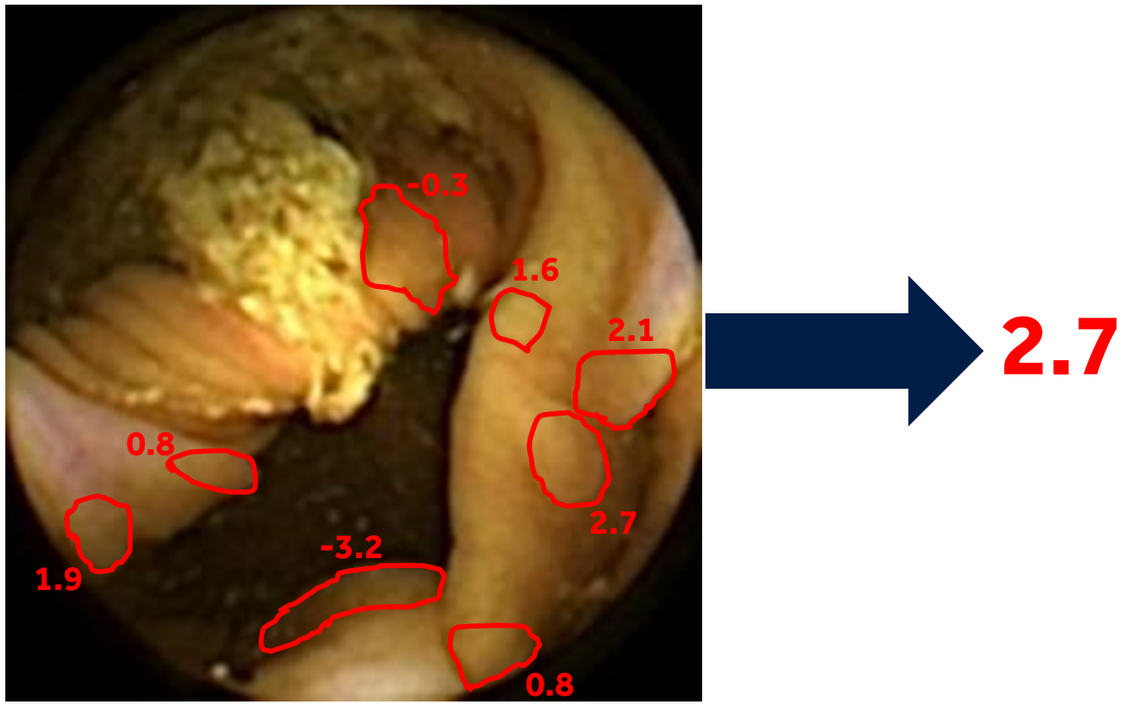
Further, Together

Medtronic

OVERVIEW OF PRESENTED APPROACH

The algorithm is composed of the following steps:

1. Segmentation-> produces about 200 candidates of polyps per image
2. Features which reflect the visual cues used by GI physician to discriminate between normal tissue and polyps.
3. Classifier-> produces a soft-margin which indicates how sure we are the candidate is a polyp.
4. Score per image-> The probability the image contains a polyp is based on maximal soft-margin of the candidates of the image.



OVERVIEW OF PRESENTED APPROACH

- Big data, small data: The database consists of ~1e8 non-polyp candidates and ~100 polyp candidates.

- Why isn't this talk about deep learning?
  - Very few positive examples: It is expensive (millions of dollars) and time consuming (years) to collect images of several hundreds of polyps.
  - While on **average** deep learning has provided superior results on many visual recognition benchmarks, deep learning is susceptible to embarrassing mistakes:
    - Missing obvious polyps may be fatal
    - Suggesting images which clearly don't have a polyp significantly reduce the physician's confidence in the system.

ALGORITHM-CLASSIFIER

Medtronic

Further, Together

Medtronic

CANDIDATE LEVEL VS. IMAGE LEVEL

- From each image, we have ~200 candidates.
- The straightforward way is to collect several thousand examples and train a classical classifier such as the SVM algorithm.
- This is optimization at the candidate level. The only level that is important is the image level.
- Often improving on the candidate level, deteriorates the performance on the image level.

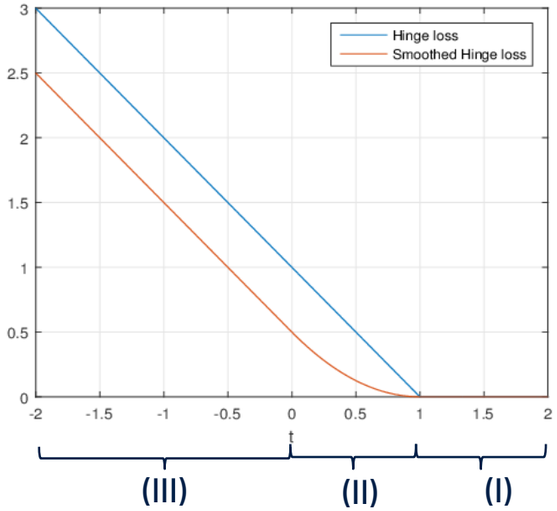
- Starting point: standard linear SVM with classifier $f(x) = w^T x + b$ and label $y \in \{-1,1\}$.
- Minimize:
  - The quadratic penalty on the norm of w: $w^T w$.
  - The hinge loss on the soft-margins multiplied by the labels: $h(t) = max\{1 - t, 0\}$, where $t = yf(x)$
- Due to the dual optimization problem⇒ The number of variables depends on the number of examples.

MODIFICATIONS FROM SVM
LINEAR CLASSIFIER

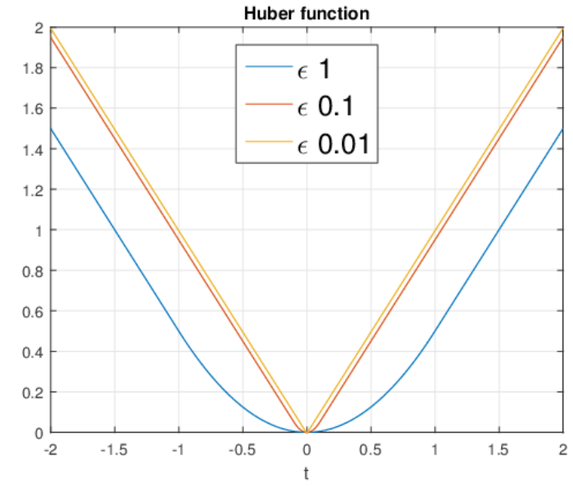
1. Balancing positive and negative:
   - Provide an equal weight for positive and negative examples without regard to the number of examples available in the training set.
2. Smoothed the hinge-loss function:

$$L_\delta(t) = \begin{cases} 0 & t \geq 1 \\ \frac{(1-t)^2}{4\delta} & 1-2\delta \leq t < 1 \\ 1-t-\delta & t < 1-2\delta \end{cases}$$

⟵ (I) No training error. Outside margin.
⟵ (II) No training error. Inside margin.
⟵ (III) Training error.



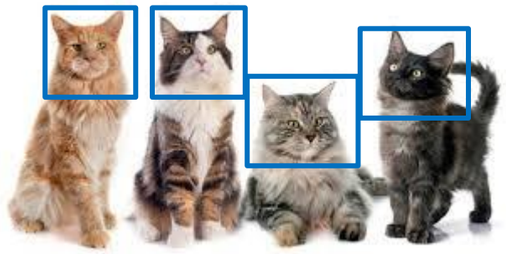
   - This is twice differentiable function.
3. Huber penalty for w:
   - Insensitive to outliers.
   - Can parametrically adjust between imposing sparsity on the features ($\varepsilon \to 0$) or utilizing all of them ($\varepsilon \to \infty$) as in the SVM.
4. Grouping examples:
   - Minimize the **maximal** (smoothed) hinge-loss error for each group. Any error which is not maximal, is unimportant.



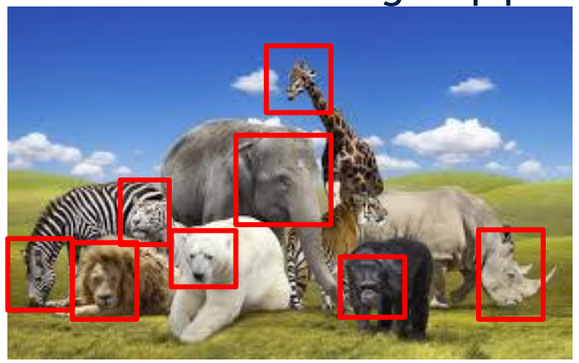
Medtronic

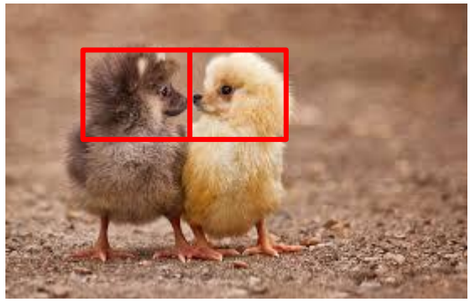
USE CASE
CAT DETECTOR- IS THERE A CAT IN AN IMAGE?

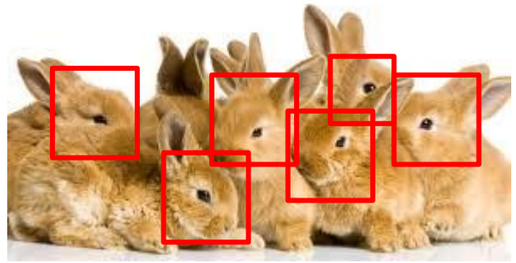
Positive examples-
Use the best candidate for the object



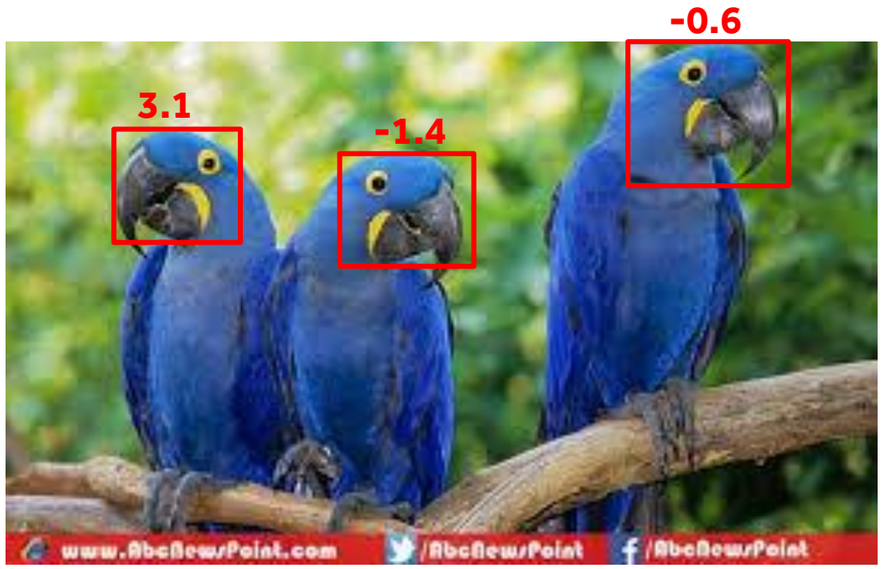
Negative examples-
Use all the candidates and group per image

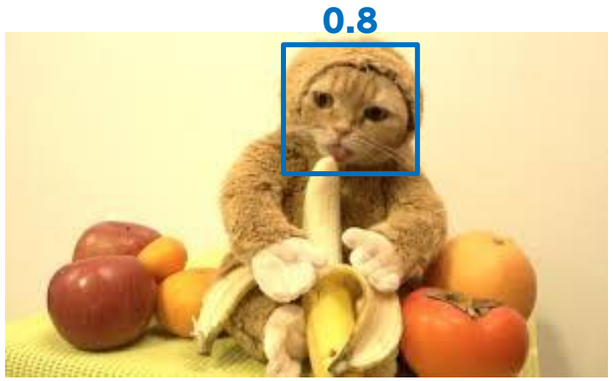






USE CASE
WHY GROUP NEGATIVE EXAMPLES?

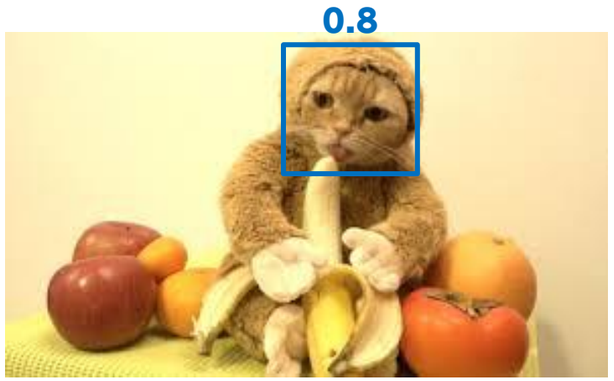


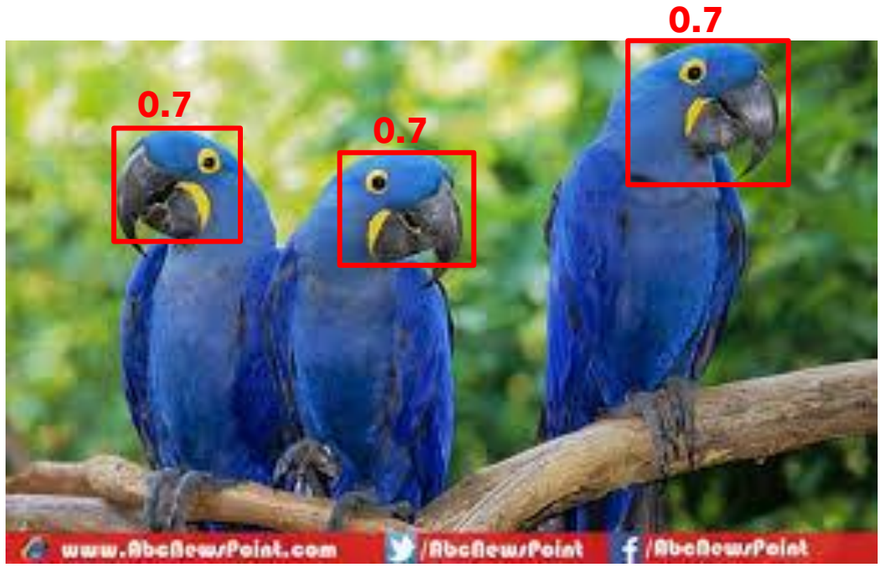


Soft-margins (positive)	2.2	0.8
Hinge loss	0	0.2

Soft-margins (negative)	3.1	-1.4	-0.6	Mean hinge loss	Max hinge loss
Hinge loss	4.1	0	0.4	1.5	4.1

USE CASE
WHY GROUP NEGATIVE EXAMPLES?





Soft-margins (positive)	2.2	0.8
Hinge loss	0	0.2

Soft-margins (negative)	0.7	0.7	0.7	Mean hinge loss	Max hinge loss
Hinge loss	1.7	1.7	1.7	1.7	1.7

WHY GROUPING ONLY NEGATIVE CLASS
ILLUSTRATIVE EXAMPLE

Focuses on increasing all the soft-margin equally

Focuses on increasing the lowest soft-margin

Focuses on increasing the largest soft-margin

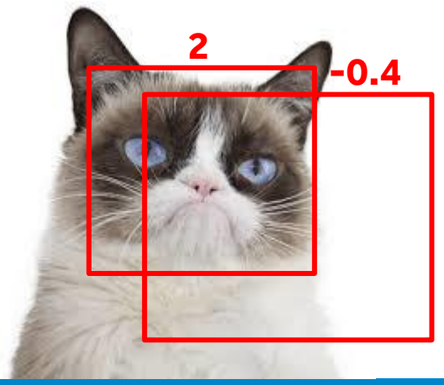
Positive class

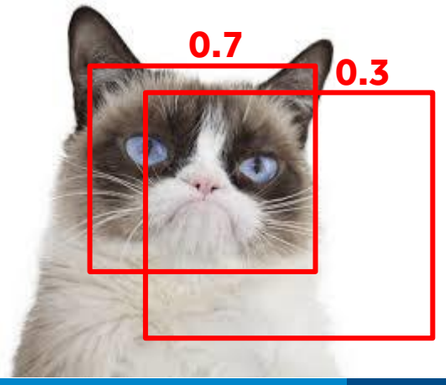
Soft-margins (positive)	2	-0.4	Min hinge loss	Mean hinge loss	Max hinge loss
Hinge loss	0	1.4	**0**	0.7	1.4

Non-convex

Soft-margins (positive)	0.7	0.3	Min hinge loss	Mean hinge loss	Max hinge loss
Hinge loss	0.3	0.7	0.3	**0.5**	**0.7**

The best solution is simply to choose a single candidate for the positive class instead of a few.





LETTING GO OF THE KERNEL TRICK

- If we don't use the kernel trick, the number of variables is the number of features and the matrix to store is of dimensions $[d \times n]$.
- However, this leaves us with a linear classifier, which typically is not versatile enough to provide the best performance.
- "Polynomial" trick: Transform the input features to all the polynomial factors up to a certain degree. The classifier will be linear on its input and polynomial in the original space.
- For example, consider two features x_1, x_2, with a maximal degree of 3. The new features will be $\left\{x_1, x_2, x_1x_2, {x_1}^2, {x_2}^2, {x_1}^2x_2, {x_2}^2x_1, {x_1}^3, {x_2}^3\right\}$ and instead of 2 weights for a linear classifier there will be 9.
- Any non-linear transformation of the features will retain the convexity of the optimization problem.

RESULTS

Medtronic

Further, Together

Medtronic

RESULTS
LINEAR CLASSIFIER EXPERIMENT

- In order to analyze if the proposed classifier is better than the SVM and why, the following experiment was prepared:
  - Both classifiers were a linear model.
  - The training set was divided into two subgroups: “Training” and “Test”.

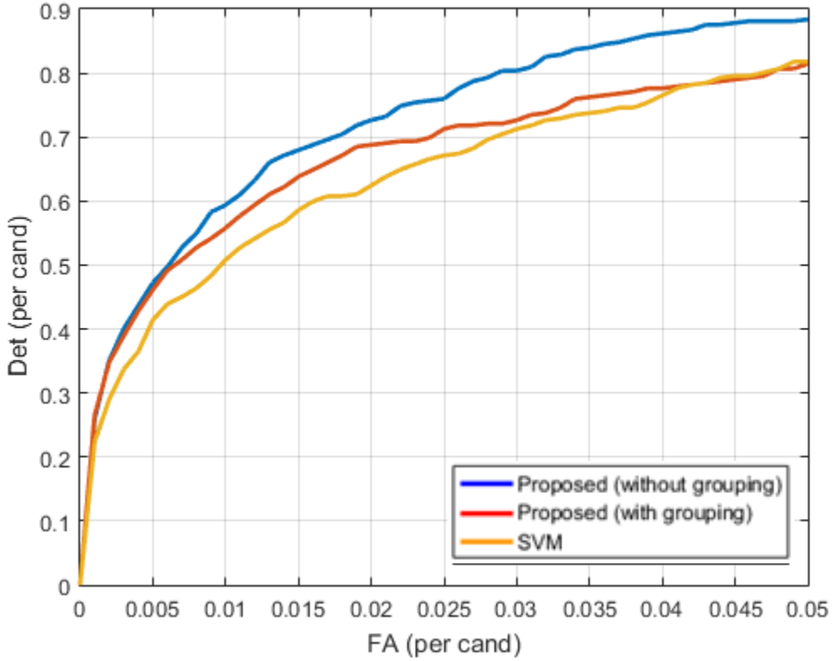
	Training		Test	
	Positive	Negative	Positive	Negative
SVM	293 (100 Polyps)	5000	362 (115 Polyps)	326M
Proposed Classifier	293 (100 Polyps)	100M	362 (115 Polyps)	326M

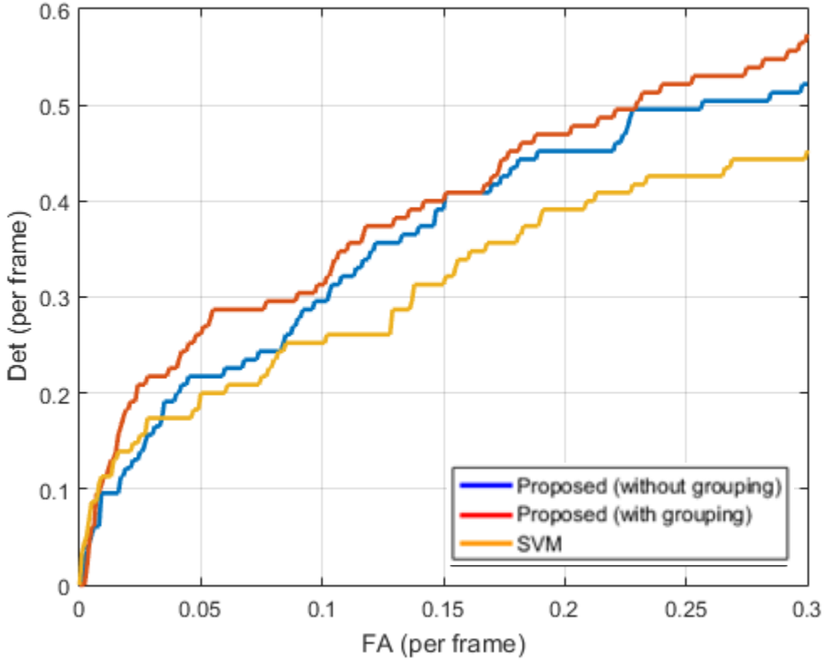
  - Note that the training set size for the Big Data classifier is 20000 times larger than the maximal for the SVM.

PERFORMANCE OF PER CANDIDATE VS. PER FRAME OPTIMIZATION

Candidate level: best to use the proposed classifier *without* grouping.

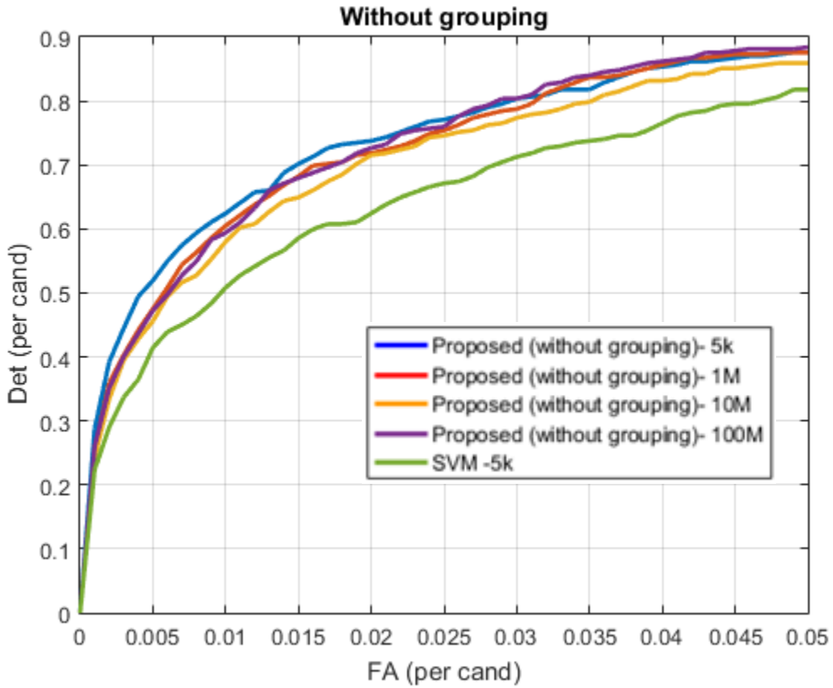
Frame level: best to use the proposed classifier over the SVM thanks to its ability to handle big data (*without* grouping) and thanks to its ability to group examples.

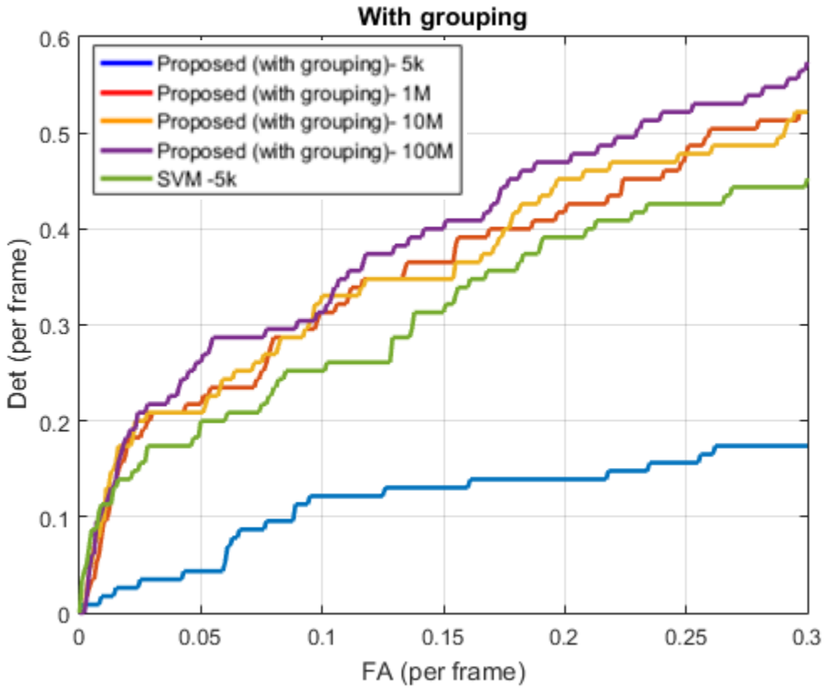




DEPENDENCY ON TRAINING SET SIZE

- For the classification task without grouping, there was no major impact due to the “big data” aspect. Increasing the training set did not improve performance.
- However, the proposed classifier outperformed the SVM classifier with exactly the same training set and without using cross validation. This means that the objective function with the different penalty terms on the generalization and training errors leads to better performance in itself.
- For the classification task with grouping, the big data plays an important role.





SUMMARY

Medtronic

Further, Together

Medtronic

FUTURE WORK

- It can be used for grouping other sets of examples: sequences of frames, non-image applications….
- It can also be used for transfer learning with a similar dataset but when the desired data doesn’t have enough (positive) examples to fine-tune the last few layers.
- In future work we will try to design non-linear transforms on the data in order to perform end-to-end learning.

- Contact me for questions: dori.peleg@medtronic.com