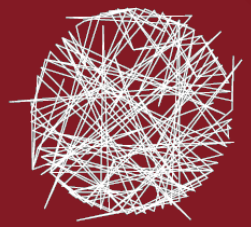
Institute for Systems and Robotics | LISBOA

A Deep Attention Model for the Hiearchical Diagnosis of Skin Lesions

C. Barata, M. E. Celebi, and J. S. Marques



LARSyS
Laboratory of Robotics and Engineering Systems

TÉCNICO LISBOA

Motivation

- Deep neural networks became the state-of-the-art

Motivation

- Deep neural networks became the state-of-the-art
- These methods rival the performance of dermatologists

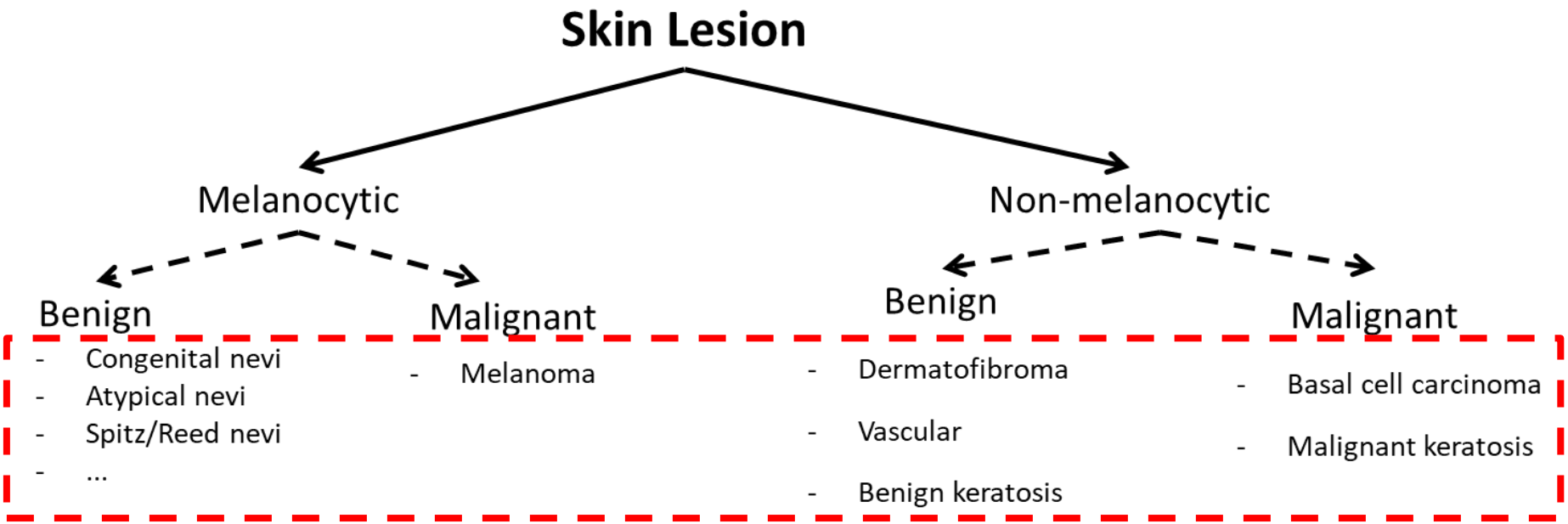
Motivation

- Deep neural networks became the state-of-the-art
- These methods rival the performance of dermatologists
- **But they lack interpretability and transparency!**

Motivation

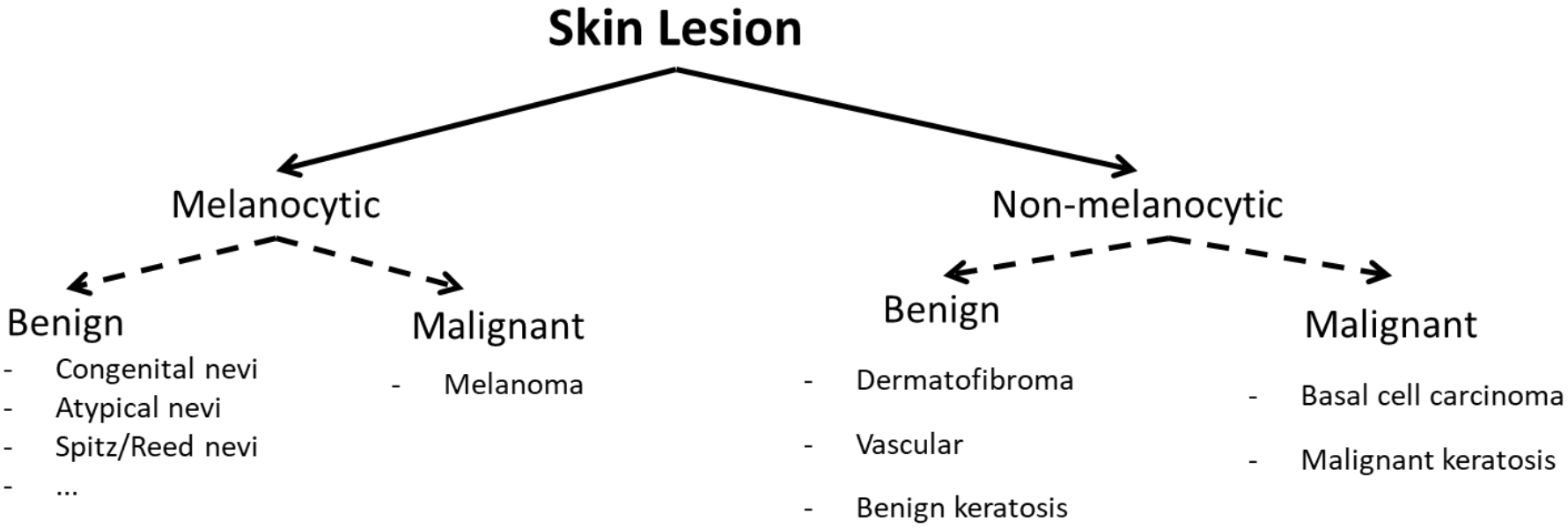
- Deep neural networks became the state-of-the-art
- These methods rival the performance of dermatologists
- **But they lack interpretability and transparency!**
- **How to incorporate medical knowledge in DNNs?**

Changing The Perspective



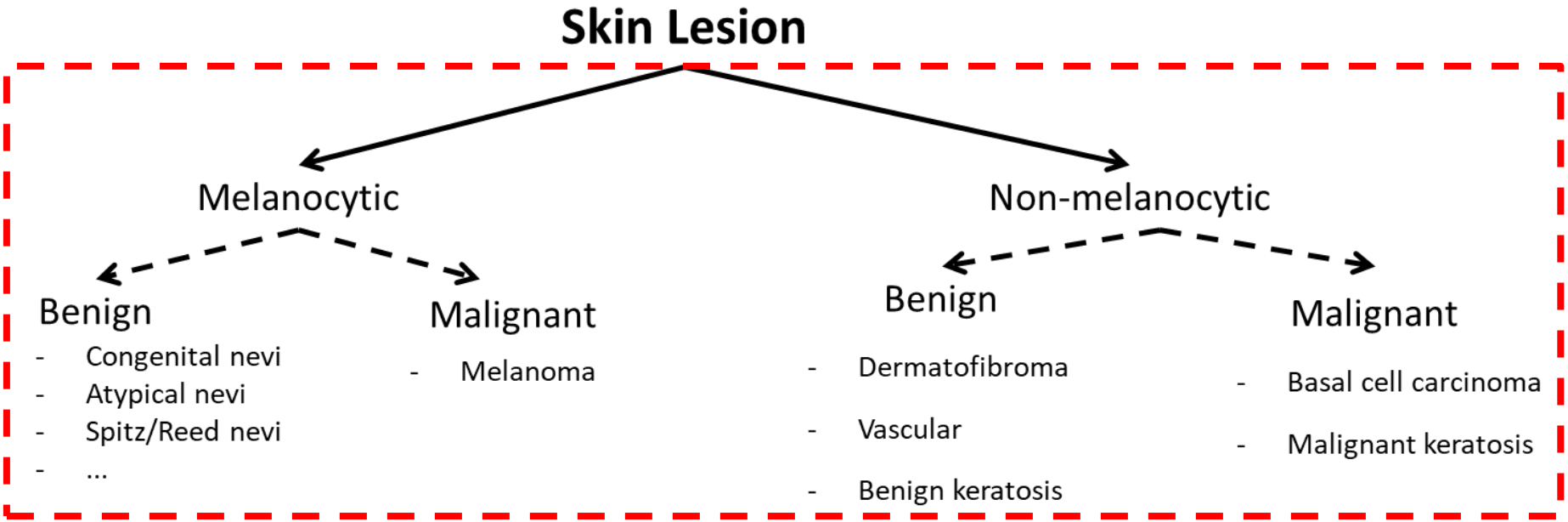
- The diagnosis of skin cancer is usually perceived as a multi-class problem.

Changing The Perspective



- The diagnosis of skin cancer is usually perceived as a multi-class problem.
- But...maybe we can look at it from a different perspective!

Changing The Perspective



- The diagnosis of skin cancer is usually perceived as a multi-class problem.
- But...we can look at it differently!
- Why not explore the hierarchical organization of the lesions?

Hierarchical Classification

- Predict the **sequence of classes C** that **better describes** the dermoscopy image I and **maximizes**

$$log\ p(C|I) = \sum_{t=0}^{T} log\ p(C_t|I, C_0, \dots, C_{t-1})$$

Image Captioning

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

Karpathy, Fei Fei, CVPR '15

- **Definition:** automatic generation of image descpritions.
- **Requirements:** recognize objects, their attributes, and relationships in the image.

Hierarchical Classification

- Predict the **sequence of classes C** that **better describes** the dermoscopy image I and **maximizes**

$$log\ p(C|I) = \sum_{t=0}^{T} log\ p(C_t|I, C_0, \dots, C_{t-1})$$

- **Image Caption:** given an image I, we want to predict the sequence of words $S = \{S_0, S_1, \dots, S_T\}$ that maximizes

$$log\ p(S|I) = \sum_{t=0}^{T} log\ p(S_t|I, S_0, \dots, S_{t-1})$$

Hierarchical Classification

- Predict the **sequence of classes C** that **better describes** the dermoscopy image I and **maximizes**

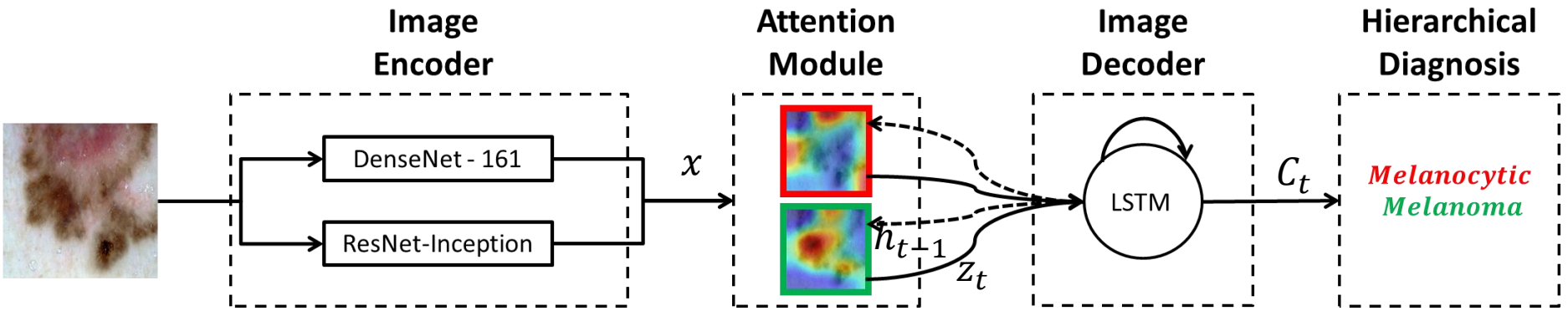
$$log\ p(C|I) = \sum_{t=0}^{T} log\ p(C_t|I, C_0, \dots, C_{t-1})$$

- **Image Caption:** given an image I, we want to predict the sequence of words $S = \{S_0, S_1, \dots, S_T\}$ that maximizes

$$log\ p(S|I) = \sum_{t=0}^{T} log\ p(S_t|I, S_0, \dots, S_{t-1})$$

- **It is the same formulation!**

Proposed Solution



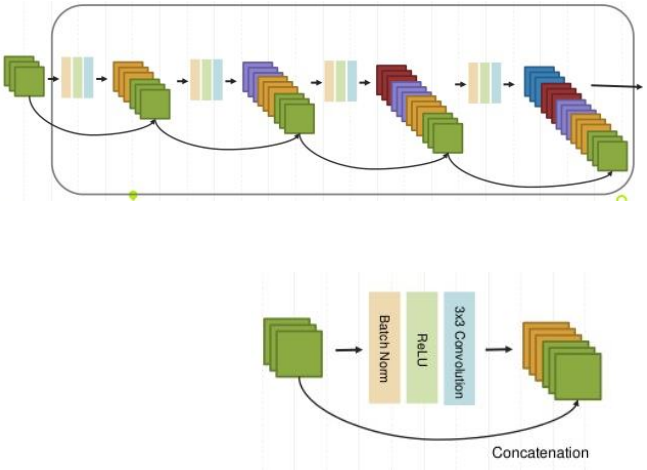
- The model uses a **encoder-decoder framework** to sequentially generate the hiearchical classes.

- An **attention module** is incorporated to provided **interpretability**.

Model Specifications

Image Encoder – Pretrained on ImageNet

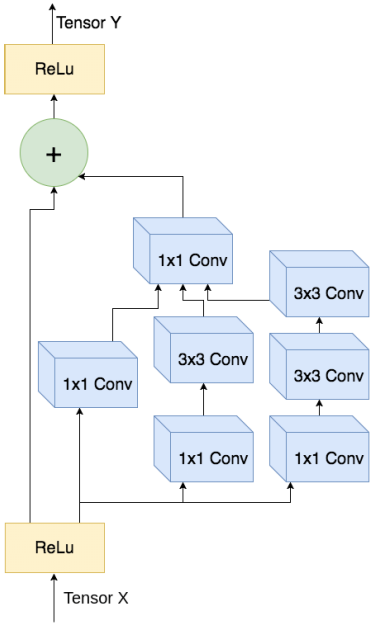
DenseNet-161



Output: 9x9x2208

+

ResNet-Inception

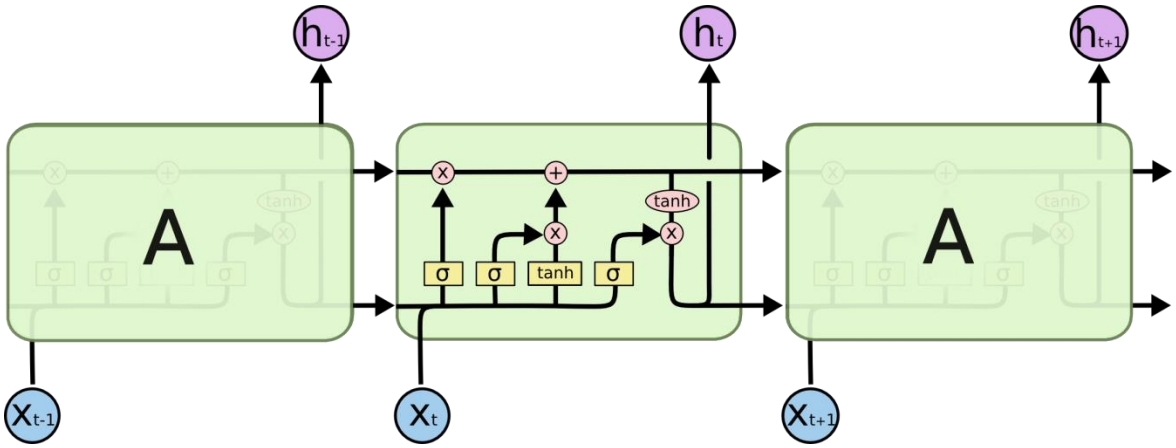


Output: 9x9x1536

Model Specifications

Language Decoder – Hierarchical Classification

Long-short term memory



$$X_t = [z_t, C_{t-1}E]$$

Words are embedded

Comes from attention module

Model Specifications

Attention Module

- Identifies regions of interest for a certain class

1. Computes a weight for each location $\alpha_t = \{\alpha_{t1}, \dots, \alpha_{t81}\}$

$$\alpha_t = softmax(\boldsymbol{W}_a(\tanh(\boldsymbol{W}_{ax}x + \boldsymbol{W}_{ah}h_{t-1}))$$

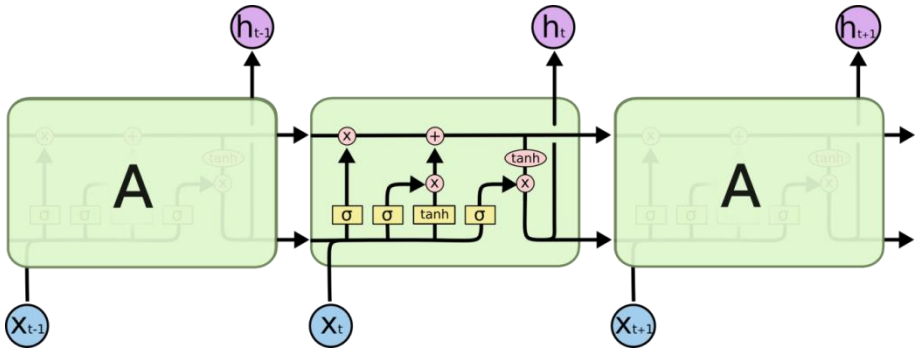
2. Determines the context input z_t **(soft attention)**

$$z_t = \sum_{i=1}^{81} a_{ti}x_i$$

Model Specifications

Language Decoder – Hierarchical Classification

Long-short term memory



$$X_t = [z_t, C_{t-1}E]$$

Words are embedded

Comes from attention module

Hierarchical Class Inference

$$p(C_t|I, C_{t-1}) = softmax(W_o(C_{t-1}E + W_z z_t + W_h h_t))$$

Experimental Setup

- The experiments were performed on two datasets
  - ISIC 2017: **non-melanocytic** (1 type) and **melanocytic** (2 types)
  - ISIC 2018: **non-melanocytic** (5 types) and **melanocytic** (2 types)
- All of the models were optimized using Adam with an adaptive learning rate ($\eta = 10^{-6}$ to start) – cross entropy loss.
- The system is evaluated using:
  - Sensitivity and Specificity
  - Area under the curv (AUC)
  - Balanced accuracy (BACC)

ISIC 2017 & 2018 Scores

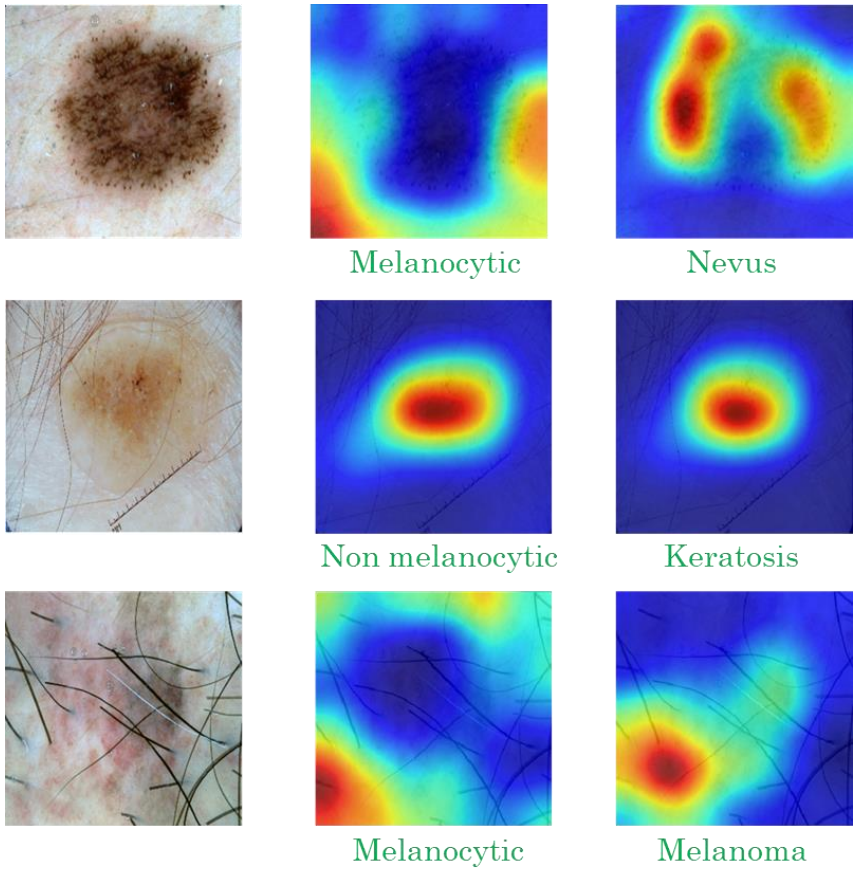
ISIC 2017

Table 1. Best performance scores for the ISIC 2017 test set, using full images.

Lesion Class	*SE*	*SP*	*BACC*	*AUC*
Melanocytic/Non-Melanocytic (#510/#90)	92.5%	70.0%	81.3%	91.9%
Keratosis (#90)	67.8%	92.1%	-	91.2%
Melanoma (#117)	65.8%	88.6%	-	85.9%
Nevus (#393)	82.2%	78.7%	-	86.5%
Average (#600)	71.9%	86.5%	71.9%	87.9%

Training Set – 2000 images
Val. Set – 150 images
Test Set – 600 images

Interpretability Examples – ISIC 2017



ISIC 2017 & 2018 Scores

ISIC 2017

Table 1. Best performance scores for the ISIC 2017 test set, using full images.

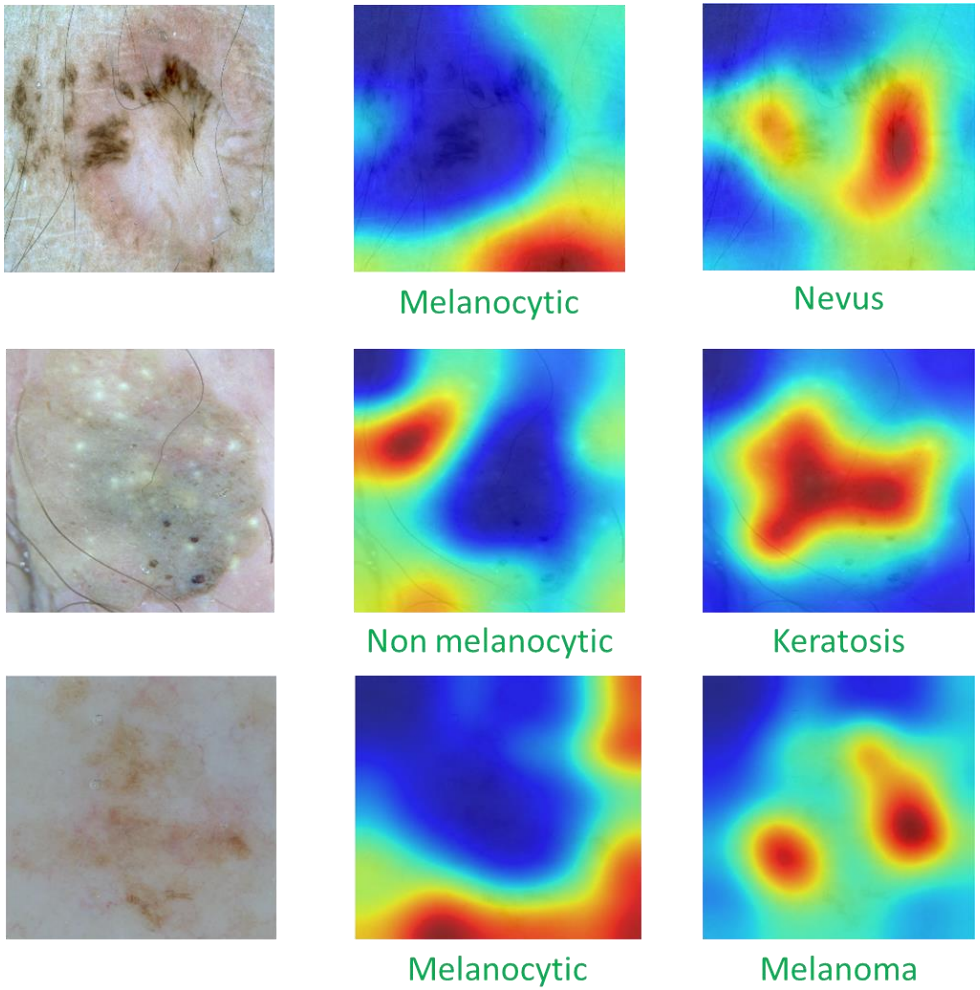
Lesion Class	*SE*	*SP*	*BACC*	*AUC*
Melanocytic/Non-Melanocytic (#510/#90)	92.5%	70.0%	81.3%	91.9%
Keratosis (#90)	67.8%	92.1%	-	91.2%
Melanoma (#117)	65.8%	88.6%	-	85.9%
Nevus (#393)	82.2%	78.7%	-	86.5%
Average (#600)	71.9%	86.5%	71.9%	87.9%

Table 2. Best performance scores for the ISIC 2017 test set, using cropped images.

Lesion Class	*SE*	*SP*	*BACC*	*AUC*
Melanocytic/Non-Melanocytic (#510/#90)	97.2%	61.1%	79.2%	93.8%
Keratosis (#90)	61.1%	97.2%	-	93.2%
Melanoma (#117)	73.5%	83.8%	-	85.5%
Nevus (#393)	82.4%	79.2%	-	88.4%
Average (#600)	72.3%	86.7%	72.3%	89.0%

Training Set – 2000 images
Val. Set – 150 images
Test Set – 600 images

Interpretability – ISIC 2017



State-of-the-art

Where do we stand?

Table 3. Comparison with other works on the ISIC 2017 test set. $*$ means that some information is missing from the paper.

			Melanoma			Keratosis			Average		
Method	**Ensembles**	**Ext. Data**	*SE*	*SP*	*AUC*	*SE*	*SP*	*AUC*	*SE*	*SP*	*AUC*
#1 [19]	Y	Y	73.5%	85.1%	86.8%	97.8%	77.3%	95.3%	85.7%	81.3%	91.1%
#2 [13]	N	Y	10.3%	99.8%	85.6%	17.8%	99.8%	96.5%	14.1%	99.8%	91.0%
#3 [21]	Y	Y	54.7%	95.0%	87.4%	35.6%	99.0%	94.3%	34.4%	97.4%	90.8%
#4 [5]	Y	Y	42.7%	96.3%	87.0%	58.9%	97.6%	92.1%	50.8%	97.0%	89.6%
Proposed Cropped	Y	N	73.5%	83.8%	85.5%	61.1%	97.2%	93.2%	67.3%	90.5%	89.4%
#5 [10]	Y	Y	35.0%	96.5%	83.6%	55.6%	97.6%	93.5%	45.3%	97.1%	88.6%
Proposed Full	Y	N	65.8%	88.6%	85.9%	67.8%	92.1%	91.2%	66.8%	90.3%	88.6%
[33]	N	N	60.7%	88.4%	84.2%	*	*	*	*	*	*
[14]	Y	N	40.2%	71.9%	85.1%	71.1%	85.1%	93.0%	55.6%	78.5%	89.1%
[34]	N	Y	65.8%	89.6%	87.5%	87.8%	86.7%	95.8%	76.8%	88.2%	91.7%

State-of-the-art

Where do we stand?

Table 3. Comparison with other works on the ISIC 2017 test set. ∗ means that some information is missing from the paper.

			Melanoma			Keratosis			Average		
Method	**Ensembles**	**Ext. Data**	*SE*	*SP*	*AUC*	*SE*	*SP*	*AUC*	*SE*	*SP*	*AUC*
#1 [19]	Y	Y	73.5%	85.1%	86.8%	97.8%	77.3%	95.3%	85.7%	81.3%	91.1%
#2 [13]	N	Y	10.3%	99.8%	85.6%	17.8%	99.8%	96.5%	14.1%	99.8%	91.0%
#3 [21]	Y	Y	54.7%	95.0%	87.4%	35.6%	99.0%	94.3%	34.4%	97.4%	90.8%
#4 [5]	Y	Y	42.7%	96.3%	87.0%	58.9%	97.6%	92.1%	50.8%	97.0%	89.6%
Proposed Cropped	Y	N	73.5%	83.8%	85.5%	61.1%	97.2%	93.2%	67.3%	90.5%	89.4%
#5 [10]	Y	Y	35.0%	96.5%	83.6%	55.6%	97.6%	93.5%	45.3%	97.1%	88.6%
Proposed Full	Y	N	65.8%	88.6%	85.9%	67.8%	92.1%	91.2%	66.8%	90.3%	88.6%
[33]	N	N	60.7%	88.4%	84.2%	*	*	*	*	*	*
[14]	Y	N	40.2%	71.9%	85.1%	71.1%	85.1%	93.0%	55.6%	78.5%	89.1%
[34]	N	Y	65.8%	89.6%	87.5%	87.8%	86.7%	95.8%	76.8%	88.2%	91.7%

ISIC 2017 & 2018 Scores

ISIC 2017

Table 1. Best performance scores for the ISIC 2017 test set, using full images.

Lesion Class	*SE*	*SP*	*BACC*	*AUC*
Melanocytic/Non-Melanocytic (#510/#90)	92.5%	70.0%	81.3%	91.9%
Keratosis (#90)	67.8%	92.1%	-	91.2%
Melanoma (#117)	65.8%	88.6%	-	85.9%
Nevus (#393)	82.2%	78.7%	-	86.5%
Average (#600)	71.9%	86.5%	71.9%	87.9%

Table 2. Best performance scores for the ISIC 2017 test set, using cropped images.

Lesion Class	*SE*	*SP*	*BACC*	*AUC*
Melanocytic/Non-Melanocytic (#510/#90)	97.2%	61.1%	79.2%	93.8%
Keratosis (#90)	61.1%	97.2%	-	93.2%
Melanoma (#117)	73.5%	83.8%	-	85.5%
Nevus (#393)	82.4%	79.2%	-	88.4%
Average (#600)	72.3%	86.7%	72.3%	89.0%

Training Set – 2000 images
Val. Set – 150 images
Test Set – 600 images

ISIC 2018

Table 4. Best performance scores for our validation set randomly selected from ISIC 2018.

Lesion Class	*SE*	*SP*	*BACC*	*AUC*
Melanocytic/Non-Melanocytic (#1564/#439)	93.7%	90.7%	92.2%	97.6%
Melanoma (#222)	75.7%	92.0%	-	93.6%
Nevus (#1342)	87.4%	94.7%	-	97.2%
Actinic (#65)	61.5%	99.4%	-	80.4%
BCC (#102)	84.3%	98.9%	-	82.0%
Keratosis (#220)	81.4%	94.9%	-	83.7%
Dermatofibroma (#24)	66.7%	99.5%	-	60.5%
Vascular (#28)	89.2%	99.6%	-	64.4%
Average (#2003)	78.0%	97.0%	78.0%	80.2%

Table 5. Best performance scores on the test set of ISIC 2018.

Lesion Class	*SE*	*SP*	*BACC*	*AUC*
Melanoma	60.8%	90.9%	-	88.1%
Nevus	84.6%	90.5%	-	94.9%
Actinic	44.2%	99.0%	-	94.4%
BCC	60.2%	98.4%	-	96.6%
Keratosis	70.0%	91.7%	-	91.0%
Dermatofibroma	65.9%	99.4%	-	94.7%
Vascular	60.0%	99.5%	-	97.3%
Average (#1512)	63.7%	95.6%	64.1%	93.9%

Training Set ~ 10,000 images
Test Set ~ 1500 images

Conclusions and Future Work

- This work proposes a diagnosis system inspired by medical knowledge
- The model uses attention maps to improve explainability
- The preliminary results are promising but it is necessary to improve the performance of the ISIC 2018 dataset

THANK YOU FOR YOUR ATTENTION!

QUESTIONS?