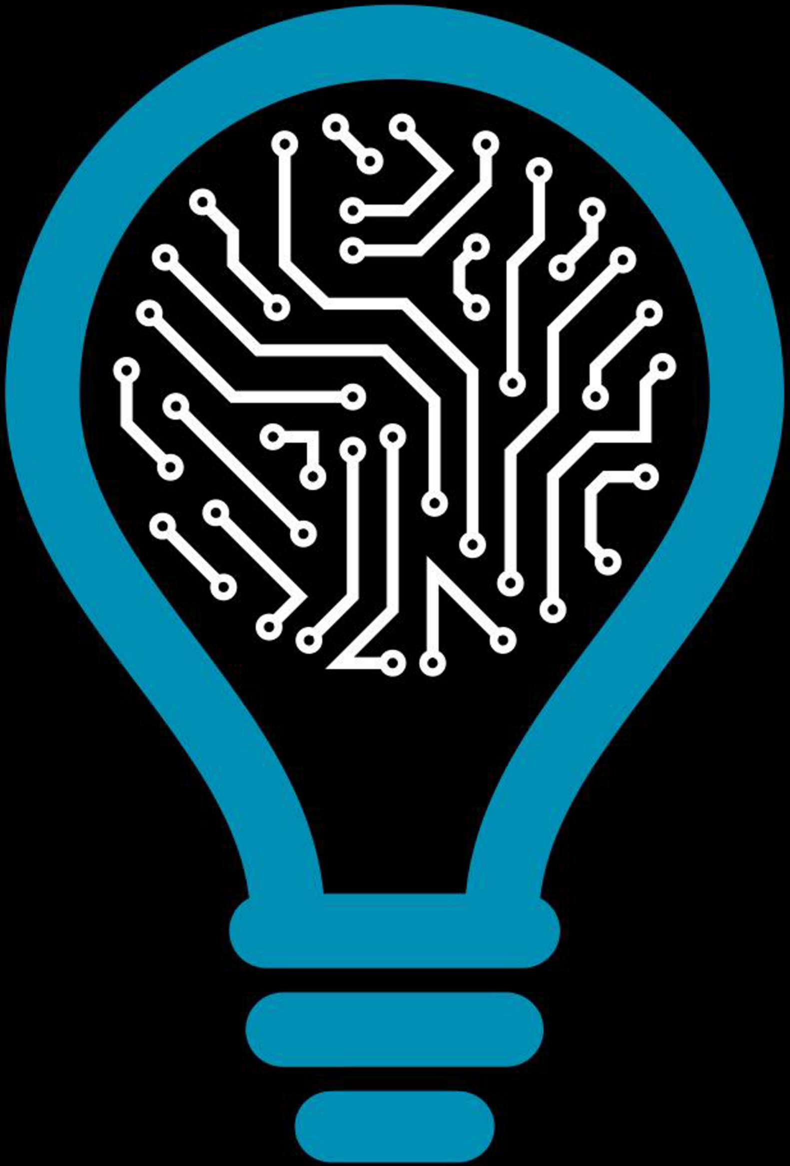
Attention Mechanisms for Medical Applications: Claims, Potentialities and Future Challenges

Deep Learning Sessions Portugal

March 15, 2023 | Startup Lisboa (Lisboa)

Tiago Filipe Sousa Gonçalves

tiago.f.goncalves@inesctec.pt



INESCTEC

INSTITUTE FOR SYSTEMS
AND COMPUTER ENGINEERING,
TECHNOLOGY AND SCIENCE

Outline

1. Introduction: Why is this a problem?

2. Attention is All You Need: Some background insights on attention mechanisms

3. A survey on attention mechanisms for medical applications: are we moving towards better algorithms?

1. Introduction: Why is this a problem?

Deep learning has been challenging human performance

- **The increase of available computational power and the democratised access to a huge amount of data** has leveraged the development of novel artificial intelligence (AI) algorithms and their applications

- Deep learning techniques **have been challenging human performance** at some specific tasks such as cancer detection in biomedical imaging[1] or machine translation in natural language processing[2]

- However, most of these models work as black boxes (i.e., their internal logic is hidden to the user) that receive data and output results **without justifying their predictions in a human understandable way**[3]

Sources: **[1]** McKinney et al. "International evaluation of an AI system for breast cancer screening", **[2]** Belinkov and Glass "Analysis Methods in Neural Language Processing: A Survey", **[3]** Cynthia Rudin "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead"

This motivated the creation of explainable AI

- The topic of explainable artificial intelligence (XAI) appeared intending to contribute to a more **transparent AI**[1]
  - There are three distinct strategies: **pre-, in- and post-model** methods

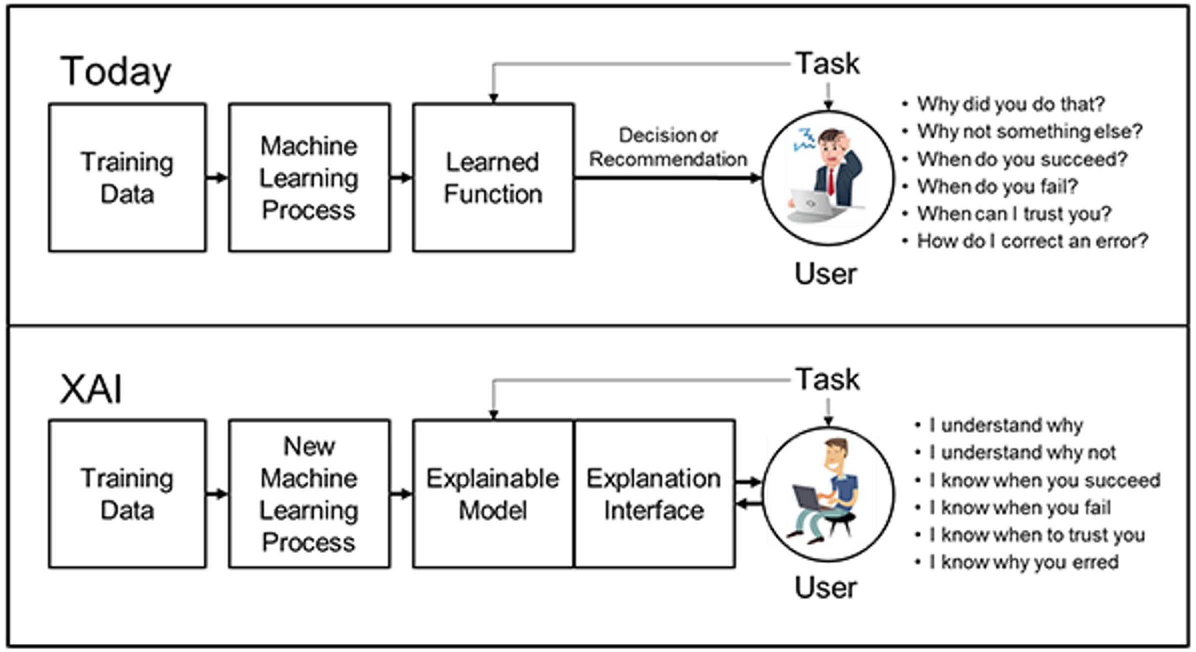


Figure: The concept of XAI, from the Defense Advanced Research Projects Agency (DARPA)[2]

Sources: [1] [L. H. Gilpin et al. "Explaining Explanations: An Overview of Interpretability of Machine Learning"](), **[2]** https://www.darpa.mil/program/explainable-artificial-intelligence

Now, we must understand the “explanations”

- The **generalised belief that complex models seem to uncover “hidden patterns”** actively contributed to the research and **development of post-model** methods

- There are **several drawbacks of exclusively investing** in a post-model strategy[1]:
  - **Explanations are just an approximation** to what the model computes
  - **Explanations may not provide enough detail** to understand what the model is doing

- It is fundamental to assess the quality of these explanations[1] and to dedicate more effort to pre- and in-model strategies
  - **Pre-model interpretability**: understanding the data distribution that we are dealing with will contribute to an increase of confidence with the posterior decisions and explanations[2]

  - **In-model interpretability**: since models that are inherently interpretable provide their explanations and are faithful to what the machine learning model actually computes[1]

Sources: [1] Cynthia Rudin “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”,
[2] Wilson Silva et al. “How to produce complementary explanations using an Ensemble Model”

2. Attention is All You Need: Some background insights on attention mechanisms

What if some parts of the data are more relevant than others?

- In AI systems, **some parts of the input data are more relevant than others** (e.g., in automatic translation systems, only a subset of words is relevant)[1]
  - In the deep learning context, the first successful implementations of attention mechanisms were accomplished with RNNs, which can learn and process data with a temporal component

- A possible taxonomy for the classification of attention mechanisms[1] proposes the following categories
  - **Number of Abstraction Levels:** single-, multi-level
  - **Number of Positions:** soft, hard, global, local
  - **Number of Representations:** single-, multi-representational, multi-dimensional
  - **Number of Sequences:** distinctive, co-attention, self-attention

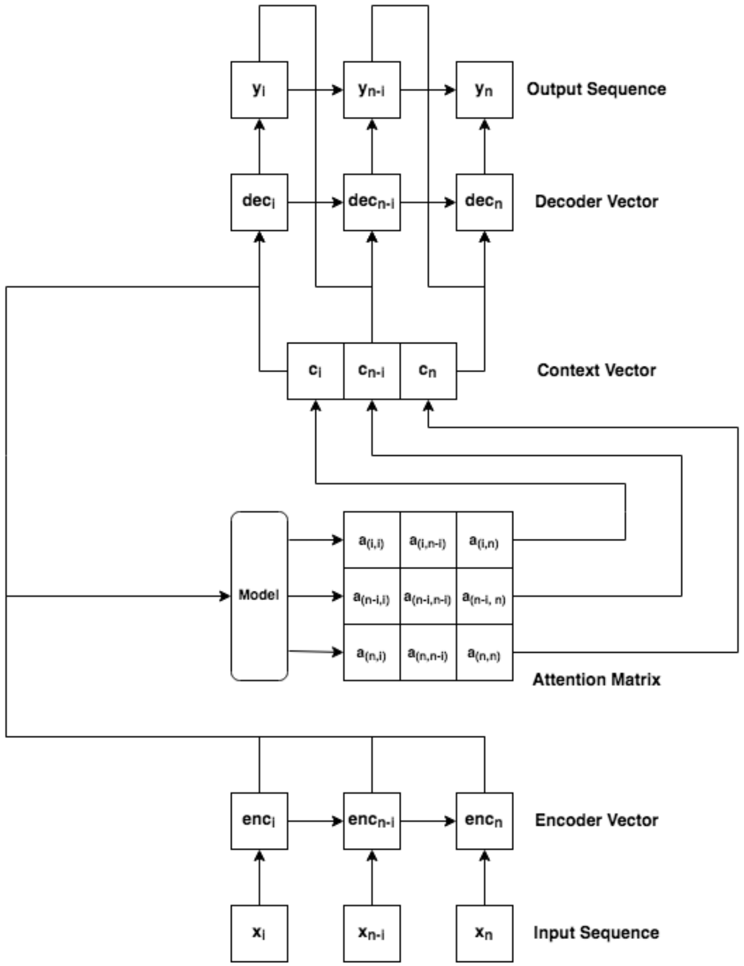
Sources: [1] S. Chaudhari et al. “An Attentive Survey of Attention Models”

It all started in the field of natural language processing

Language, Text and Speech

- **The first paradigm of attention** was based on long short-term memory network (LSTM) applied to the task of **neural machine translation** and allowed the study of **local and global attention mechanisms**[1]

- The introduction and the success of the **Transformer** architecture[2] which is based **solely on attention mechanisms** (dot-product and multi-head), allowed the creation of a new paradigm for the study of attention mechanisms



Sources: [1] Minh-Thang Luong et al "Effective Approaches to Attention-based Neural Machine Translation",
[2] Ashish Vaswani et al. "Attention is All you Need"

But rapidly permeated to the field of computer vision

Computer Vision

- Aligned with the task of image captioning, several **attention-based approaches** (multiple, soft and hard) have been proposed to generate **meaningful semantic representations**[1]
- Several **channel and spatial attention** modules have been developed[2] to facilitate the **regression of finer features** and to **efficiently model relationships** between widely separated spatial regions
- Recently, a successful application of the Transformer architecture on the computer vision domain has been proposed: the **Vision Transformer**[3]

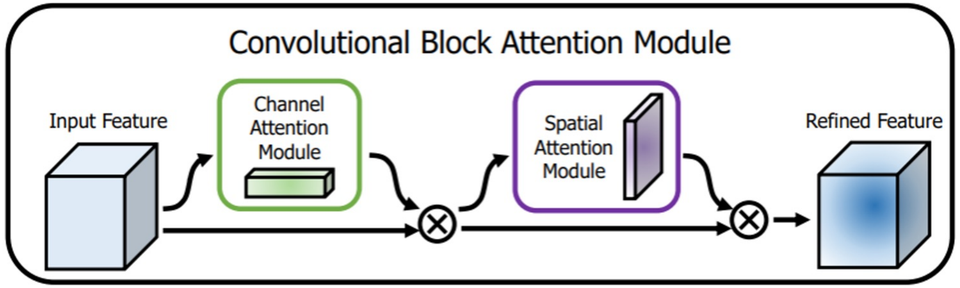


Figure: Channel and Spatial Attention[2]

Sources: [1] Peng Wang et al. "Multi-Attention Network for One Shot Learning", [2] Sanghyun Woo et al. "CBAM: Convolutional Block Attention Module", [3] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale"

The medical domain is an interesting use case that benefits from both

- Most of the use-cases focus on **medical image segmentation or classification** using **different modalities** (e.g., computed tomography, magnetic resonance imaging, ultrasound[1], positron emission tomography)

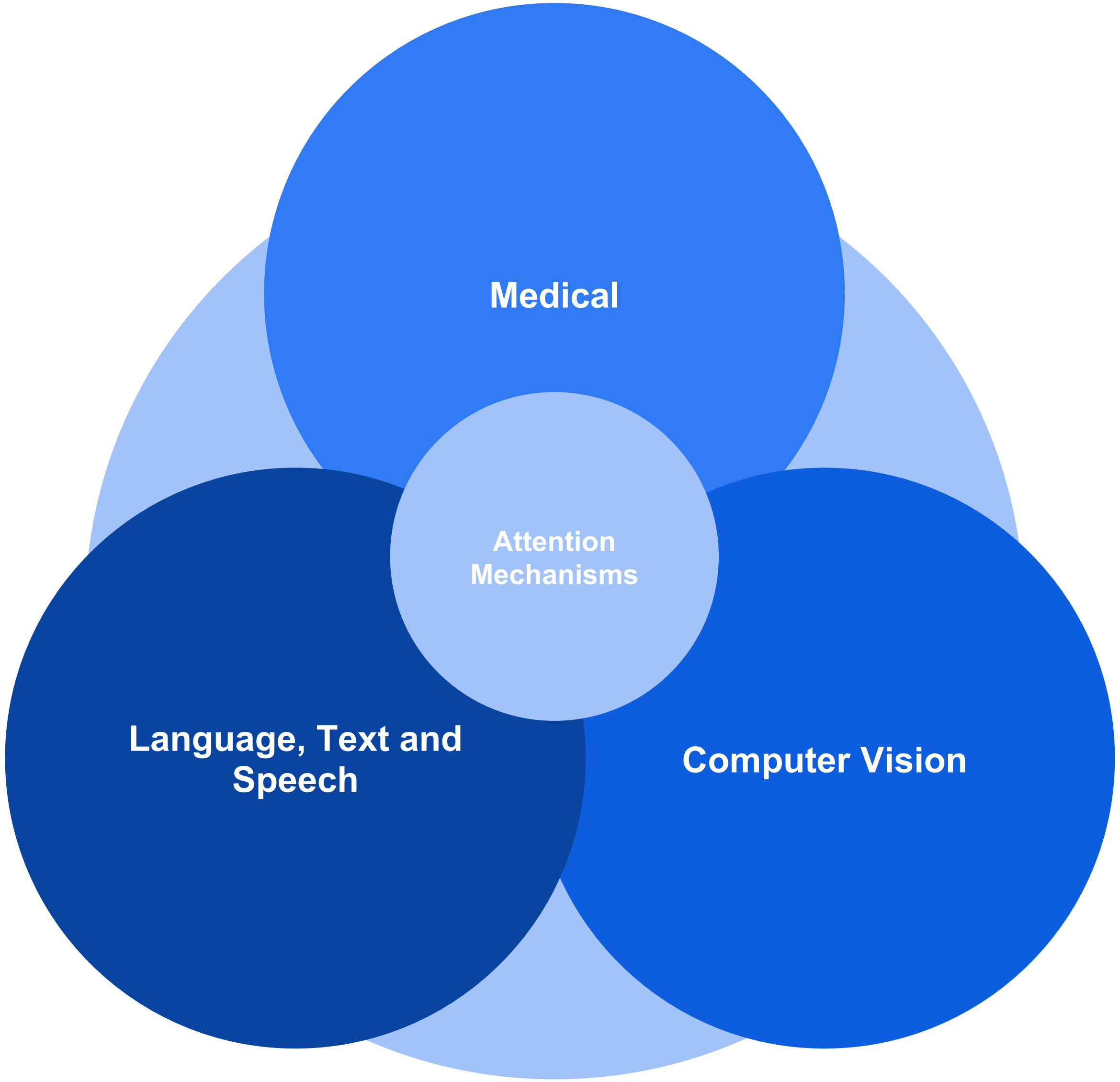
- In automatic report generation, different attention methodologies (**contrastive**, **variational topic inference**) have been proposed to represent better the visual features of abnormal regions or to align image and language modalities in a latent space, thus improving the quality of the generated reports[2]

- The potential of Transformer-based architectures is also being explored in the medical context, as more recent methodologies on medical image segmentation are taking advantage of a hybrid use of the **Vision Transformer and the U-Net**[3]

Sources: [1] Aleksandar Vakanski et al. "Attention Enriched Deep Learning Model for Breast Tumor Segmentation in Ultrasound Images", [2] Fenglin Liu et al. "Contrastive Attention for Automatic Chest X-ray Report Generation", [3] Bingzhi Chen et al. "TransAttUnet: Multi-level Attention-guided U-Net with Transformer for Medical Image Segmentation"

INESCTEC

In the last years, attention mechanisms have played a key role in all these domains



3. A survey on attention mechanisms for medical applications: are we moving towards better algorithms?

We performed an extensive review of attention mechanisms for medical applications with a strong focus on methodologies and applications[1]

We approached this topic with a critical view and asked the following research questions:

- Will attention mechanisms automatically improve the predictive power of deep learning algorithms for medical image applications?
- What is the impact of integrating attention mechanisms on model complexity?
- Can we improve the degree of interpretability of deep learning models solely through attention mechanisms?
- How practical is it to design and build attention mechanisms for deep learning applications?

Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

We performed experiments on different medical use cases[1]

APTOS2019

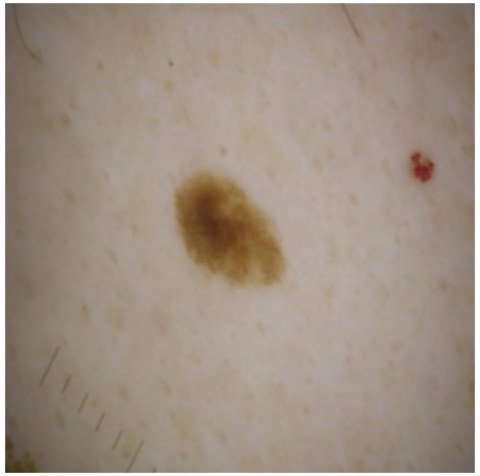
Retinography data related to retinopathy severity score

In our paper, we worked on a binary case: ”Normal” vs “Diabetic Retinopathy”

ISIC2020

Dermoscopic images of benign and malignant skin lesions

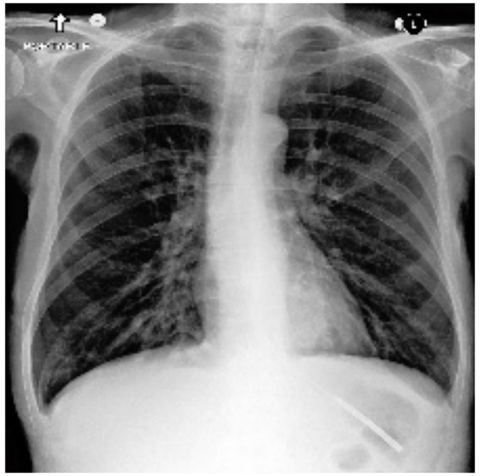
In our paper, we worked on the binary case: “Benign” vs “Malign”



MIMIC-CXR

Chest radiographs database

In our paper, we worked on the binary case: “Normal” vs “Pleural Effusion”

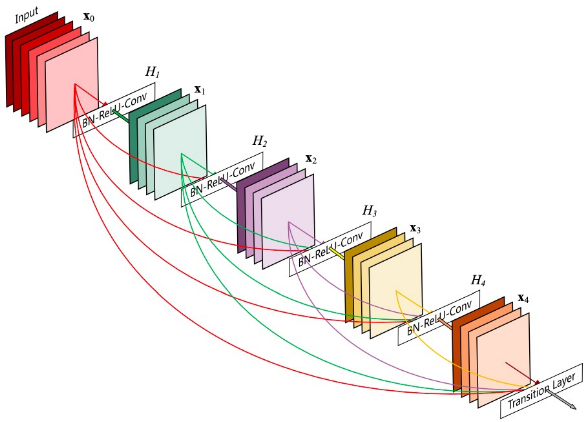


Sources: [1] Tiago Gonçalves et al. “A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

Using two different backbone architectures[1]

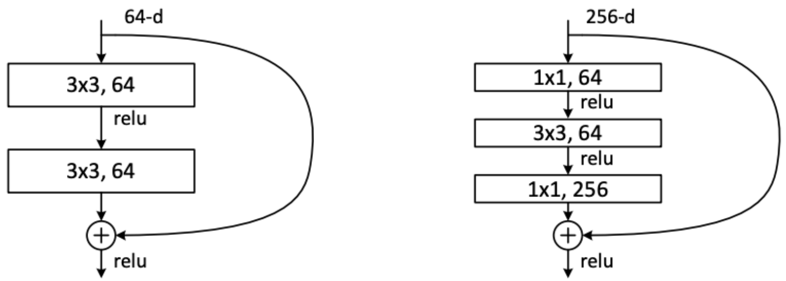
DenseNet-121[2]

Allows connecting all layers directly with each other, thus improving the flow of information and gradients throughout the network, and facilitating their training



ResNet-50[3]

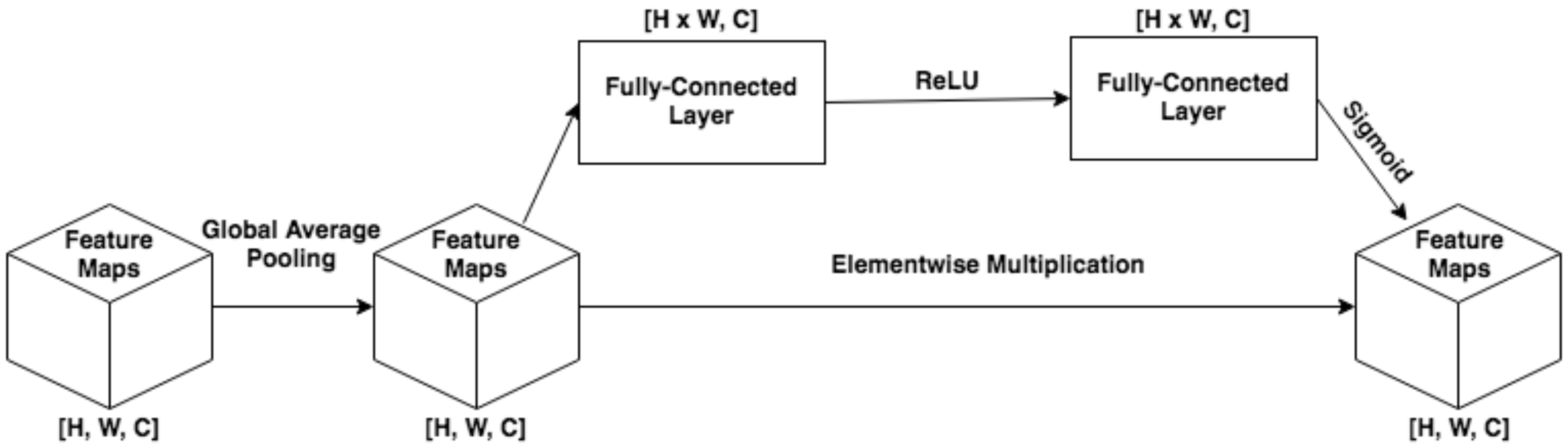
Introduced the deep residual learning framework, which consists of adding skip connections that perform identity mapping and adding their outputs to the outputs of the stacked layers



Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?", [2] Gao Huang et al. "Densely Connected Convolutional Networks", [3] Kaiming He et al. "Deep Residual Learning for Image Recognition"

We integrated the *Squeeze-and-Excitation (SE)* block into the backbones[1]

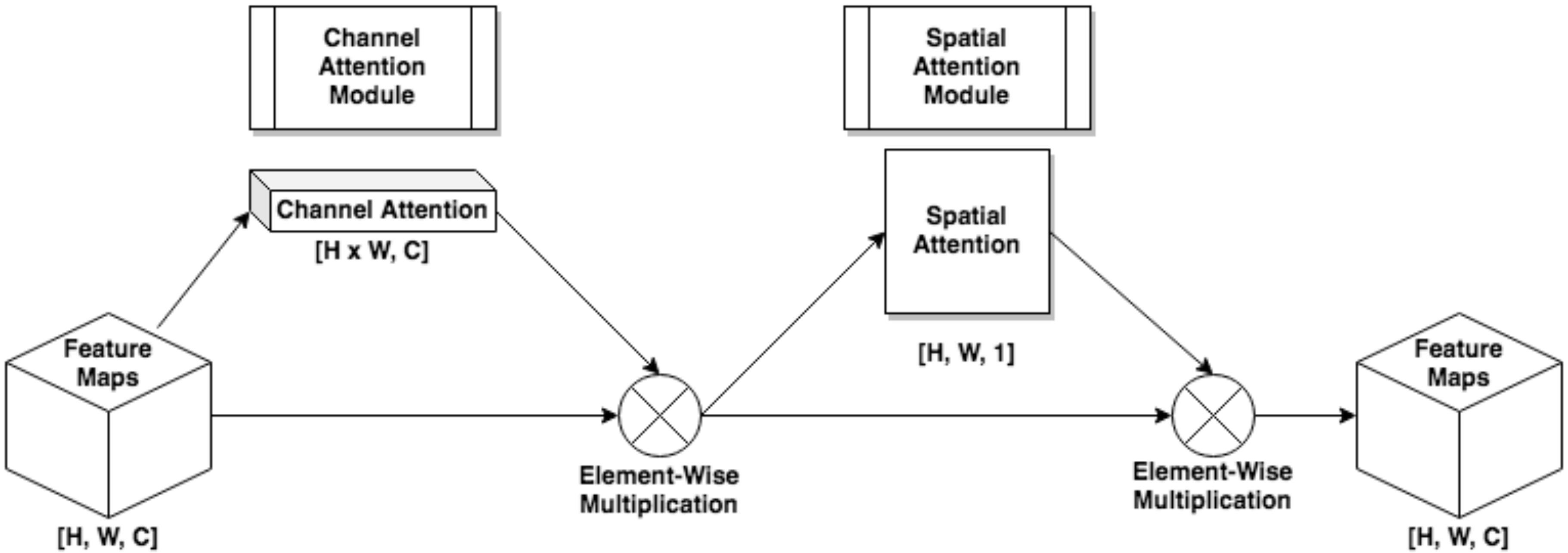
This attention block was designed to adaptively recalibrate channel-wise feature responses by explicitly modeling interdependencies between channels[2]



Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?", [2] Jie Hu et al. "Squeeze-and-Excitation Networks"

We also integrated the *Convolutional Block Attention Module (CBAM)* block into the backbones[1]

This attention mechanism integrates two specific attention blocks[2]:

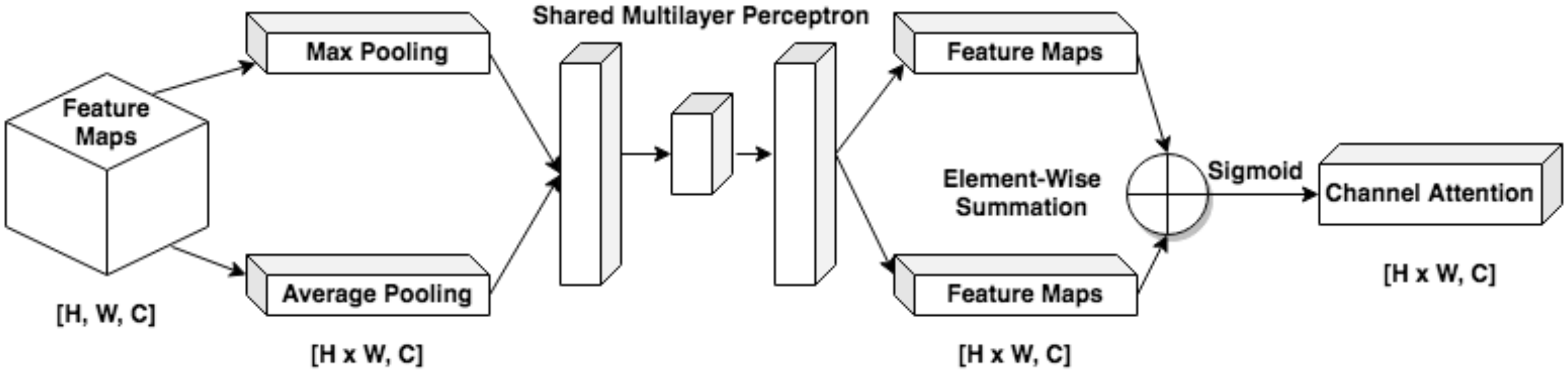


Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?", [2] Sanghyun Woo et al. "CBAM: Convolutional Block Attention Module"

We also integrated the *Convolutional Block Attention Module (CBAM)* block into the backbones[1]

This attention mechanism integrates two specific attention blocks[2]:

- **The channel attention module**, which aims to produce a channel attention map by exploiting the inter-channel relationship of features and is considered as a feature detector

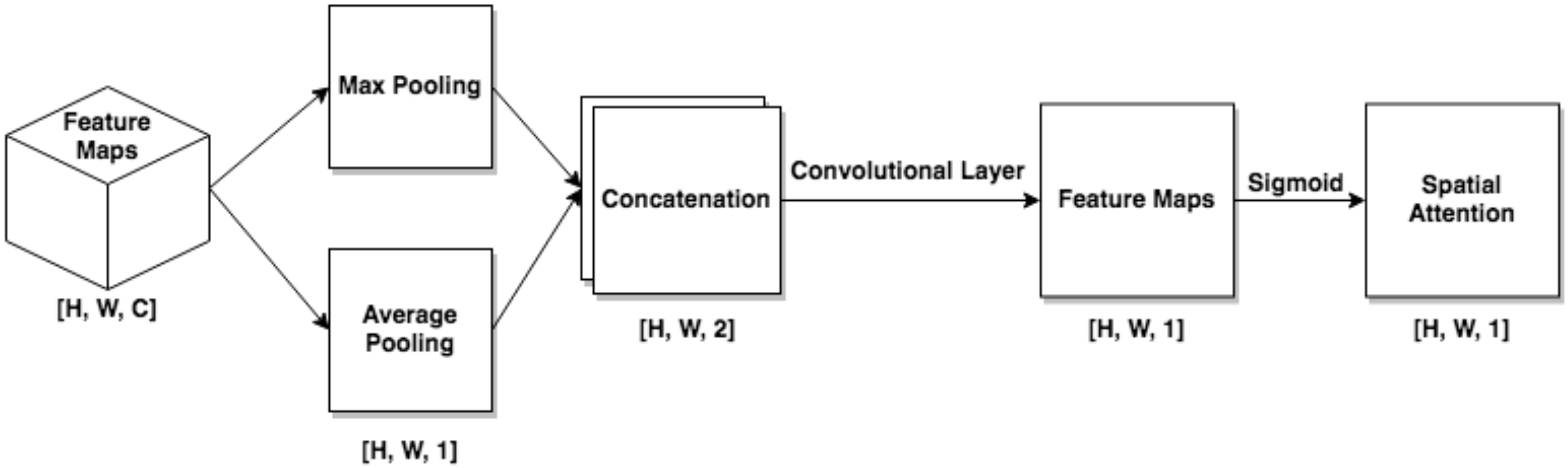


Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?", [2] Sanghyun Woo et al. "CBAM: Convolutional Block Attention Module"

We also integrated the *Convolutional Block Attention Module (CBAM)* block into the backbones[1]

This attention mechanism integrates two specific attention blocks[2]:

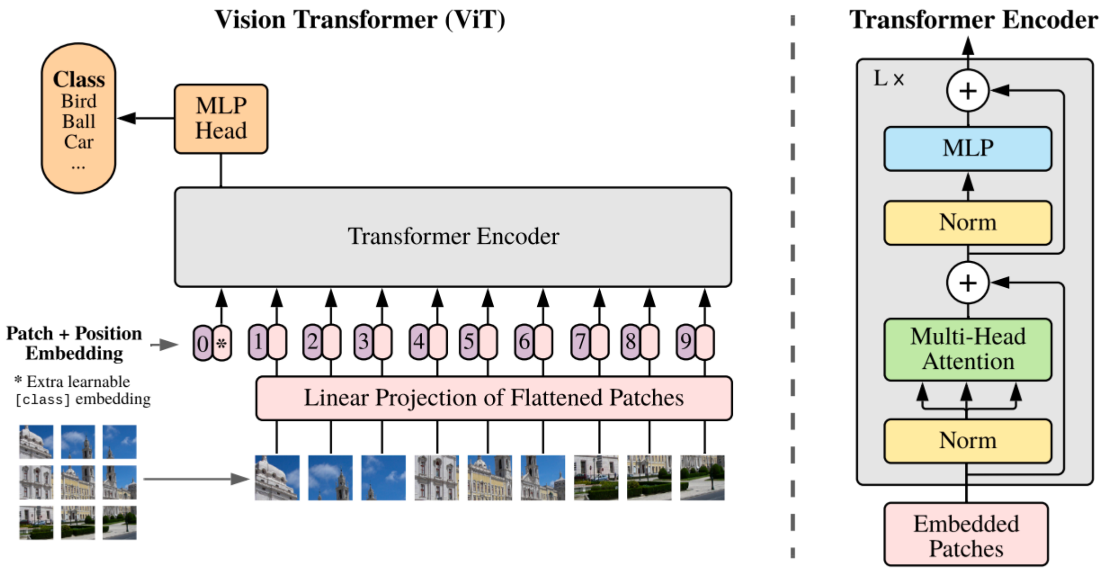
- **The spatial attention module**, which aims to generate a spatial attention map by utilizing the inter-spatial relationship of features, thus being complementary to the channel attention

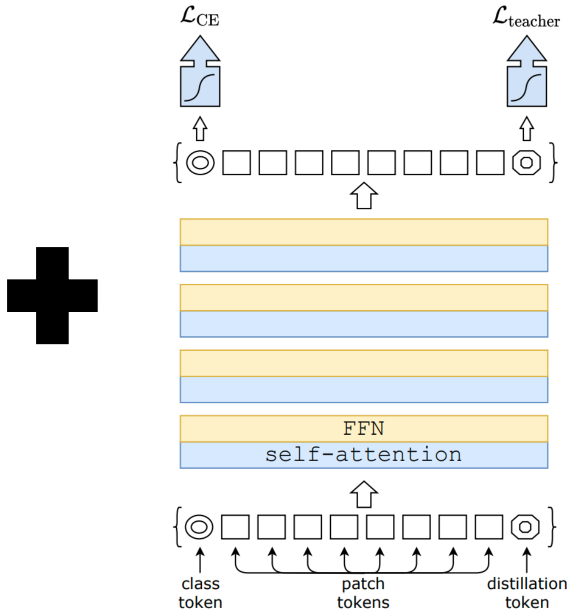


Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?", [2] Sanghyun Woo et al. "CBAM: Convolutional Block Attention Module"

Finally, we tested a Transformer-based architecture composed solely of attention mechanisms[1]

The Data-efficient image Transformer (DeiT)[2] is an architecture inspired by the Vision Transformer[3] and trained with fewer parameters. In this case, we used the DeiT-Ti variation[2], which has a comparable number of parameters against the chosen CNN backbones





Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?", [2] Hugo Touvron et al. "Training data-efficient image transformers & distillation through attention", [3] Alexey Dosovitskiy "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale"

The level of interpretability was measured using post-hoc methods[1]

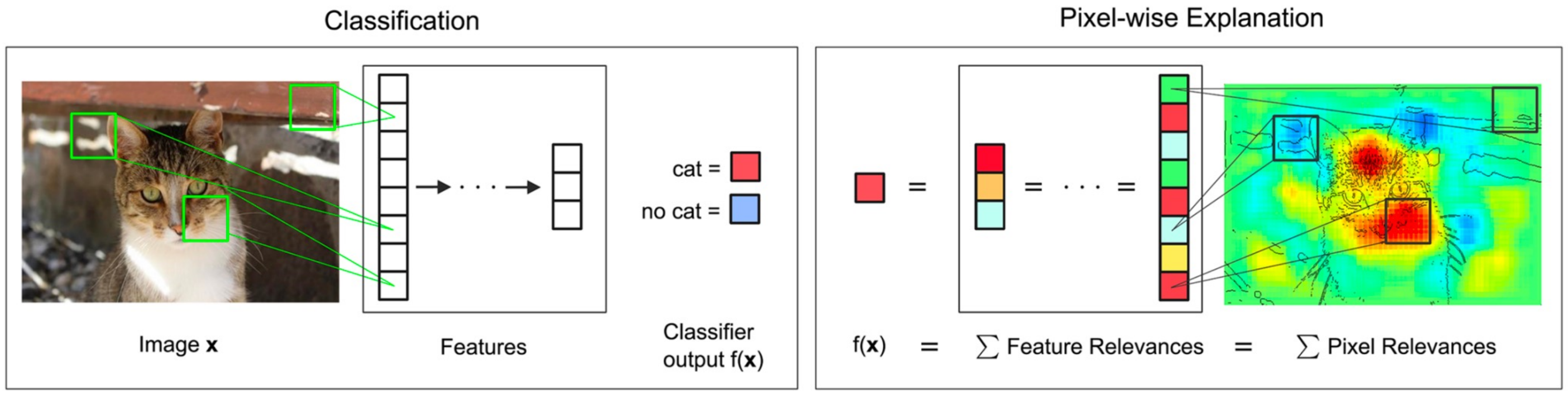
DeepLIFT[2]

The *Deep Learning Important FeaTures* (DeepLIFT) compares the activation of each neuron to its related reference activation and assigns contribution scores according to the difference

LRP[3]

The *Layer-wise Relevance Propagation* (LRP) is a methodology that aims to create visualizations of the contributions of single pixels to predictions

Image below provides an intuition



Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?", [2] Avanti Shrikumar et al. "Learning Important Features Through Propagating Activation Differences", [3] Sebastian Bach et al. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation"

Hence, the final experimental protocol looks like this[1]:

Data

- APTOS2019
- ISIC2020
- MIMIC-CXR

Models

- DenseNet-121
- ResNet-50
- SEDenseNet-121
- SEResNet-50
- CBAMDenseNet-121
- CBAMResNet-50
- DeiT

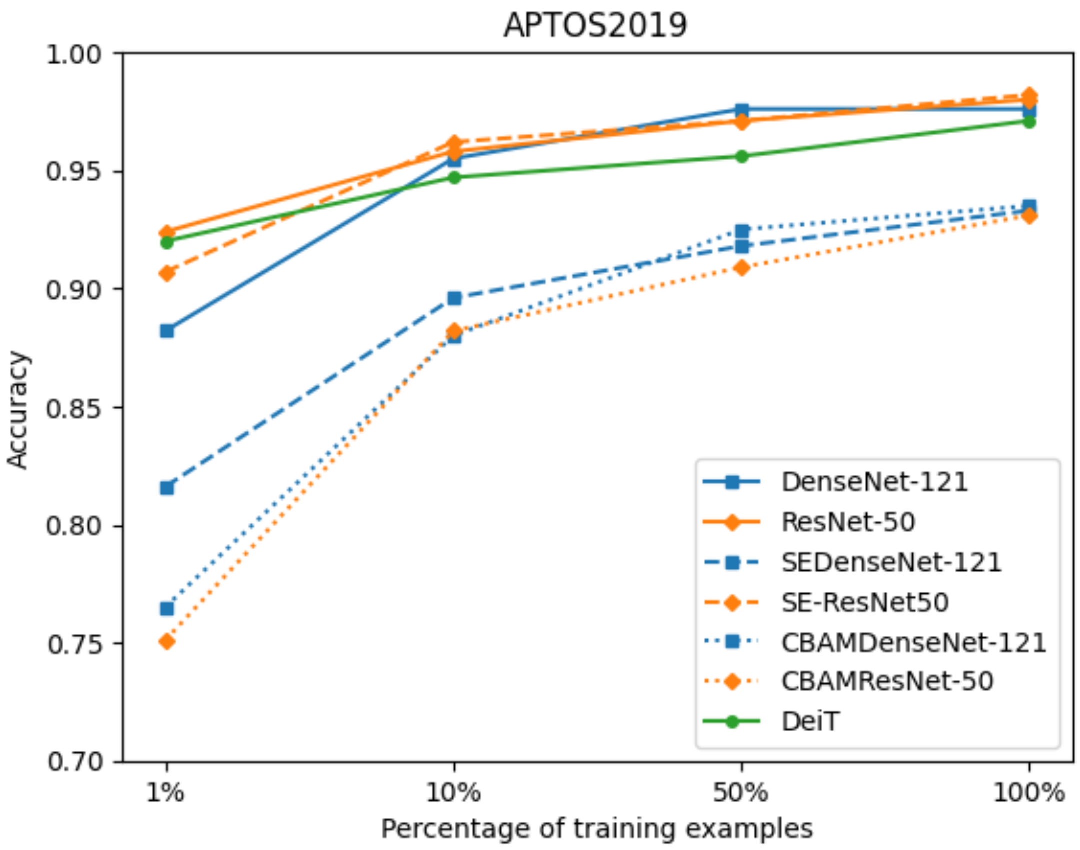
xAI

- DeepLIFT
- LRP

We also employed several training data regimes (1%, 10%, 50% and 100%)

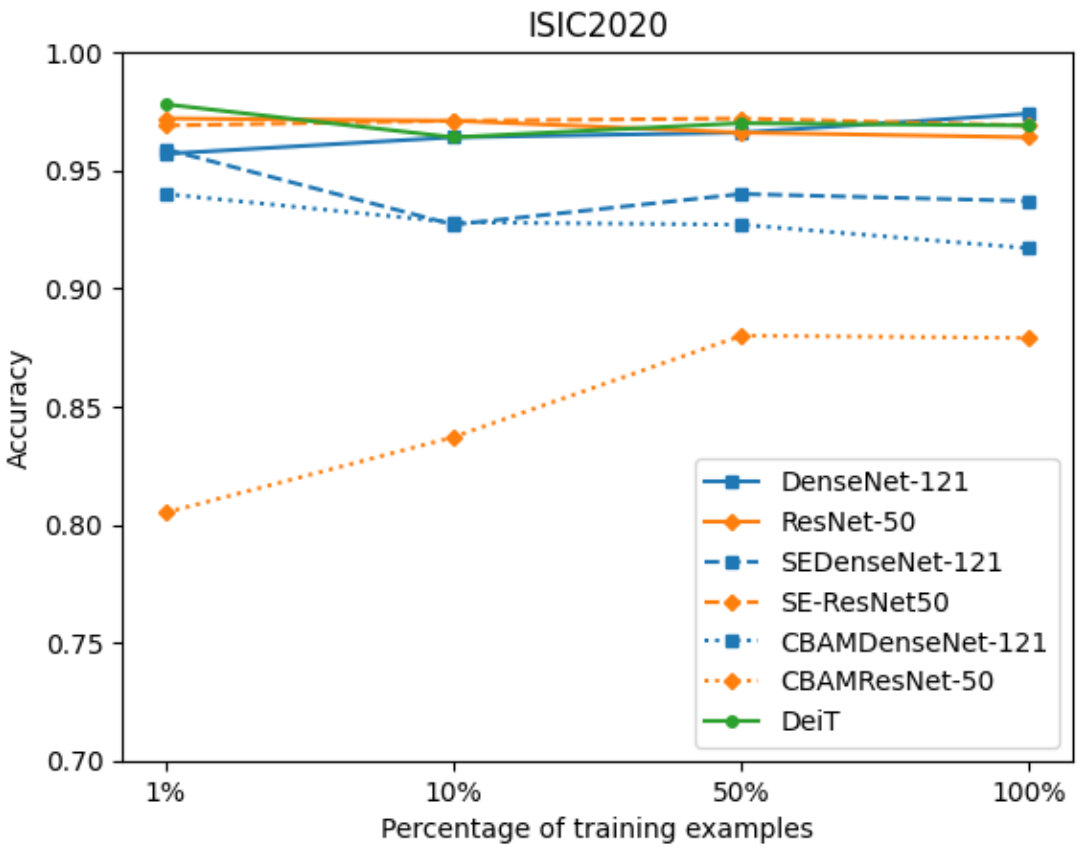
Sources: [1] [Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"]

What are the expectations concerning predictive performance?[1]



Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

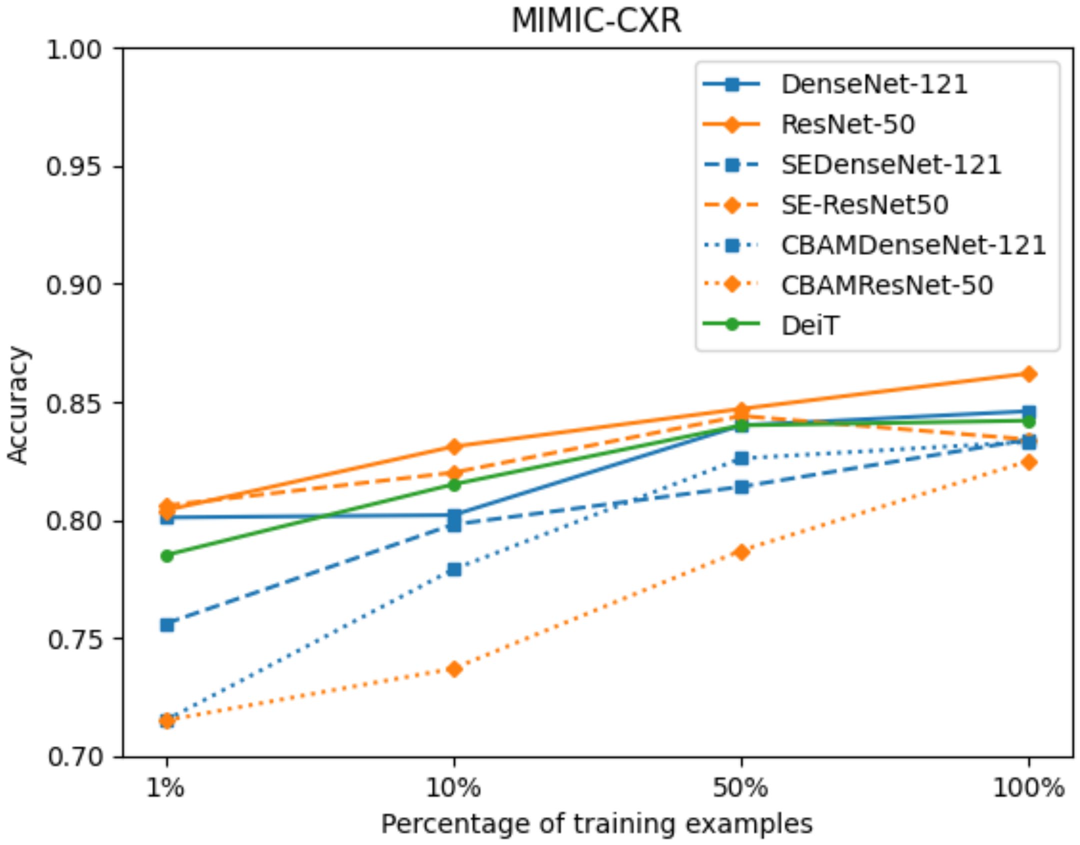
What are the expectations concerning predictive performance?[1]



Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

INESCTEC

What are the expectations concerning predictive performance?[1]



Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

What can we conclude regarding predictive performance?[1]

- All the experiments related to the predictive performance of deep learning models on the different data sets suggest that it is not clear that one should expect improvements in their accuracy when using attention mechanisms

- Given that the intuition behind attention mechanisms is that these end up learning the most relevant features, one might expect that attention-based architectures would perform better when trained in low data regimes. However, results obtained in all data sets suggest that this might not be the case

- The results reported in the literature often relate to marginal or residual improvements in the state-of-the-art backbone networks. Given that the training of deep learning algorithms is generally a stochastic process, there is a need to assess these reported improvements with a more critical view and with robust statistical tests

Sources: [1] Tiago Gonçalves et al. “A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

Are we decreasing the complexity of our models?[1]

Model	Number of parameters
DenseNet-121	7,054,210
ResNet-50	23,512,130
SEDenseNet-121	7,357,314
SEResNet-50	26,027,074
CBAMDenseNet-121	7,360,706
CBAMResNet-50	26,044,722
DeiT	5,486,786

Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

What can we conclude regarding model complexity?[1]

- The integration of attention mechanisms increases the number of parameters of the deep learning models, thus increasing their complexity. This information allows us to conclude that, at least for computer vision applications, it is not necessarily true that the use of attention mechanisms contributes to the decrease of model complexity

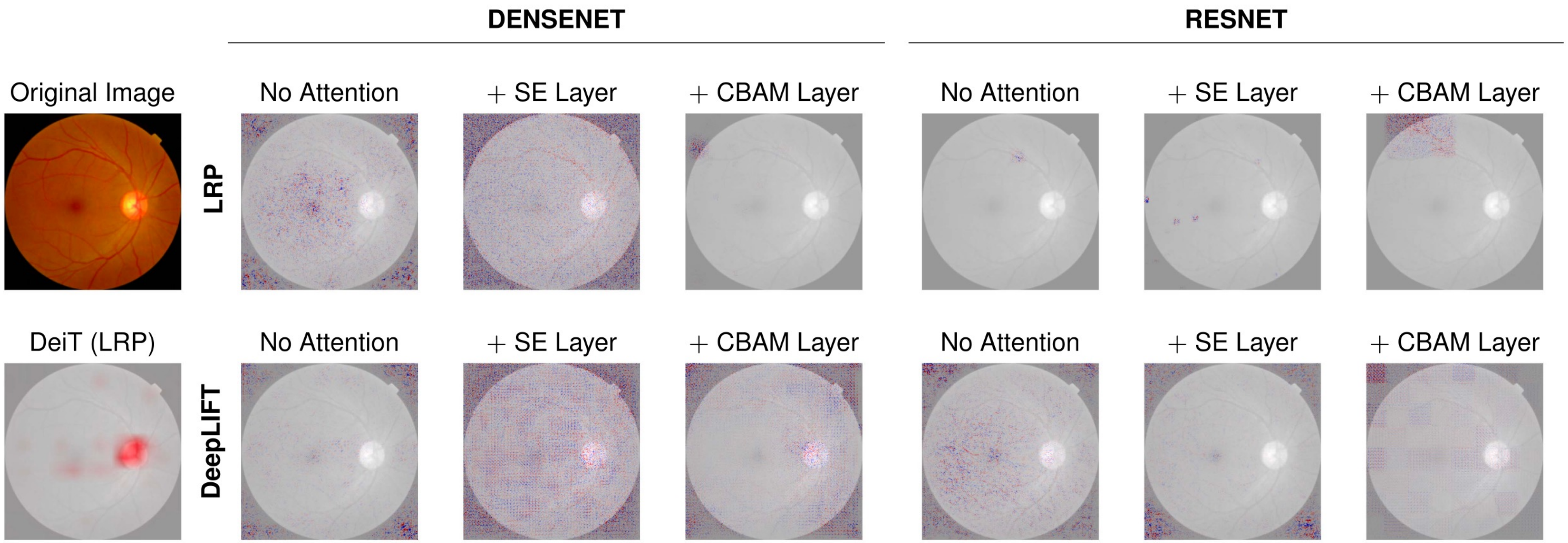
- On the other hand, since these attention mechanisms often rely on simple operations (e.g., matrix multiplications), such as convolutions, we acknowledge that their use may reduce the training time of deep learning algorithms (similarly to what happened when the community started using CNNs)

- Another question arises from these results: are these attention-based algorithms allegedly performing well because of the inner-functioning of their attention mechanisms themselves or just because we are increasing the number of model parameters? While one may report that this issue is nonsense, we point out that some Transformer-based architectures have a considerably high number of parameters

Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

What about explainability?[1]

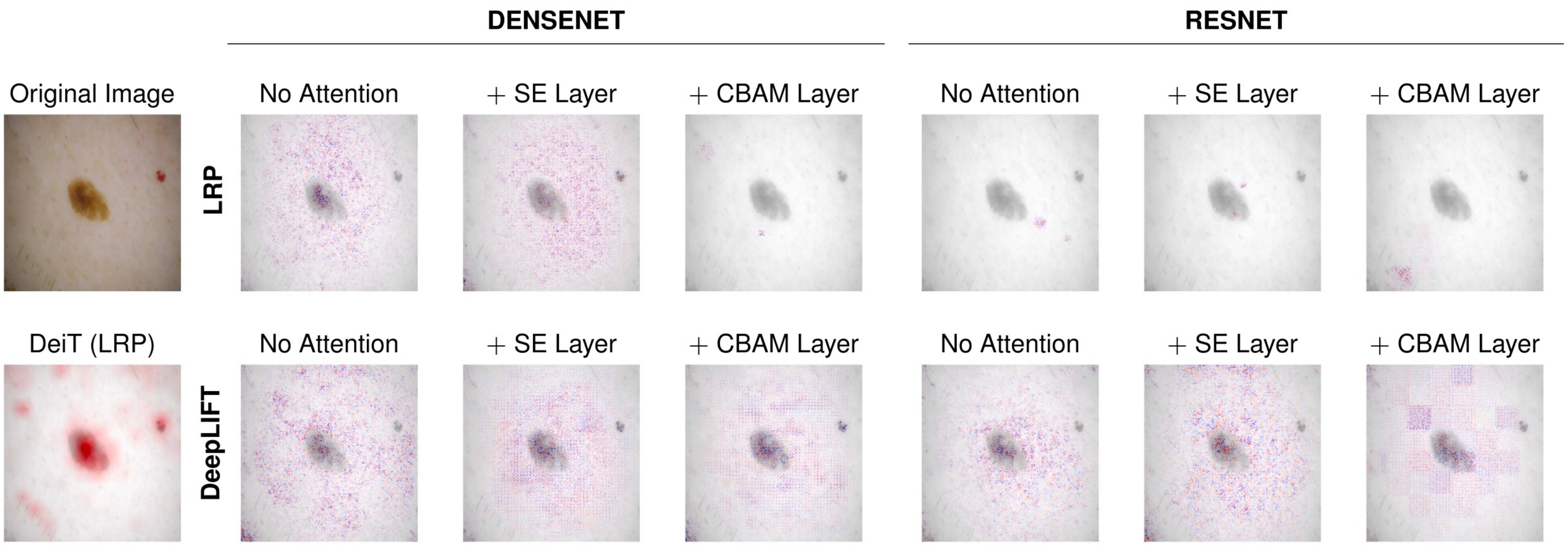
TABLE 5. **Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the APTOS2019 data set with the label 0 correctly classified as 0 by all models.**



Sources: [1] Tiago Gonçalves et al. “A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

What about explainability?[1]

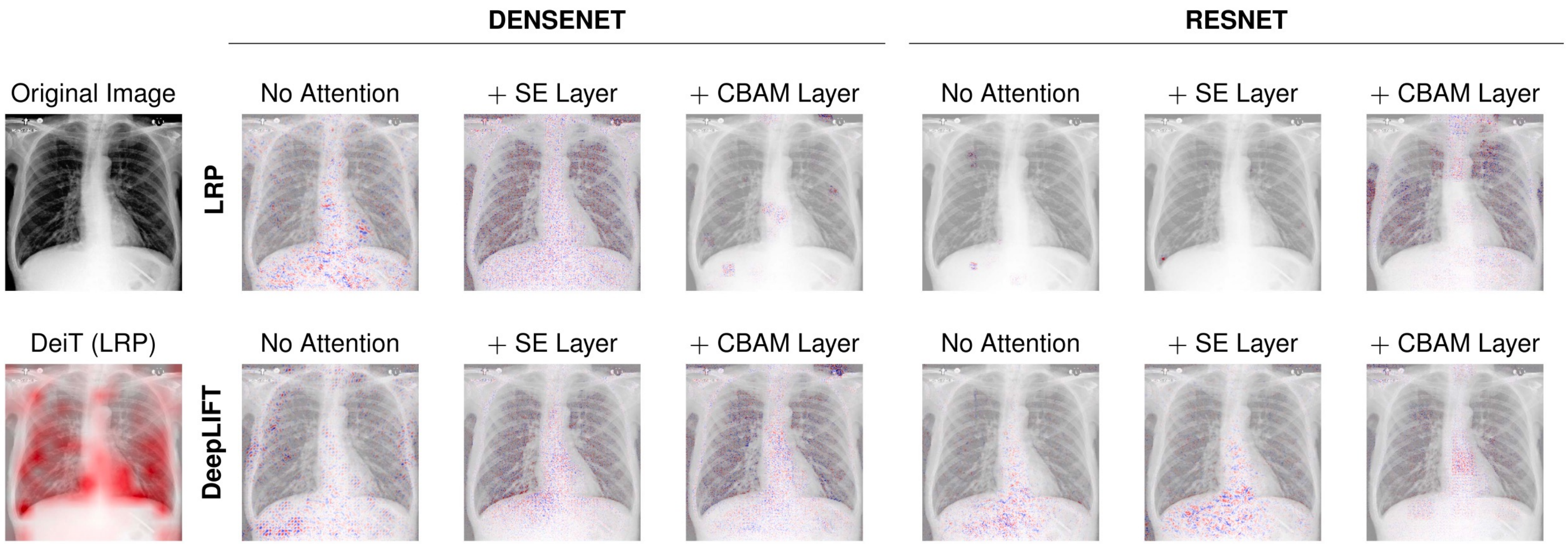
TABLE 8. **Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the ISIC2020 data set with the label 0 correctly classified as 0 by all models.**



Sources: [1] Tiago Gonçalves et al. “A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

What about explainability?[1]

TABLE 12. **Example of LRP and DeepLIFT *post-hoc* saliency maps for an image of the MIMIC-CXR data set with the label 0 correctly classified as 0 by all models.**



Sources: [1] Tiago Gonçalves et al. “A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

Are there any valuable insights regarding explainability?[1]

- When analyzing the results obtained with the baseline models, we would expect that, with the increase in complexity of the attention mechanism, the distribution of the important pixels around the image would also be more focused. However, that does not seem to happen in our use cases. Interestingly, besides the inherent properties of the data and the task, the results drastically change when we use a different backbone

- Besides, it is also important to remind the you that these frameworks allow us to generate explanations even for the cases where the model miss-classifies. We also stress that one of the limitations of such analysis is that there does not exist an objective ground truth of what a high-quality visual explanation is

- It is not trivial, in computer vision tasks such as this, to conclude with complete confidence that, even for the cases where the model succeeds, it learned the right correlations. Hence, can we believe the narrative that attention mechanisms are learning the most relevant features of the image?

Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

Are we moving towards better algorithms?[1]

- We found that backbone models can attain equivalent predictive performances to Transformer-based architectures with equivalent model complexity (i.e., number of parameters)

- When using a post-hoc framework to visually assess what type of features these models can extract, we can conclude that there is still a high degree of subjectivity in such analysis (i.e., results are very noisy, even for the cases of attention mechanisms, which is counter-intuitive)

- The community is moving toward using attention mechanisms (specially Transformer-based ones) and arguing that these frameworks increase the quality, transparency, and interpretability of deep learning architectures. However, we state that this is not true

Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

A New Hope (or Future Challenges)[1]

1. **Attention Mechanisms: Past or Future?** Even if attention mechanisms are pushing deep learning algorithms towards the limits of their predictive power, we must start thinking about creating interpretable frameworks that allow us to audit and assess these algorithms concerning the specific conditions of their domains

2. **Design and Integration of Attention Mechanisms** If we look at the topographies of these deep learning algorithms, it is not always clear for the users where they should place these modules, and why it makes sense to put them in a specific place. Another question arises: are these attention modules dependent on the backbone into which they are integrated to?

3. **The Rise of Transformers** While there is hype on the use of these structures, it is not clear whether they are more interpretable or not, or if their generalization power is superior to the other deep models

4. **Interpretability is the Path to Better Algorithms** Even if we intend to keep using visual saliency maps to explain our models, we must achieve a clear standard, validated by the clinical community, of what these maps should look like and what is their effective meaning

Sources: [1] Tiago Gonçalves et al. "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?"

Attention Mechanisms for Medical Applications: Claims, Potentialities and Future Challenges

Deep Learning Sessions Portugal

March 15, 2023 | Startup Lisboa (Lisboa)

Tiago Filipe Sousa Gonçalves

tiago.f.goncalves@inesctec.pt

INESCTEC

INSTITUTE FOR SYSTEMS
AND COMPUTER ENGINEERING,
TECHNOLOGY AND SCIENCE