Assessing Knee OA Severity with CNN attention-based end-to-end architectures

Marc Górriz[1]   ALGAYON2@GMAIL.COM
Joseph Antony[2]   JOSEPH.ANTONY@DCU.IE
Kevin McGuinness[2]   KEVIN.MCGUINNESS@DCU.IE
Xavier Giró-i-Nieto[1]   XAVIER.GIRO@UPC.EDU
Noel E. O'Connor[2]   NOEL.OCONNOR@DCU.IE

[1] *Universitat Politècnica de Catalunya (UPC), Barcelona, Catalonia/Spain*
[2] *Insight Centre for Data Analytics, Dublin City University (DCU), Dublin, Ireland*

Editors: Under Review for MIDL 2019

Abstract

This work proposes a novel end-to-end convolutional neural network (CNN) architecture to automatically quantify the severity of knee osteoarthritis (OA) using X-Ray images, which incorporates trainable attention modules acting as unsupervised fine-grained detectors of the region of interest (ROI). The proposed attention modules can be applied at different levels and scales across any CNN pipeline helping the network to learn relevant attention patterns over the most informative parts of the image at different resolutions. We test the proposed attention mechanism on existing state-of-the-art CNN architectures as our base models, achieving promising results on the benchmark knee OA datasets from the osteoarthritis initiative (OAI) and multicenter osteoarthritis study (MOST). All code from our experiments will be publicly available on the github repository: https://github.com/marc-gorriz/KneeOA-CNNAttention

Keywords: Convolutional Neural Network, End-to-end Architecture, Attention Algorithms, Medical Imaging, Knee Osteoarthritis.

1. Introduction

Knee osteoarthritis (OA) is the most common articular disease and a leading cause of chronic disability (Heidari, 2011), and mainly affects the elderly, obese, and those with a sedentary lifestyle. Degenerative processes of the articular cartilage as a result of excessive load on the joint, and the aging process, contributes to the natural breakdown of joint cartilage with joint space narrowing (JSN) and osteophytes (Li et al., 2013). Knee OA causes excruciating pain and often leads to joint arthroplasty in its severe stages. An early diagnosis is crucial for clinical treatment to be effective in curtailing progression and mitigating future disability (Oka et al., 2008) (Shamir et al., 2009a). Despite the introduction of several imaging modalities such as MRI, OCT, and ultrasound for augmented OA diagnosis, X-Ray is still the method of choice in diagnosing knee OA, although clinical evidence also contributes.

Previous work has approached the challenge of automatically assessing knee OA severity as an image classification problem (Shamir et al., 2009a) (Shamir et al., 2008) (Yoo et al., 2016) using the Kellgren and Lawrence (KL) grading (Kellgren and Lawrence, 1957). KL

grading quantifies the degree of degeneration on a five-point scale (0 to 4): KL-0 (no OA changes), KL-1 (doubtful), KL-2 (early OA changes), KL-3 (moderate), and KL-4 (end-stage). Assessment is based on JSN, presence of osteophytes, sclerosis, and bone deformity. Most methods in the literature use a two step process to automatically quantify knee OA severity: 1) localization of knee joints; and 2) quantification of severity. Separate models for knee joint localization, either using hand-crafted features (Shamir et al., 2008), (Yoo et al., 2016) or CNNs (Antony et al., 2016) are not always highly accurate, affecting the subsequent quantification accuracy and adding extra complexity to the training process.

To overcome this problem, this work proposes a novel end-to-end architecture incorporating trainable attention modules that act as unsupervised fine-grained ROI detectors, which automatically localize knee joints without a separate localization step. The proposed attention modules can be applied at different levels and scales across an arbitrary CNN pipeline. This helps the network to learn attention patterns over the most informative parts of the image at different resolutions, achieving improvements in the quantification performance.

2. Related work

Much of the literature has proposed image classification-based solutions to assess knee OA severity using radiography-based semi-quantitative scoring systems, like KL gradings, which are based on the study of anatomical features such as variations in joint space width or osteophytes formation (Shamir et al., 2009b) (Shamir et al., 2009a) (Shamir et al., 2008). Shamir et al. (Shamir et al., 2008) proposed WND-CHARM: a multi purpose medical image classifier to automatically assess knee OA severity in radiographs using a set of features based on polynomial decompositions, contrast, pixel statistics, textures, and features from image transforms. Recently, Yoo et al. (Yoo et al., 2016) proposed a self-assessment scoring system associating risk factors and radiographic knee OA features using multivariable logistic regression models, additionally using an Artificial Neural Network (ANN) to improve the overall scoring performance. Shamir et. al. (Shamir et al., 2009a) proposed template matching to automatically detect knee joints from X-ray images. This method is slow to compute for large datasets and gives poor detection performance. Antony et al. (Antony et al., 2016) introduced an SVM-based approach for automatically detecting the knee joints. Later, Antony et al. (Antony et al., 2017) proposed an FCN-based approach to improve the localization of the knee joints. Although more accurate, the aspect ratio chosen for the extracted knee joints affects the overall quantification.

Recently, the emergence of deep learning has enabled the development of new intelligent diagnostics based on computer vision. CNNs outperform many state-of-the-art methods based on hand-crafted features in tasks such as image classification (Krizhevsky et al., 2012), retrieval (Babenko et al., 2014) and object detection (Lawrence et al., 1997) (Wei et al., 2011). Antony et al. (Antony et al., 2016) showed that the off-the-shelf CNNs such as the VGG 16-layer network (Simonyan and Zisserman, 2014), the VGG-M-128 network (Chatfield et al., 2014), and the BVLC reference CaffeNet (Jia et al., 2014) (Karayev et al., 2013) pre-trained on ImageNet LSVRC dataset (Russakovsky et al., 2015) can be fine-tuned for classifying knee OA images through transfer learning. They argued that it is appropriate to assess knee OA severity using continuous metrics like mean-squared error together with

binary or multi-class classification losses, showing that predicting the continuous grades through regression reduces the error and improves overall quantification. They proposed a novel pipeline (Antony et al., 2017) to automatically quantify knee OA severity using a FCN for localization and a CNN jointly trained for classification and regression. The work consolidates the state-of-the-art baseline for the application of CNNs in the field, opening a range of research lines for further improvements. Tiulpin et al. (Tiulpin et al., 2018) presented a new computer-aided diagnosis method based on using deep Siamese CNNs, which are originally designed to learn a similarity metric between pairs of images. However, rather than comparing image pairs, the authors extend this idea to similarity in knee x-ray images (with 2 symmetric knee joints). Splitting the images at the central position and feeding both knee joints into a separate CNN branch allows the network to learn identical weights for both branches. They outperform the previous approaches by achieving an average multi-class testing accuracy score of 66.71 % on the entire OAI dataset, despite also needing a localization step to focus the network branches on the knee joint areas.

This work mainly focuses on designing an end-to-end architecture with attention mechanisms. There are similar methods reported in the literature. Xiao et al. (Xiao et al., 2015) propose a pipeline to apply visual attention to deep neural networks by integrating and combining attention models to train domain-specific nets. In another approach, Liu et al. (Liu et al., 2016) introduce a reinforcement learning framework based on fully convolutional attention networks (FCAN) to optimally select local discriminative regions adaptive to different fine-grained domains. The proposed weakly-supervised reinforcement method combined with a fully-convolutional architecture achieves fast convergence without requiring expensive annotation. Recently, Jetley et al. (Jetley et al., 2018) introduce an end-to-end-trainable attention module for CNN architectures built for image classification. The module takes as input the 2D feature vector maps, which forms the intermediate representations of the input image at different stages in the CNN pipeline, and outputs a matrix of scores for each map. They redesign standard architectures to classify the input image using only a weighted combination of local features, forcing the network to learn relevant attention patterns.

3. Method

This section describes the proposed methods, detailing the design, implementation, and training of the attention modules. Several strategies are investigated to integrate the attention mechanism into standard CNN architectures, proposing experimental approaches to classify the knee images.

3.1. Trainable Attention Module for CNNs

The selected attention module is inspired by the work of Kevin Mader (Mader, 2018), in which a trainable attention mechanism is designed for a pretrained VGG-16 network to predict bone age from hand X-Ray images. Figure 1 illustrates this idea.

Given an input volume D^l from a convolutional layer l with N feature maps, several 1×1 convolutional layers are stacked to extract spatial features. The output is then passed to a 1×1 locally connected layer (Chen et al., 2015) (convolution with unshared weights) with sigmoidal activation to give an attention mask A^l. The original feature maps are element-wise multiplied by the attention mask, generating a new convolutional volume
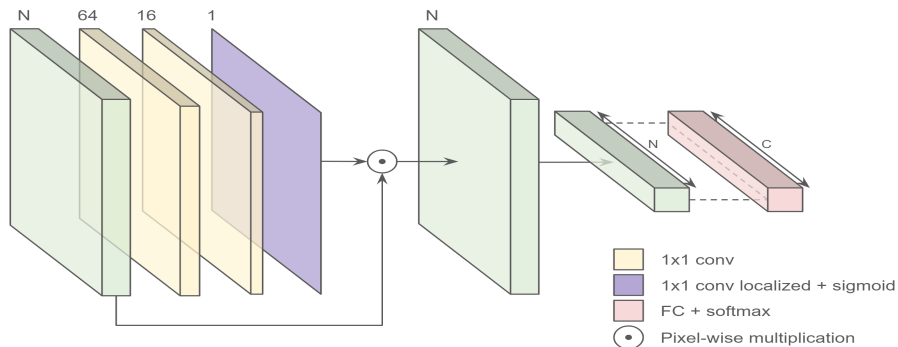
3

Figure 1: Attention module scheme.

$\tilde{D}^l$ accentuating informative areas. A spatial dimensionality reduction is performed by applying global average pooling (GAP) on the masked volume, generating a N-dimensional feature vector F^l, which is then normalized by the average value of the attention mask. Additionally, a *softmax* layer can be applied to yield a C-dimensional vector with the output class probabilities $\{p_1, p_2, ..., p_C\}$.

3.2. Module Integration to CNN Pipeline

Standard CNN architectures typically stack several convolutional layers with occasional pooling operations that reduces the spatial dimension and increase the receptive field. Therefore, the degree of abstraction of the attention modules is closely related to their location in the CNN, focusing on more global details as depth increases. We define the concept of an attention branch as the location of an attention module in a specific convolutional block, applying a *softmax* operation at the top to produce independent class probabilities based on the KL scores. Each attention branch will be seen as a new model by itself that could be trained end-to-end. Figure 2 shows a sample architecture integrating the attention modules to the VGG-16 pipeline. Fixing an input size of 320×224 pixels, as fixed in Section 3.4, we build the branches $att\{i\}_{i=0..2}$ taking as input volumes the feature maps belonging to the pooling layers after the convolutional blocks 3, 4, and 5. Following the methodology in Section 3.3, a combinational module is applied to fuse the local features from all the branches into a global feature vector and then generate the KL grades probability distribution by applying a *softmax* layer at the top.

3.3. Combining Multiple Attention Branches

Several strategies are investigated to merge features from multiple branches with the aim to combine attention patterns at different resolutions. Our first strategy is performing early fusion of features from different branches. Each attention module generates a N-dimensional feature vector F^l with the average values of the N masked feature maps conforming the input convolutional volume. A channel-wise concatenation is applied to fuse all the branches, generating a new vector $F_A = [F^{l_1} \ F^{l_2} \ ... \ F^{l_B}]$ with $F_A \in R^{1x(\sum N_b)}$, being B the total of attention branches and N_b the dimension of F^{l_b}. In addition, a fully connected layer is added at the top to perform early fusion of the concatenated features, while a *softmax*
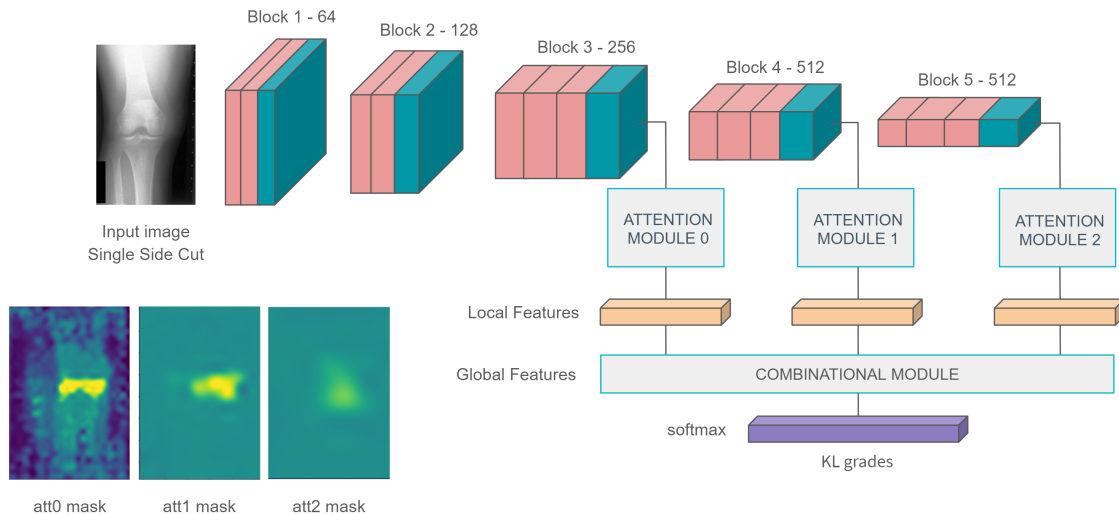
Figure 2: Sample architecture integrating the attention modules in the VGG-16 pipeline.

operation is applied to generate the C-dimensional output class probabilities. As shown above, the complexity of the attention modules correlates with their location in the CNN pipeline, which biases the convergence behavior. This can be critical for a combined model that attempts to train modules with different convergence rates at the same time: deeper branches quickly overfit while waiting for the convergence of the slower ones. In contrast, by reducing the overall learning time, the shallower branches with more complex modules may decrease their performance due to under training.

Our next strategy is to simplify the multiple branch learning process. We propose the use of multi-loss training, which aims to improve the learning efficiency and prediction performance by learning multiple objectives from a shared representation. Each attention branch makes separate predictions via a *softmax* to to generate their class probabilities and we linearly combine the individual categorical cross-entropies $\mathcal{L}_b$ into a global loss: $\mathcal{L} = \sum_b w_b \mathcal{L}_b$, with $\mathcal{L}_b = -\sum_c y_c^{(b)} \log p_c^{(b)}$. This allows to control the rate of convergence by weighting the contribution of each branch, assigning low weights for those branches with faster convergence to reduce their influence at the initial stages of training and attenuate updates in shallower attention modules. There are previous approaches that propose the use of multi-loss training to address different machine learning tasks such as dense prediction (Kokkinos, 2017), scene understanding (Eigen and Fergus, 2015), natural language processing (Collobert and Weston, 2008) or speech recognition (Huang et al., 2013). However, the model performance is extremely sensitive to the weight selection w_b, that needs an expensive and time-consuming hyper-parametrization process.

Several multi-branch combinations were tested by applying multidimensional cross-validation to find the optimum branch locations and multi-loss weights. We used a 2D grid search, validating the $att0$ and $att1$ loss weights between a range of 0.5 to 1 with a step size of 0.1, and using the validation loss as monitor. The best performance was achieved with $att0$, $att1$ weights $w_0 = 1$, $w_1 = 0.8$, slightly reducing the contribution of the deeper attention modules and decreasing their overfitting tendency while the shallower branches are still learning.
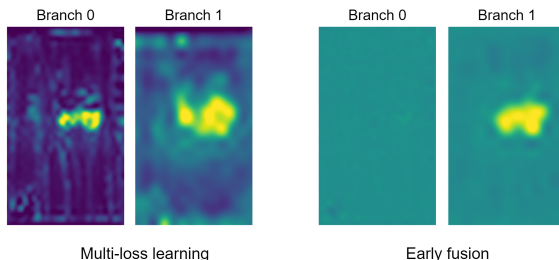
Figure 3: Comparison between merging solutions in the VGG-16 pipeline, visualizing the generated masks in the attention branches $att0$ and $att1$ and observing a large improvement in the shallower modules using multi-loss.

3.4. Public Knee OA Datasets

The data used for this work are bilateral PA fixed flexion knee X-ray images. The datasets are from the Osteoarthritis Initiative (OAI) and Multicenter Osteoarthritis Study (MOST) in UCSF, being standard public datasets widely used in knee OA studies. The baseline cohort of the OAI dataset contains MRI and X-ray images of 4,476 participants. From this entire cohort, we selected 4,446 X-ray images based on the availability of KL grades for both knees as per the assessments by Boston University X-ray reading center (BU). The MOST dataset includes lateral knee radiograph assessments of 3,026 participants. From this, 2,920 radiographs are selected based on the availability of KL grades for both knees as per baseline assessments. As a pre-processing step, all the X-ray images are manually split in the middle, generating two vertical sections from the left and right sides, resizing them to a fixed mean size of 320×224 pixels by keeping the average aspect ratio. Histogram equalization is performed for intensity level normalization, and eventually data augmentation is applied by performing horizontal right-left flips to generate more training data. The training, validation, and test sets were split based on the KL grades distribution. A 70-30 train-test split was used and 10% of the training data was kept for validation.

3.5. Training

All models were trained from scratch using categorical cross-entropy with the ground truth KL grades. Regarding multi-branch training, all target data were duplicated for each attention branch. We used Adam (Kingma and Ba, 2014) with a batch size of 64, $\beta_1 = 0.9$, $\beta_2 = 0.999$, an initial learning rate of 10^{-5} scaled by 0.1 every 2 epochs without improvement in validation loss, and early stopping after 3 epochs without improvement.

4. Results

Although the attention mechanism can be integrated to any CNN pipeline, not all the architectures are well-suited for assessing knee OA severity. We explored several architectures in the literature including state-of-the-art models from previous works of Antony et al. (Antony et al., 2017) and more complex architectures such as ResNet-50 (He et al., 2016), and we analyzed their performance. We found that the same level of abstraction in $att1$ and $att2$ for Antony et. al models can be achieved in shallower branches $att0$, $att1$ for deeper

architectures, implying the best branch location depends on model complexity. After testing different branch combinations, the best performing locations presented in Table 1, and detailed in the Tables 3, 4, 5, specifying the output resolution of their convolutional blocks and the location of the attention branches. From the evaluation of multi-loss models, since every single attention branch produces an independent prediction, the top performing one is used at test time. A more sophisticated ensemble approach was considered but not included in this paper. This approach involves averaging the pre-activation outputs (i.e. values before the softmax) of each of the model branches and then passing the result through a softmax to enforce a valid probability distribution over the classes. This idea is often effective in test time data augmentation and ensemble methods and may improve performance over the single best model referred here. As Table 1 shows, the VGG-16 attention branch $att0$ with multi-loss learning achieved the best overall classification accuracy (64.3%).

Table 1: Evaluation overview for different CNN pipelines.

Models	Antony Clsf.	Antony Extended	ResNet-50	VGG-16
$att0$	41.76%	40%	59.3%	62.2%
$att1$	41.9%	53.26%	58.67%	61.7%
$att2$	44.53%	53.83%	56.47%	58.8%
Early fusion	52.5%	56.17%	59%	63%
Multi-Loss	$att1$: 43.9% $att2$: 46.6%	$att1$: 55.61% $att2$: 55.68%	$att0$: **60%** $att1$: 56.88%	$att0$: **64.3%** $att1$: 63.2%

We also compared the attention mechanism with related knee OA classification-based solutions in the literature. First, we retrained the Antony et. al models with the same training data from the previous experiments, applying the FCN introduced by the authors to address the knee joints extraction (Antony et al., 2017). The results (Table 2) show that the attention-based models with end-to-end architectures clearly outperform the state-of-the-art frameworks. We further compared our results to human level accuracy using the radiologic reliability readings from the OAI (Klara et al., 2016). Although the data used to compute the reliability grading does not match our test set, we followed the methodology of previous works in the literature (Tiulpin et al., 2018), with the aim to dispose of a panoramic view of the current gold standard for diagnosing OA involving human performance. Cohen's *kappa* coefficient (Cohen, 1960) was used to evaluate the agreement between non-clinician readers and experienced radiologists by classifying N items with C mutually exclusive categories. Considering the following grading: $\kappa < 0.20$ slight agreement, 0.21-0.40 fair, 0.41-0.60 moderate, 0.61-0.80 substantial, and 0.81-1.0 almost perfect agreement, their inter-reader reliability for the KL scores was moderate to substantial, with κ values between 0.5 and 0.8. In the case of automatic assessments, by considering a CNN model as a non-clinician X-Ray reader, we can apply the κ coefficient to evaluate the inter-reader reliability between its predictions and the corresponding ground truth annotations, provided by experienced radiologists. As Table 2 shows, the $att0$ branch of the VGG-16 trained by multi-loss together with the $att1$ branch ($w_1 = 0.8$), improves the reliability of related works with a substantial $\kappa = 0.63$ agreement, reaching the margins of human accuracy.

Table 2: Comparative with related frameworks

	Test Acc.	Test Loss	Kappa	# parameters
Antony Clsf	59%	1.13	0.42	~ 7.51 M (FCN: ~ 212.7 K)
Antony Joint Clsf & Reg	62.29%	0.89	0.48	~ 5.77 M (FCN: ~ 212.7 K)
VGG-16: Multi-loss (*att*0)	**64.3%**	0.817	**0.63**	~ 7.7 M

5. Conclusions

This work proposed a novel end-to-end architecture that incorporates trainable attention modules acting as unsupervised fine-grained ROI detectors. The proposed attention modules can be applied at different levels and scales across the CNN pipeline, helping the network to learn relevant attention patterns over the most informative parts of the image at different resolutions. The results obtained for the public knee OA datasets OAI and MOST were satisfactory despite having a considerable scope for further improvement.

The proposed attention mechanism can be easily integrated to any convolutional neural network architecture, being adaptable to any input convolutional volume. However, after exploring different off-the-shelf base models for classification with different complexities, we observed that the best performance is achieved in those models with a balanced ratio between the complexity of the overall architecture and the depth of the convolutional volumes, avoiding overfitting while getting abstraction in the local features used to train the attention modules. On the other hand, we propose the use of multi-loss training to manage the training of multiple attention branches with different velocities of convergence at the same time, boosting the overall performance by fusing attention features with different levels of abstraction. The best performance was achieved by slightly reducing the contribution of the deepest attention branches, improving then the precision of the shallower attention masks and reaching the effectiveness of related approaches with a test accuracy of 64.3% and Kappa agreement of 0.63. Although our method does not surpass the state-of-the-art and could be interpreted as challenging to implement, the overall aim was to reduce the training complexity using an end-to-end architecture. As mentioned in Section 4, without an end-to-end design, the models require a localization step to focus the classifier to the knee joint regions of interest. For instance, previous work of Antony et. al. (Antony et al., 2017) needed a manual annotation process for training a FCN to automatically segment the input knee joints. Our approach, in contrast, requires no such annotation of knee joint locations in the training data. Finally, we observed that localizing the knee joints in an unsupervised way can reduce performance by adding noise in the attention masks and thus into the overall process. A more robust attention module can improve the results and have a bigger impact in the future. As future work, it may be interesting to design better base networks for the attention mechanism and then to test new fine-grained methods in the state-of-art, with the aim to improve the performance of the attention modules towards reducing their dependence on the complexity of the base model.

Acknowledgments

References

Joseph Antony, Kevin McGuinness, Noel E O'Connor, and Kieran Moran. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 1195–1200. IEEE, 2016.

Joseph Antony, Kevin McGuinness, Kieran Moran, and Noel E O'Connor. Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 376–390. Springer, 2017.

Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *European conference on computer vision*, pages 584–599. Springer, 2014.

Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

Yu-hsin Chen, Ignacio Lopez-Moreno, Tara N Sainath, Mirkó Visontai, Raziel Alvarez, and Carolina Parada. Locally-connected and convolutional neural networks for small footprint speaker recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Behzad Heidari. Knee osteoarthritis prevalence, risk factors, pathogenesis and features: Part i. *Caspian journal of internal medicine*, 2(2):205, 2011.

Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7304–7308. IEEE, 2013.

Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. *arXiv preprint arXiv:1311.3715*, 2013.

JH Kellgren and JS Lawrence. Radiological assessment of osteo-arthrosis. *Annals of the rheumatic diseases*, 16(4):494, 1957.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kristina Klara, Jamie E Collins, Ellen Gurary, Scott A Elman, Derek S Stenquist, Elena Losina, and Jeffrey N Katz. Reliability and accuracy of cross-sectional radiographic assessment of severe knee osteoarthritis: role of training and experience. *The Journal of rheumatology*, pages jrheum–151300, 2016.

Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *CVPR*, volume 2, page 8, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 8(1): 98–113, 1997.

YongPing Li, XiaoChun Wei, JingMing Zhou, and Lei Wei. The age-related changes in cartilage and osteoarthritis. *BioMed research international*, 2013, 2013.

Xiao Liu, Tian Xia, Jiang Wang, Yi Yang, Feng Zhou, and Yuanqing Lin. Fully convolutional attention networks for fine-grained recognition. *arXiv preprint arXiv:1603.06765*, 2016.

Kevin Mader. Attention on pretrained-vgg16 for bone age. https://www.kaggle.com/kmader/attention-on-pretrained-vgg16-for-bone-age, 2018.

H Oka, S Muraki, T Akune, A Mabuchi, T Suzuki, H Yoshida, S Yamamoto, K Nakamura, N Yoshimura, and H Kawaguchi. Fully automatic quantification of knee osteoarthritis severity on plain radiographs. *Osteoarthritis and Cartilage*, 16(11):1300–1306, 2008.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Lior Shamir, Nikita Orlov, D Mark Eckley, Tomasz Macura, Josiah Johnston, and Ilya G Goldberg. Wndchrm–an open source utility for biological image analysis. *Source code for biology and medicine*, 3(1):13, 2008.

Lior Shamir, Shari M Ling, William Scott, Marc Hochberg, Luigi Ferrucci, and Ilya G Goldberg. Early detection of radiographic knee osteoarthritis using computer-aided analysis. *Osteoarthritis and Cartilage*, 17(10):1307–1312, 2009a.

Lior Shamir, Shari M Ling, William W Scott Jr, Angelo Bos, Nikita Orlov, Tomasz J Macura, D Mark Eckley, Luigi Ferrucci, and Ilya G Goldberg. Knee x-ray image analysis method for automated detection of osteoarthritis. *IEEE Transactions on Biomedical Engineering*, 56(2):407–415, 2009b.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Aleksei Tiulpin, Jérôme Thevenot, Esa Rahtu, Petri Lehenkari, and Simo Saarakkala. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. *Scientific reports*, 8(1):1727, 2018.

Jun Wei, Heang-Ping Chan, Chuan Zhou, Yi-Ta Wu, Berkman Sahiner, Lubomir M Hadjiiski, Marilyn A Roubidoux, and Mark A Helvie. Computer-aided detection of breast masses: Four-view strategy for screening mammography. *Medical physics*, 38(4):1867–1876, 2011.

Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850, 2015.

Tae Keun Yoo, Deok Won Kim, Soo Beom Choi, and Jee Soo Park. Simple scoring system
and artificial neural network for knee osteoarthritis risk prediction: a cross-sectional study.
PloS one, 11(2):e0148724, 2016.

Appendix A. CNN Architectures, Learning Curves and Visualizations

Table 3: Antony et al. base architecture for classification

Layer	Kernels	Kernel Size	Strides	Output shape
conv1	32	11×11	2	$100 \times 150 \times 32$
pool1	-	3×3	2	$49 \times 74 \times 32$
conv2	64	5×5	1	$49 \times 74 \times 64$
($att0$) pool2	-	3×3	2	$24 \times 36 \times 64$
conv3	96	3×3	1	$24 \times 36 \times 96$
($att1$) pool3	-	3×3	2	$11 \times 17 \times 96$
conv4	128	3×3	1	$11 \times 17 \times 128$
($att2$) pool4	-	3×3	2	$5 \times 8 \times 128$

Table 4: Antony et al. extended base architecture for classification and regression

Layer	Kernels	Kernel Size	Strides	Output shape
conv1	32	11×11	2	$100 \times 150 \times 32$
pool1	-	3×3	2	$49 \times 74 \times 32$
conv2-1	64	3×3	1	$49 \times 74 \times 64$
conv2-2	64	3×3	1	$49 \times 74 \times 64$
($att0$) pool2	-	3×3	2	$24 \times 36 \times 64$
conv3-1	96	3×3	1	$24 \times 36 \times 96$
conv3-2	96	3×3	1	$24 \times 36 \times 96$
($att1$) pool3	-	3×3	2	$11 \times 17 \times 96$
conv4-1	128	3×3	1	$11 \times 17 \times 128$
conv4-2	128	3×3	1	$11 \times 17 \times 128$
($att2$) pool4	-	3×3	2	$5 \times 8 \times 128$

Table 5: ResNet-50 base architecture for classification

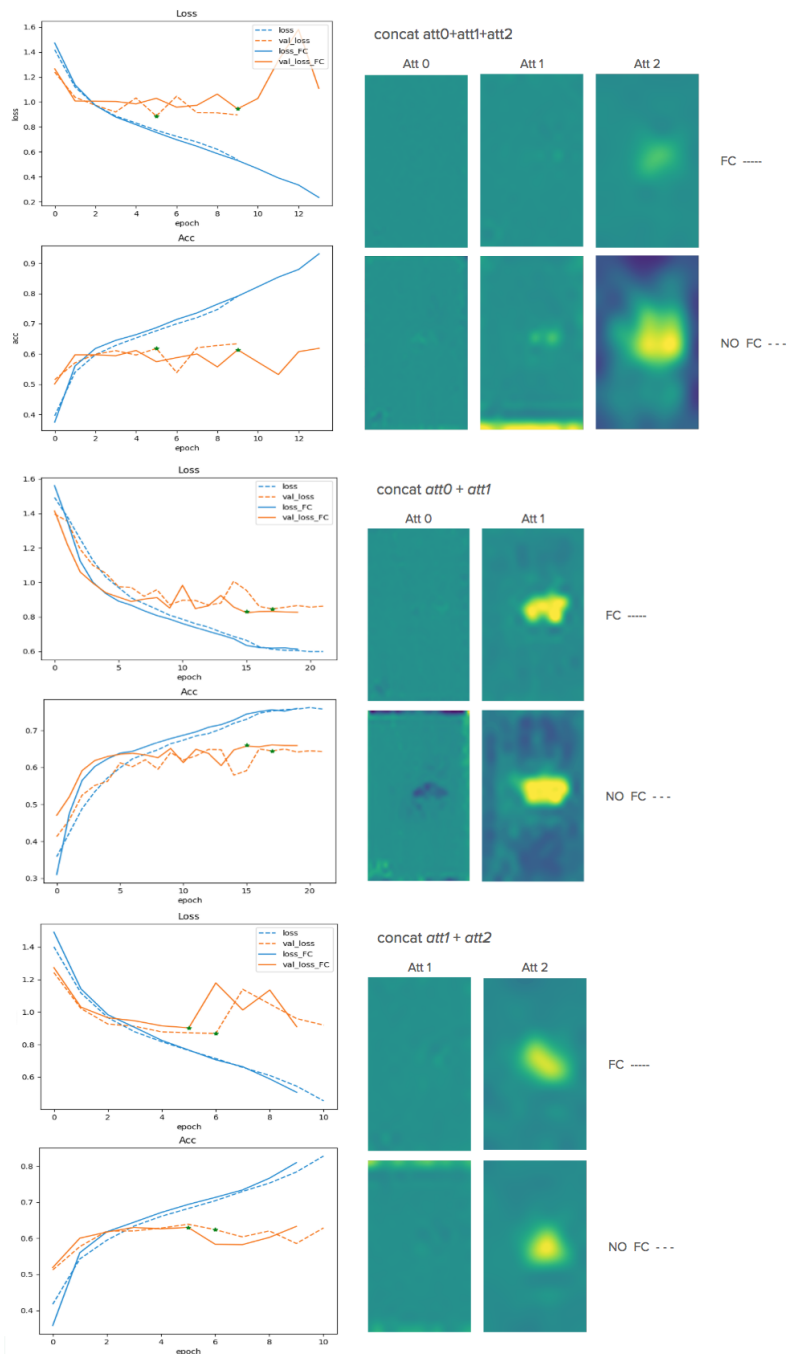| Layer	Kernels	Kernel Size	Strides		Output shape
conv1	64	7×7	2		$224 \times 224 \times 64$
maxpool	-	3×3	2		$112 \times 112 \times 64$
conv2_$\times$	64	1×1	1	$(\times 3)$	$56 \times 56 \times 256$
	64	3×3	1		
	256	1×1	1		
$(att0)$ conv3_$\times$	128	1×1	2	$(\times 4)$	$28 \times 28 \times 512$
	128	3×3	1		
	512	1×1	1		
$(att1)$ conv4_$\times$	256	1×1	2	$(\times 6)$	$14 \times 14 \times 512$
	256	3×3	1		
	1024	1×1	1		
$(att2)$ conv5_$\times$	512	1×1	2	$(\times 3)$	$7 \times 7 \times 512$
	512	3×3	1		
	2048	1×1	1		

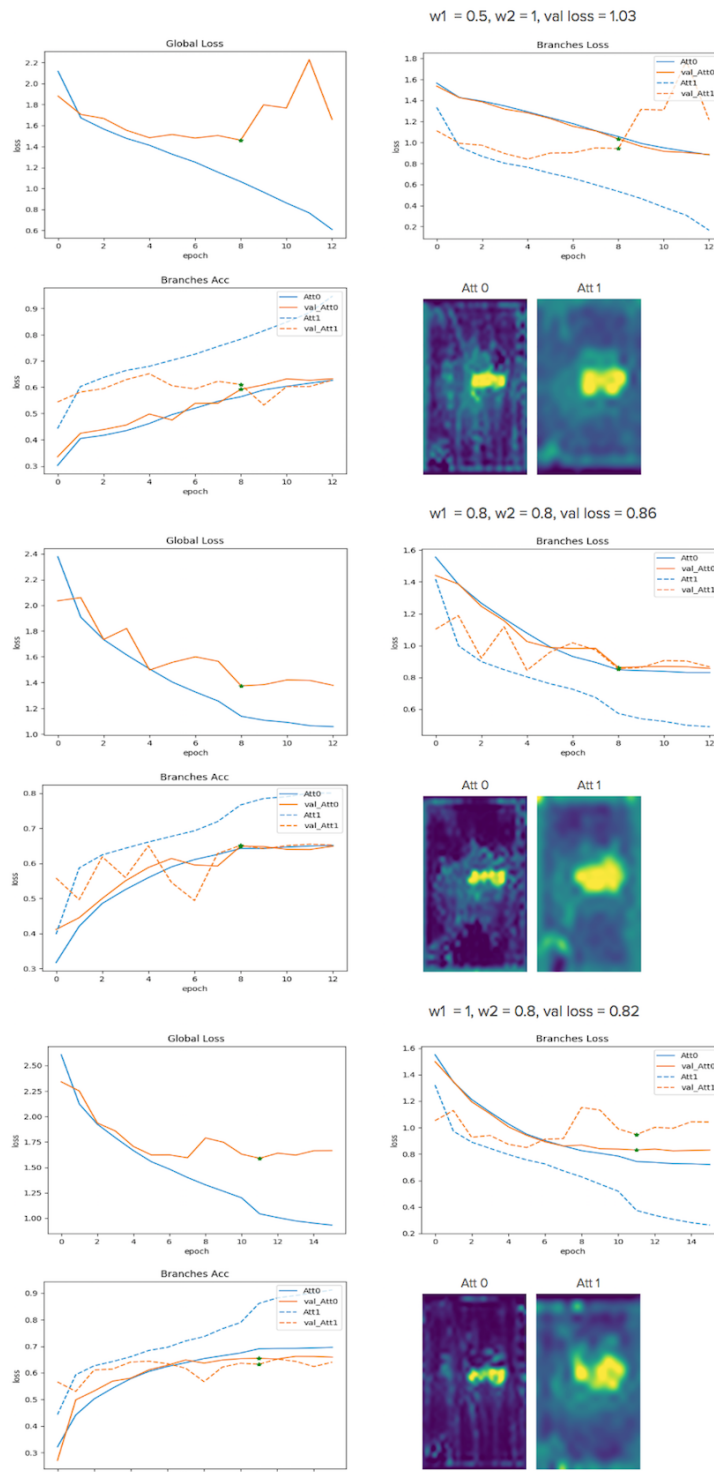Figure 4: Learning curves and visualization for early fusion experiment in VGG-16.

Figure 5: Learning curves and visualization for multi-loss experiment in VGG-16.
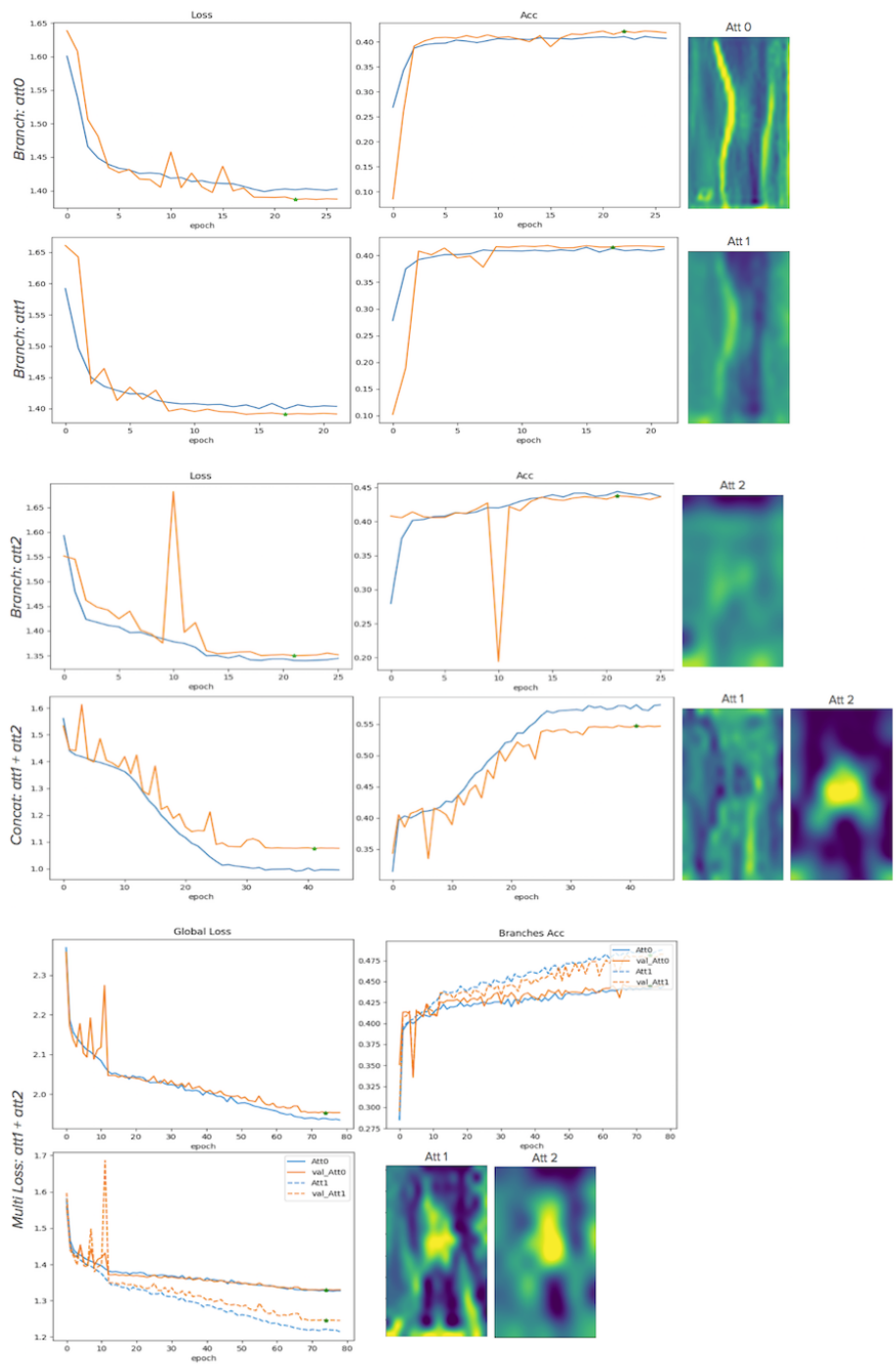
Figure 6: Learning curves and visualization for Antony et al. pipeline for classification.
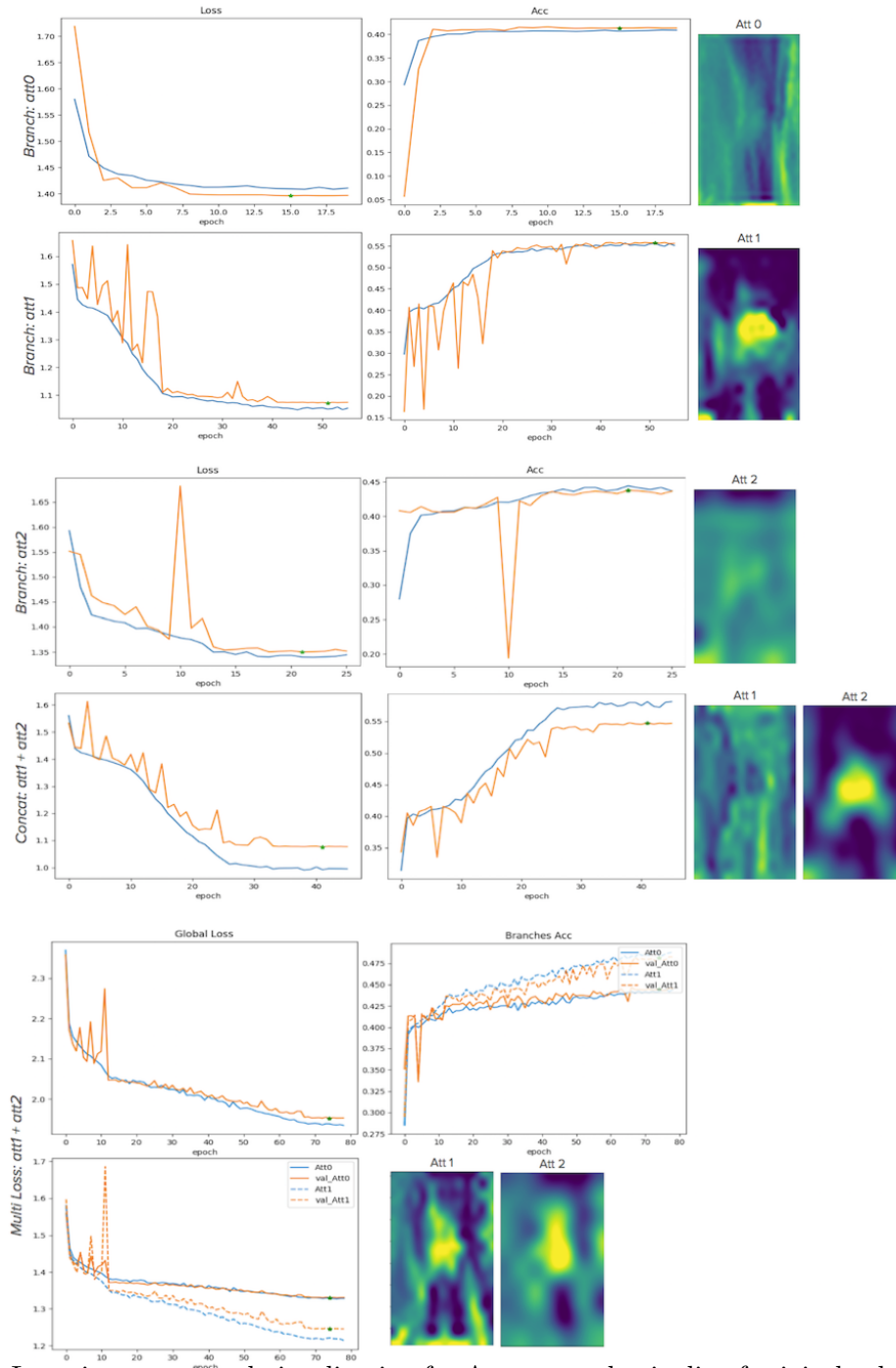
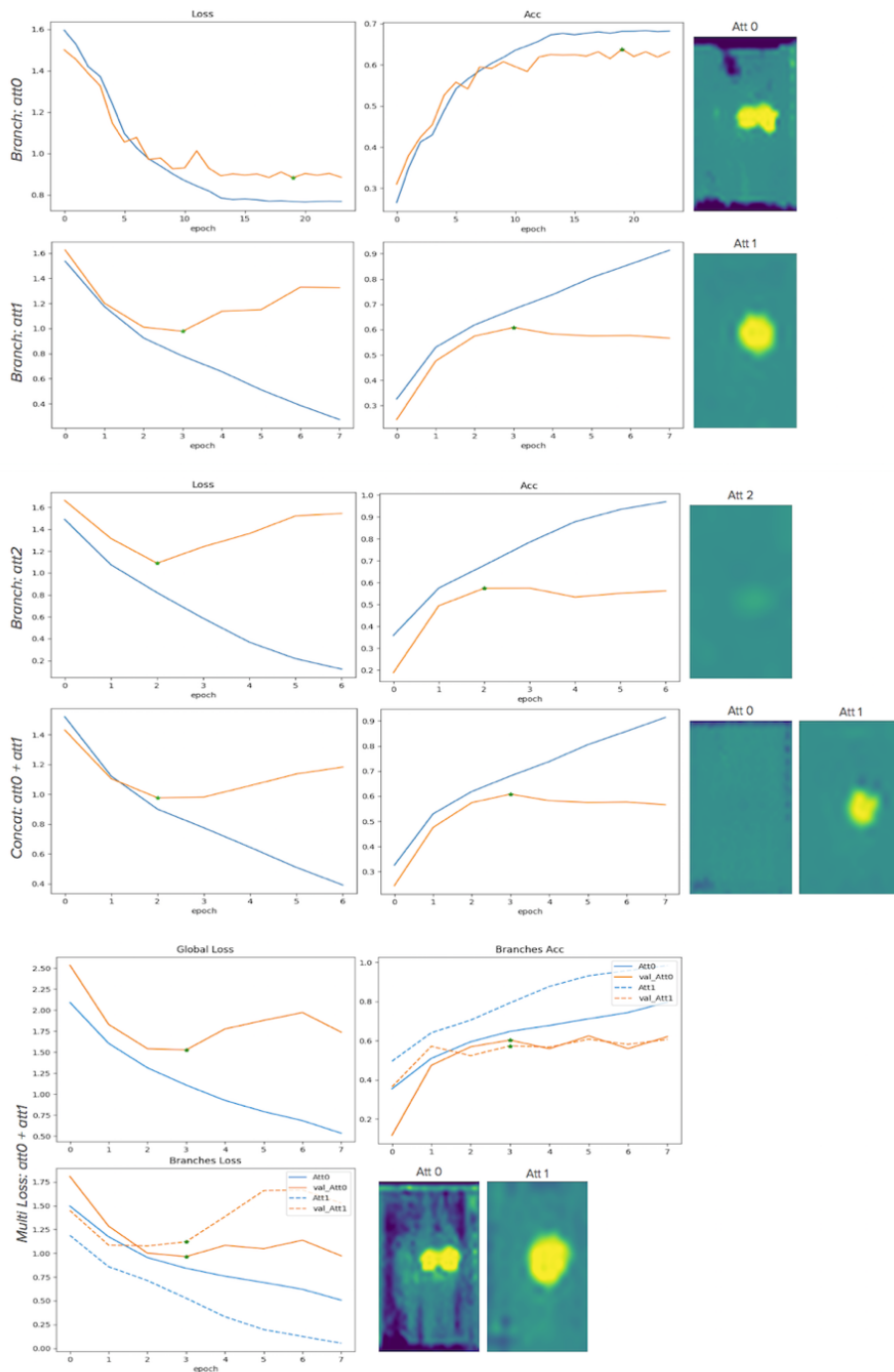Figure 7: Learning curves and visualization for Antony et al. pipeline for jointly classification and regression.

Figure 8: Learning curves and visualization for ResNet-50 pipeline.