They demonstrated

Attention-based learning was first introduced by Xu Et Al for the purposes of generating image captions [7]. They proposed that image captioning tasks could be enhanced through hard attention, which is focused on the expected value of specific words, and soft attention which focuses on the expected value of context vectors. The proposed attention mechanism offered an ability to bring specific regions of an image to the fore for multi-object detection. As multi-label classification involves identifying labels associated with different pixel groupings and configurations extending attention to multi-label classification seemed appropriate.

With respect to image classification

The use of attention modules in image classification was explored further by Wang Et Al, who (at the time) demonstrated that attention modules could be embedded into very deep neural networks to achieve state of the art performance on CIFAR-10, CIFAR-100 and ImageNet LSVRC 2012 [6]. Rather than encoding in context vectors, they defined attention layers as feature masks. The feature masks are built up from image convolutions which are used to capture regions of an image that are relevant to both foreground and background features.

labels

took a different approach to attention modules. They challenged the ability of convolutional layers to identify context-dependent labels because CNNs are built from pixel neighbourhoods.

Bello et Al explore how attention modules can be used to idenitfy contextual information, whereas standard convolutional layers focus on pixel regions/neighbourhoods [2]. Rather than defining attention as a form of pixel map, they define attention as an augmented convolutional layer with learnable parameters. To capture the various regions of the image, queries, keys and values are utilized, which is often used in encoder-decoder architectures to create context vectors to represent the input for decoder steps [5]. In that sense multi-headed attention can be used to encode different features of an image to be used in downstream feature extraction.

that capture the relationships between different pixel regions. This is an example of multi-headed attention that demonstrate resuls.

significant

Despite there being constant research into attention modules, there has been limited analysis of their application to multi-label problems compared to single-label problems. In 2015, Ba Et Al [1]. They combined convolutions with recurrent models to capture a ”glimpse”, or image feature. The glimpses are then forward propagated so the final layer classification layer contains all the previously learned information. This approach was successfully tested for constrained multi-label tasks such as house number identification that don't require the context specific feature extraction explored by Bello and Vaswani.

Applying attention modules to more complicated object detection was recently completed by Wei Et Al [3]. They propose an adaptive attention mechanism where max pools are used to identify small, region specific objects and average pools are used to identify less region specific/background objects. The weighting of these pooled units is learnt during training and is thus data specific. When the module was into YOLOv3 and MobileNetV2 and tested on KITTI and PASCAL_VOC accuracy improved significantly.

was incorporated, accuracy improved by **

cross-attention, correlated features.

feature and label covariance is used to support the binary classification of each label

It's also worth mentioning some of the concerns and proposed limitations of attention networks on their own. The state of the art for NUS utilizes a cross-attention to identify features that are correlated or related [4]. This is then fed into a transformer architecture where the relationship between features and labels are encoded and utilized for binary classification. Exploring transformer architectures is outside the scope of this report but incorporating attention modules into transformers seems to support the development of generalizable multi-image classifiers.

but based on the Query2Lablel architecture.

References

[1] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention, 2015.

[2] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. Attention augmented convolutional networks, 2020.

[3] Wei Li, Kai Liu, Lizhe Zhang, and Fei Cheng. Object detection based on an adaptive attention mechanism. *Scientific Reports*, 10(1):11307, 2020.

[4] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification, 2021.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[6] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification, 2017.

[7] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2016.