Attention-based learning was first introduced by Xu Et Al for the purpose of image caption generation[7]. They developed hard-attention modules which focused on the expected value of specific captions and soft attention modules which encoded different levels of information in encoder-decoder architectures. When captions were idenfified, both attention mechanism allowed for different pixel regions to be brought to the fore. As multi-label classification involves assigning global labels based on different pixel configuration extending attention to image classification is logical.

With respect to image classification Wang Et Al (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 attention 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.

Bello et Al departed from the use of feature masks and standard convolutions because of their focus on specific pixel neighbourhoods [2]. They defined attention as an augmented convolutional layer with learnable parameters. They proposed a multi-headed approach to attention, where multiple linear transformations are concatenated into one context vector that captures the various regions of an image. This approach is more aligned with the encoder-decoder architectures initially suggested by Xu Et Al, and popularized by Vaswani et al in 2017 [5].

Despite there being constant research into attention modules, most of the research has focused on single-label classification. In 2015, Ba Et Al combined convolutions with recurrent models to capture a "glimpse", or image feature [1]. 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.

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 version of multi-headed 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.

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.