**TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification**

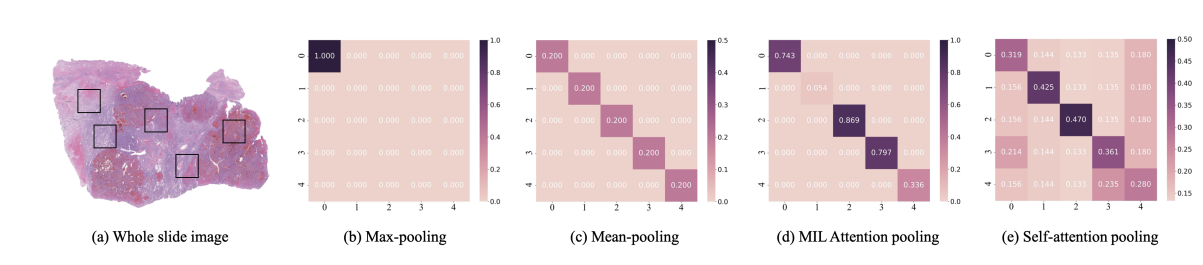
**What is Whole slide image?**

* The advent of whole slide image (WSI) scanners, which convert the tissue on the biopsy slide into a gigapixel image fully preserving the original tissue structure, provides a good opportunity for the application of deep learning in the field of digital pathology.
* However, the deep learning-based biopsy diagnosis in WSI has to face a great challenge due to the huge size and the lack of pixel-level annotations.
* To address this problem, multiple instance learning (MIL) is usually adopted to take diagnosis analysis as a weakly supervised learning problem.

**What is TransMIL?**

* MIL (Multiple Instance Learning) - is a powerful tool to solve the weakly supervised classification in whole slide image (WSI) based pathology diagnosis.
* However, the current MIL methods are usually based on independent and identical distribution hypothesis, thus neglect the correlation among different instances.
* To address this problem, we proposed a new framework, called correlated MIL, and provided a proof for convergence.
* Based on this framework, we devised a Transformer based MIL (TransMIL), which explored both morphological and spatial information. The proposed TransMIL can effectively deal with unbalanced/balanced and binary/multiple classification with great visualization and interpretability.

**Different Types of pooling?**

* ****

The difference between different Pooling Matrix P. Suppose there are 5 instances sampled from WSI in (a), P ∈ R 5×5 is the corresponding Pooling Matrix, where the values in the diagonal line indicate the attention weight for itself and the rest indicate correlation between different instances. (b,c,d) all neglect the correlation information, hence the P is diagonal matrix. In (b), the first instance was chosen by Max-pooling operator, so there is only one non-zero value in the first diagonal position. In (c), all the values within diagonal line are the same due to the Mean-pooling operator. In (d), the values within diagonal line can be varied due to the introduction of bypass attention. (e) obeys the correlation assumption, so there are non-zero values in off-diagonal position indicating correlation between different instances.