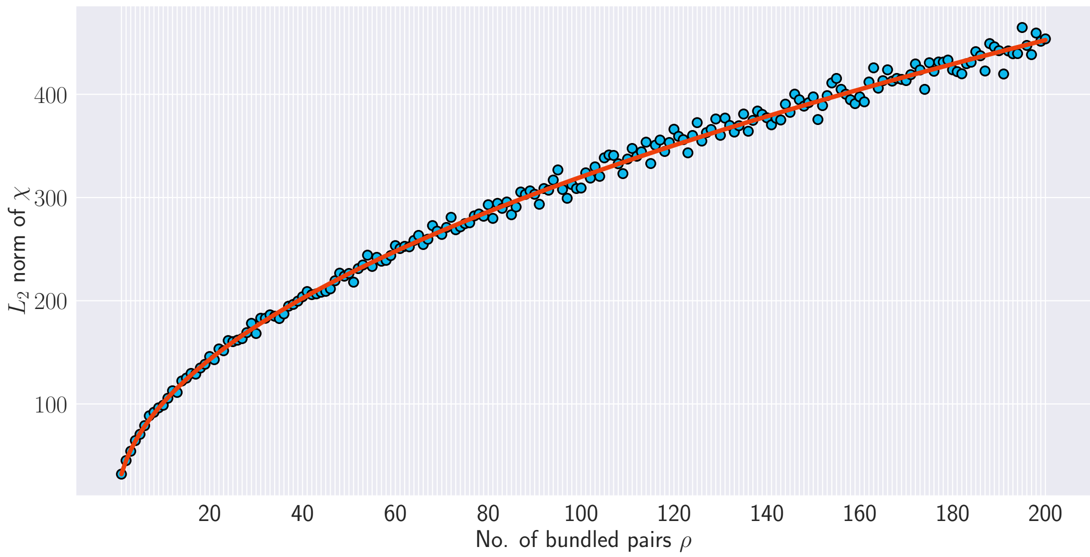

Figure 1: L_2 norm of the composite representation χ for the number of bundled vector pairs ρ varied from 1 to 200. The figure shows the L_2 norm of χ can be approximated to $\sqrt{d \cdot \rho}$ with a R-square value of 0.9865. Hence, we can estimate the number of bundled pairs from the norm of the composite representation.

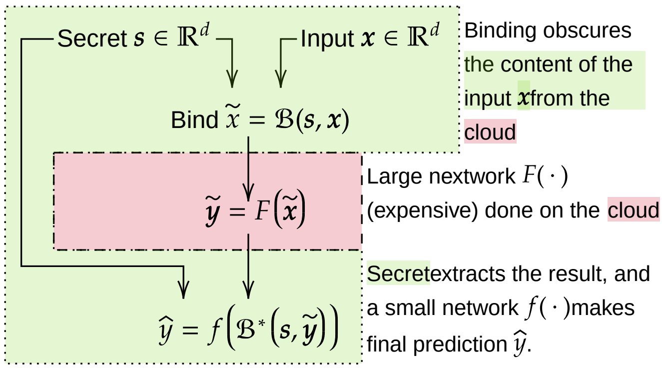

Figure 2: Diagram of how CSPS works, which will be added to the relevant section.

For XML classification, we have a set of K classes that will be present for a given input, where $K \approx 10$ is the norm. Yet, there will be L total possible classes where $L \geqslant 100,000$ is quite common. Forming a normal linear layer to produce L outputs is the majority of computational work and memory use in standard XML models, and thus the target for reduction. A VSA can be used to side-step this cost, as shown by [8], by leveraging the symbolic manipulation of the outputs. First, consider the target label as a vector $\boldsymbol{s} \in \mathrm{R}^d$ such that $d \ll L$. By defining a VSA vector to represent "present" and "missing" classes as $\mathbf{p}$ and $\mathbf{m}$, where each class is given it's own vector $\boldsymbol{c}_{1,\ldots,L}$, we can shift the computational complexity form $\mathcal{O}(L)$ to $\mathcal{O}(K)$ by manipulating the "missing" classes as the compliment of the present classes:

$$\boldsymbol{s} = \overbrace{\sum_{i \in y_i = 1} \mathcal{B}(\boldsymbol{p}, \boldsymbol{c}_i)}^{\text{Labels Present}\mathcal{O}(dK)} + \overbrace{\sum_{j \in y_j = -1} \mathcal{B}(\boldsymbol{m}, \boldsymbol{c}_j)}^{\text{Labels Absent}\mathcal{O}(dL)} = \overbrace{\mathcal{B}\left(\boldsymbol{p}, \left(\boldsymbol{a} =: \sum_{i \in y_i = 1} c_i\right)\right)}^{\text{Labels Present}\mathcal{O}(d\ \ K)} + \overbrace{\mathcal{B}\left(\boldsymbol{m}, \left(\boldsymbol{a} - \sum_{i \in y_i = 1} c_i\right)\right)}^{\text{Labels Absent}\mathcal{O}(dK)}$$

Similarly, the loss to calculate the gradient can be computed based on the network's prediction $\hat{\boldsymbol{s}}$ by taking the cosine similarity between each expected class and one cosine similarity for the representation of all missing classes. The excepted response of 1 or 0 for an item being present/absent from the VSA is used to determine if we want the similarity to be 0 (1-cos) or 1 (just cos), as shown in the below question.

$$loss = \overbrace{\sum_{i \in y_i = 1} \left(1 - \cos\left(\mathcal{B}^*(\boldsymbol{p}, \hat{\boldsymbol{s}}), \boldsymbol{c}_i\right)\right)}^{\text{Present Classes } \mathcal{O}(d\ \ K)} + \overbrace{\cos\left(\mathcal{B}^*(\boldsymbol{m}, \hat{s}), \sum_{i \in y_i = 1} \boldsymbol{c}_i\right)}^{\text{Asbsent classes } O(d\ \ K)}$$