- Training with label smoothing : minimize cross-entropy between y_k^{LS} and p_k instead of y_k and p_k where (soft label) $y_k^{LS} = y_k(1-\alpha) + \alpha/K$, parameterised by α.
- Training with LS encourages logits of correct and incorrect classes be separated by a constant distance parameterised by α, but with hard labels, the logits of correct class tend to be higher than all others and allows others to be different from each other.
- $x^T.w_k$ can be looked as a measure of L2 distance between x and w_k ($x^T.x + w_k^T.w_k - 2*x^T.w_k$) where $x^T.x$ is factored out during softmax, and generally $w_k^T.w_k$ are constant across classes.
- Therefore, label smoothing encourages the activations of the penultimate layer to be close to the template of the correct class and equally distant to the templates of the incorrect classes.
- For visualising the activations of penultimate layer, the projections of these activations onto a plane crossing 3 of the class templates (w_k) were used.
    - The projections with LS of each class were tightly clustered and more equidistant from incorrect class templates compared to without LS.Despite the differences, accuracies were close.
    - Scales of activations without LS were higher because the model was overconfident in predicting, but with LS, the distances are constrained.
    - Semantically similar classes were more separated with LS, and formed an arc shape w.r.t the 3rd class showing equal distance.
    - Without LS we could measure how close a particular instance is to its incorrect classes cluster, but after applying LS, that information is erased.
- Calibration is a measure of how good the predicted probabilities actually represent the confidence values.Predicted probabilities that match the expected distribution of probabilities for each class are referred to as calibrated.
- Reliability curves : Relative counts of positive examples with predicted probability in the range of each bin is plotted. Perfectly calibrated model should be x=y line. Temperature (Platt) scaling was used for calibration by multiplying the logits with a scalar (learnable parameter) before softmax.
- LS also gave similar effects on reliability curve as platt scaling and calibrates the network. Despite trying to form tight clusters on training penultimate activations, the validation activations are spread towards the centre giving confidence values for all classes.
- Though calibration does not affect accuracy directly in image classification, it affects the beam search algorithm which depends on the soft output of the model(Transformer).
- Knowledge distillation (training smaller network to emulate the layer activations of larger network to get teacher guided student convergence) is worse when LS is used.
- In knowledge distillation, we replace the cross-entropy term H(y, p) by the weighted sum $(1 - \beta)H(y, p) + \beta H(p^t(T), p(T))$, where $p_k(T)$ and $p_k^t(T)$ are the outputs of the student and teacher after temperature scaling with temperature T, respectively
- Distilled models performed better than student baseline with LS, but for some α, the student performed better as relative information between logits was erased when teacher was trained with LS.

- Distillation with LS gave worse results than distillation with temperature scaling. The information erasure can be seen in the visualisations where LS constrains all examples of each class to be together and hence removing info of how far an instance is to other classes.
- The estimated mutual information (difference between logits of 2 specific classes) plotted against number of iterations shows that with LS, the difference increases rapidly at first, but then keeps decreasing, supporting the observation in the visualisations.
- For visualising machine translation, the next token prediction is taken as equivalent to image classification. But it has some differences :
    - Next token class distributions are highly imbalance.
    - Token prediction accuracy is very less compared to image classification accuracy, so there may be errors in visualisation and examples may tend to form tighter clusters.
    -