Summary of ‘Contrastive Representation Distillation’, Kian et al.

Knowledge Distillation (KD) is defined as the transfer of knowledge from one deep learning model to another. In this paper, the initial larger deep learning model is called the teacher and the knowledge is transferred to a student model. Transferring knowledge about the ‘representation’ that a model follows is oftentimes tougher. This might include decreased representation of a larger model with no change in generalization, or transfer in modality-based representation i.e., image to sound or sound to text etc.

When the authors say that representation knowledge is structured, they mean the dimensions of the representation exhibit complex interdependencies. Due to these interdependencies, the dimensions cannot be effectively transferred with the assumption that they are free and independent. Any change in one dimension affects others, so better methods for effective knowledge transfer and generalization had to be constructed.

To overcome this interdependency problem with representation transfer, the authors have used an objective that accounts for the correlations and higher-order output dependencies. This family of objectives that are capable of capturing these are called contrastives.

In the paper, the authors have considered three modes of knowledge transfer

* compression of a model (larger to smaller)
* cross-modal transfer (modality to modality)
* distilling an ensemble of models into a single model (ensemble to single)

How contrastive objective works -

It encourages the teacher and student models to map the same inputs for close representations and different inputs for distant representations. The contrastive objective is to maximise the lower bound of the mutual information shared by the teacher and the student. The key idea in contrastive objective-based learning is to learn a representation that is close in some metric space and push apart the representations for negative pairs.

Conclusion -

The paper contributes to the knowledge distillation and transfer problem with a contrastive-based objective (CRD) for transferring knowledge between deep neural network models and it has been found to outperform the state-of-the-art knowledge distillation methods. CRD has immense application in model compression, cross-modal transfer and ensemble distillation.

Knowledge Distillation objective as proposed before by Hinton et al. consistently performs second best and prior objectives only surpass KD when combined with KD.

Maximising Lower Bound over Mutual Information between Teacher and Student

The method employed by the authors in maximising this mutual information between the student and teacher representations is by maximising the KL divergence between the joint distribution of the two representations and, the product of the marginal distributions of the probability of the two representations.

A new representation is assumed to have a binary latent variable C where C=0 is when the input tuple is drawn from the product of marginals or C=1 when it is drawn from the joint distribution of representations.

With these definitions established, the mutual information bounds are calculated. Subsequently, the objective losses for the three applications are determined. These losses help in training and inference of the respective student models.

The authors approximate the true values of representations in a critic value h, and the objective is to learn a representation that maximises/optimizes the critic’s value. In practice, a memory buffer is maintained that contains the features of data samples of all previous batches and during training, sampling is done from this memory buffer. This helps in loosening the bound over mutual information.