0.1.1 Deep Aligned Representations

The motivation for the knowledge distillation driven approach in this work is provided by the paper "See, Hear, and Read: Deep Aligned Representations" by Aytar et al. (2017) [1]. For simplicity, this paper will be referred to as "SHRe" (for **S**ee, **H**ear, **Re**ad) in this work.

In SHRe, the authors propose a method to align representations of image, text, and audio through knowledge distillation from a supervised image model. The student model is a multimodal model with separate modality-specific encoders for image, text, and audio, with a shared encoder on top. The approach utilizes 1D convolutions for audio and text, and 2D convolutions for images. The output feature maps of these encoders are flattened and then passed separately through a shared encoder, consisting of 3 linear layers [1]. The approach is generally independent of the specific architecture of the components (encoders), meaning that any architecture can be used. Notice how the aforementioned exactly matches the definition of a multimodal model as defined in (TODO: cite multimodal_models).

The teacher model was trained in a supervised manner, with the authors utilizing a model pretrained on ImageNet-1K, though it is not specified what exact model was used. The training objective is to minimize the KL-Divergence between the teacher and student models.

Specifically, the method involves using image-text $\{\boldsymbol{x}^v, \boldsymbol{x}^w\}$ and image-audio $\{\boldsymbol{x}^v, \boldsymbol{x}^a\}$ pairs. $\boldsymbol{x}^v$ is a 2D image, $\boldsymbol{x}^w$ a sequence of text tokens, and $\boldsymbol{x}^a$ a spectrogram of audio. For each pair, the image $\boldsymbol{x}^v$ is passed through the teacher model $g(\cdot)$, producing a probability distribution over the ImageNet-1K classes (1000 classes), denoted as $g(\boldsymbol{x}^v)$. The same image $\boldsymbol{x}^v$ is also passed through the image encoder $f_v(\cdot)$ of the student model, followed by the shared encoder $s(\cdot)$, also resulting in a probability distribution over the ImageNet-1K classes, defined as $s(f_v(\boldsymbol{x}^v))$.

The other part of the pair, for example, the text $\boldsymbol{x}^w$ in an image-text pair, is passed through the text encoder $f_w(\cdot)$ of the student model, and then through the shared encoder $s(\cdot)$. Since the shared encoder is the same as the one used for the image, the output is, again, a probability distribution over the ImageNet-1K classes, represented as $s(f_w(\boldsymbol{x}^w))$.

The probability distribution generated by the teacher model for the image can be compared with the probability distribution produced by the student model for the same image, using KL-Divergence. This is the usual approach to knowledge distillation, defined in (TODO: cite knowledge distillation section). What makes the approach unique, however, is that the probability distribution of the teacher model for the image can be compared with the probability distribution of the student model for the text. For a single image-text pair, the loss is defined as:

$$\mathcal{L}_{\text{KD}}^{vw} = \frac{1}{2} * D_{\text{KL}}(g(\boldsymbol{x}^v) \parallel s(f_v(\boldsymbol{x}^v))) + \frac{1}{2} * D_{\text{KL}}(g(\boldsymbol{x}^v) \parallel s(f_w(\boldsymbol{x}^w))) \tag{1}$$

With D_{KL} being the KL-Divergence. The loss changes accordingly for image-audio pairs, where the probability distribution over audio is defined as $s(f_a(\boldsymbol{x}^a))$.

$$\mathcal{L}_{\text{KD}}^{va} = \frac{1}{2} * D_{\text{KL}}(g(\boldsymbol{x}^v) \parallel s(f_v(\boldsymbol{x}^v))) + \frac{1}{2} * D_{\text{KL}}(g(\boldsymbol{x}^v) \parallel s(f_a(\boldsymbol{x}^a))) \tag{2}$$

The goal of this approach is to make the probability distributions between teacher and student as similar as possible. Since an image and its corresponding text in an image-text pair describe the same real-world concept, the distribution of the teacher model for the image, over the ImageNet-1K classes, can directly be transferred to the caption of the image. That way, the model can learn to output the same probabilities over the ImageNet-1K classes for both the image and the text. This enables the alignment of modalities at the level of real-world objects. The same process can be

applied to image-audio pairs, allowing the model to align representations across multiple modalities. A visualization of this will be shown when we apply this approach in (TODO: Transformer SHRe section).

Even though all modalities share the same shared encoder $s(\cdot)$, the output of the intermediate layers in the shared encoder will still differ for each modality. This is because KL-Divergence only ensures alignment at the level of classes, which corresponds to the output layer (the last fully-connected layer of the shared encoder outputs the probability distribution over ImageNet-1K classes). The internal representations in $s(\cdot)$, meaning the first two layers, can still be different between the modalities of a pair. They can vary, as long as the resulting probability distribution of the last fully-connected/linear layer is the same as the teacher model's output.

However, the shared encoder is meant to have the same internal representation for e.g. an image and its caption/text: Since they describe the same concept, the activations in the shared encoder should be similar, which is, as described in (TODO: cite image-text contrast), crucial for tasks such as retrieval. To achieve this, the authors add a ranking loss to the training, which functions similarly to a contrastive loss. This ranking loss drives the representations of inputs from the same pair closer together, while pushing the representations of inputs from different pairs further apart. It is defined as:

$$\mathcal{L}_{\text{Rank}} = \sum_{i=1}^{B} \sum_{j \neq i} \max\{0, \Delta - \cos(\boldsymbol{x}_i^v, \boldsymbol{x}_i) + \cos(\boldsymbol{x}_i^v, \boldsymbol{x}_j)\} \tag{3}$$

Here, B represents the batch size, $\boldsymbol{x}_i^v$ is an image, and $\boldsymbol{x}_i$ is the corresponding text or audio, depending if an image-text or image-audio pair is used. j iterates over negative samples in the batch ($j \neq i$).

Different from contrastive loss, for a given input, e.g. an image, the ranking loss does not normalize the similarity scores of a positive pair (e.g. image-text) with respect to all other possible pairings (all other texts) for a sample (image) in the batch. The authors did not provide intuitions for the choice of the ranking loss over the contrastive loss, and we can only assume that since the paper was published in 2017 [1], the contrastive loss was not as widely adapted as it is today.

The final loss is a combination of the KL-Divergence loss and the ranking loss:

$$\mathcal{L}_{\text{SHRe}} = \mathcal{L}_{\text{KD}} + \mathcal{L}_{\text{Rank}} \tag{4}$$

The authors evaluate SHRe on retrieval tasks, and the results (Table 1) show that SHRe performs significantly better than a random baseline. Interestingly, even though the model is only trained on image-text and image-audio pairs, the alignment also generalizes to text-audio pairs, and the model can retrieve text-audio pairs, albeit not as well as between the modalities it was trained on [1]. This indicates that the image modality acts as an anchor between text and audio, enabling the model to align representations between modalities it was not explicitly trained on. The alignment between modalities becomes transitive.

Model	MSCOCO		Flickr (Custom)[1]		Unspecified[2]	
	Image ↓ Text	Text ↓ Image	Image ↓ Sound	Sound ↓ Image	Text ↓ Sound	Sound ↓ Text
Random	500	500	500	500	500	500
SHRe	5.8	6.0	47.5	47.8	135.0	140.5

Table 1: Retrieval results of SHRe on different datasets. Each dataset contains 5k sample pairs (e.g. image-text pairs) for evaluation, and is splitted into 5 chunks of 1k samples each. Retrieval is then performed on each chunk, and metric used is the median rank of the correct pair in the ranked list. The median rank is averaged over all chunks for each datasets, so the results seen describe the average median rank over all chunks for each dataset. The results are taken from the SHRe paper [1].

The approach is illustrated in Figure 1. It is important to note that SHRe is only trained with image-text and image-audio pairs, and not, how it might seem from the figure, with image-text-audio triplets.

The SHRe approach is a crucial foundation for this work, as it demonstrates how the knowledge from a **supervised** unimodal (image) model can be *extracted* and *transferred* to a multimodal model.

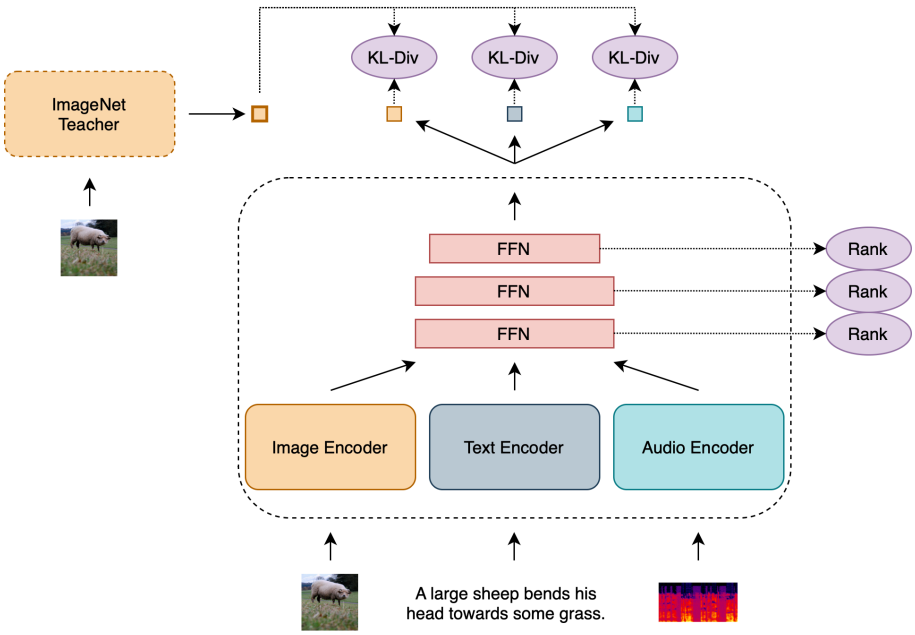

Figure 1: Illustration of the SHRe approach. The model is trained to output the same probability distribution over ImageNet-1K classes between images, image-text pairs, and image-audio pairs. Internal representations are aligned using a ranking loss [1]. Image, text, and audio are always passed individually through the model. The output of the model, shown as colored squares, represent 1000-dimensional vectors, with each element representing the probability of the input belonging to a specific ImageNet-1K class. The figure does not originate from the original paper, but is a custom visualization of the concept. Image and text example is taken from the MSCOCO train set [3], the spectogram originates from the SHRe paper [1].

[1]Datasets used consists of videos collected from Flickr, from which frames were extracted and used as images with the corresponding audio [1].

[2]Data has been collected and annotated using Amazon Mechanical Turk [1], [2]. Where the data originates from is not specified in the paper.

Bibliography

[1] Y. Aytar, C. Vondrick, and A. Torralba, "See, Hear, and Read: Deep Aligned Representations," *arXiv preprint arXiv:1706.00932*, 2017, [Online]. Available: https://arxiv.org/abs/1706.00932

[2] A. Sorokin and D. Forsyth, "Utility data annotation with Amazon Mechanical Turk," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2008, pp. 1–8. doi: 10.1109/CVPRW.2008.4562953.

[3] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, D. J. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., in Lecture Notes in Computer Science, vol. 8693. Springer, 2014, pp. 740–755.