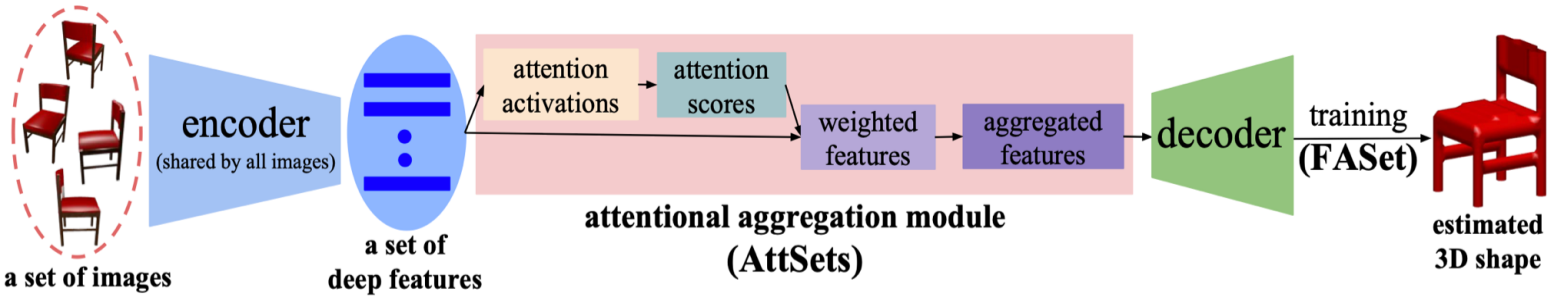
Robust Attentional Aggregation of Deep Feature Sets for Multi-view 3D Reconstruction

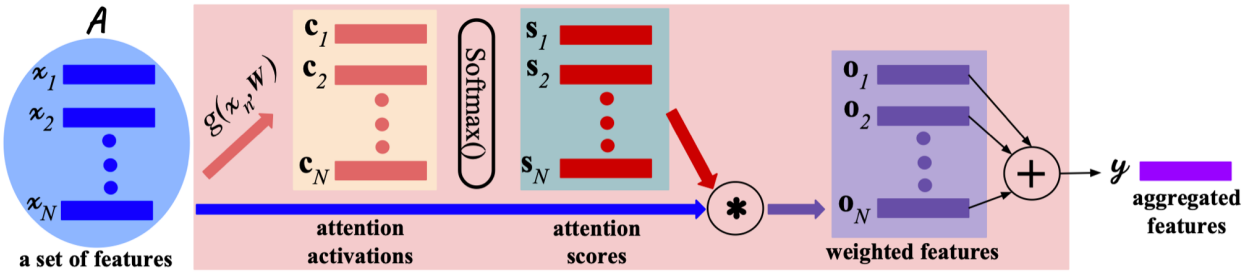
Report



Dataset : ShapeNet, ModelNet, Blobby Dataset
Pytorch Code : https://github.com/Yang7879/AttSets
Paper: https://arxiv.org/pdf/1808.00758.pdf
Key Points to note :

AttSets

- First, the AttSets take the deep features from the encoder as an input and learns the attention activations first. Then it learns the attention scores.
- The attention scores are multiplied with the deep features so that the less important features are removed and more important features are captured
- Unlike in MaxPooling/AveragePooling where only partial information is captured and many important features are lost. In AttSets, it learns to attentively select and weight important deep features, thereby being more effective to aggregate useful information.
- Compared with RNN approaches like 3D-R2N2, AttSets is permutation invariant and computationally efficient.



- Basically, AttSets solves the problem of generalizability of the number of images inputted. It tries to output the feature vector y which is always independent of the number of images used.
- The function g can consist of multiple convolution layers. More the number of layers, more the capability of the Attsets module

FASet

- The encoder-decoder net tends to generalise the type of input whereas the attention module tries to generalise the number of images
- Obviously, the encoder-decoder part is optimised only if 1 image is given whereas the attention module is optimised only if the number of images is greater than 1
- Consider the 2 parts of the network separately. The base encoder and decoder part and the attention module part. So the FASet algorithm says that first we train using only 1 image with the encoder-decoder part and we first optimise it
- Now we consider the whole network and train only the Attention module using multiple images. This gives robustness and generalising capability to the model

Basically, this Attention module can be fitted to any encoder-decoder model which is used for Multi-View 3D Reconstruction. That is, in between the encoder and decoder model we can fit this Attention Module to increase its accuracy.

For example: In the 3D-R2N2 model we can replace the LSTM Part with this attention module which will drastically increase its accuracy

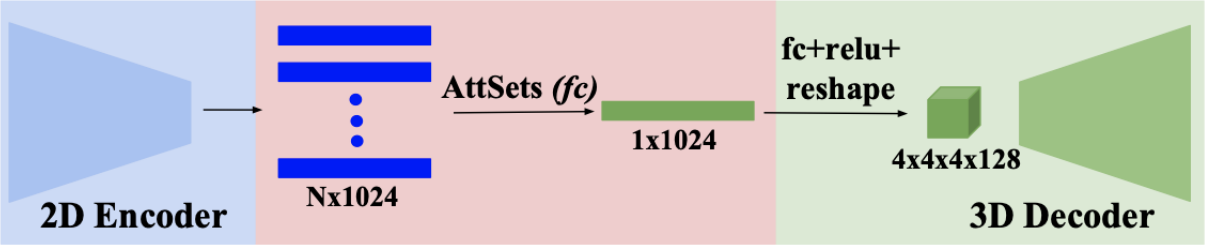


Fig. 4 The architecture of $\text{Base}_{\text{r2n2}}$-AttSets for multi-view 3D reconstruction network. The base encoder-decoder is the same as 3D-R2N2.

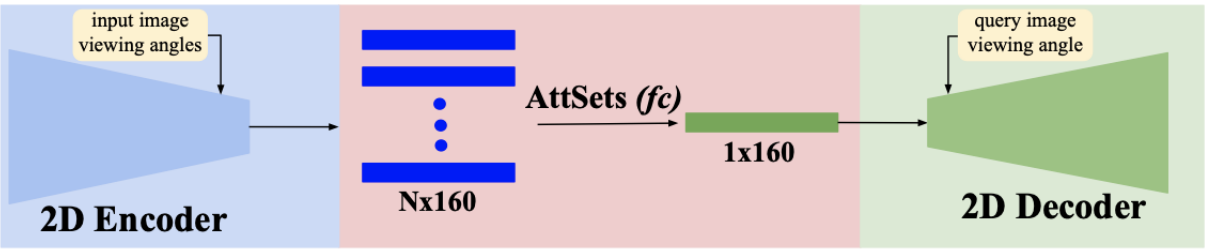


Fig. 5 The architecture of $\text{Base}_{\text{silnet}}$-AttSets for multi-view 3D shape learning. The base encoder-decoder is the same as SilNet.