6. Appendix

6.1. Architecture details

As a further explanation of the model architecture, we plot the conceptual attention module in Figure 5.

6.2. Experiments

All experiments are implemented by Python 3.8 and Pytorch 1.5.1. We run each of the experiments on an NVIDIA GeForce RTX 2080Ti GPU. Primarily we use the default Adam optimizer to train our model. The other training details are revealed in Table 5.

6.3. Comparison experiments

In Table 6 and 7, we reveal the settings for the baseline experiments. For the ablation study in Section 4.5, we use the same settings as in the previous section while altering the considered factors.

6.4. Dataset statistics extended

As extended from Table 1, in this section we provide some further details about the datasets and their label graphs.

1. **Pet datasets**. We construct 14 augmented nodes in addition to the 39 label nodes extracted from both datasets. The total edge number is 119.
2. **Flower datasets**. We augment the label graph by 18 intermediate nodes with a total edge number 129.
3. **Arxiv**. We adopt the original Arxiv label hierarchy graph structuring the website [4].

Due to the size issues of these graphs, we only visualize the moderate-size graph of the Pet datasets in Figure 6. We will release the other label graphs in the format of metafiles upon publication.

[4] http://arxiv.org/

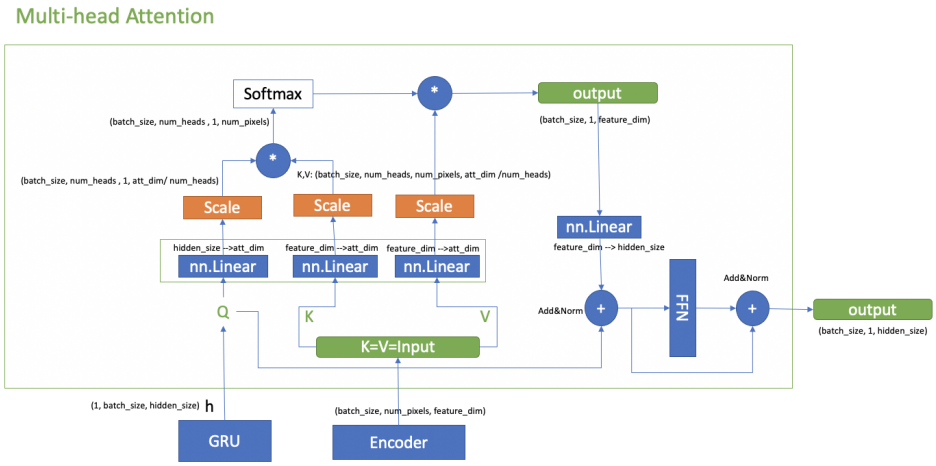

Figure 5. Multi-head attention module architecture.

Table 5. Training settings for the classification experiments of our label space augmentation approach (***label-aug***). ***LRE*** and ***LR*** indicate the learning rate adopted for the encoder backbone (*EfficientNet-B4* and *BERT/LSTM* respectively for image and text experiments) and the rest of model parameters. ***Dymanic-Reduce(n)*** means we lower the learning rate by half when the dev set performance does not improve for a consecutive *n* epochs.

DATASET	*Batch Size*	*Initial LRE*	*Initial LR*	*Decay Period(epoch)*	*Attention Size*	*Hidden Size*	*Input*
Oxford-IIIT Pet	8	0.00005	0.0004	10	1024	1024	380*380
PetFusion	24	0.0001	0.0004	10	1024	1024	380*380
102 Category Flower	12	0.0001	0.0005	*Dynamic Reduce(5)*	1024	1024	380*380
FlowerFusion	12	0.0002	0.0006	*Dynamic Reduce(5)*	1024	1024	380*380
ArxivOriginal	128	0.000015	0.000015	*Dynamic Reduce(5)*	768	768	$\leq$512
ArxivFusion	128	0.000015	0.000015	*Dynamic Reduce(5)*	768	768	$\leq$512
ArxivOriginal-LSTM	128	0.0007	0.0007	*Dynamic Reduce(5)*	768	768	$\leq$512

Table 6. Training settings for baseline experiments. Namely the *EfficientNet-B4 + FFN* and *BERT/LSTM + FFN* experiments for image and text domain respectively.

DATASET	*Epoch*	*Batch Size*	*Input*	*Initial LR*	*Decay Period(epoch)*
Oxford-IIIT Pet	60	16	380*380	0.00005	10
102 Category Flower	60	16	380*380	0.0001	*Dynamic Reduce(5)*
ArxivOriginal	40	128	$\leq$ 512	0.000015	*Dynamic Reduce(5)*
ArxivFusion	40	128	$\leq$ 512	0.000015	*Dynamic Reduce(5)*
ArxivOriginal-LSTM	80	128	$\leq$ 512	0.0007	*Dynamic Reduce(5)*

Table 7. Training settings for *Pseudo Label* method.

DATASET	*Epoch*	*Batch Size*	*Input*	*Initial LR*	*Decay Period(epoch)*
Oxford-IIIT Pet	60	16	380*380	0.0001	10
PetFusion	60	16	380*380	0.0001	10
102 Category Flower	60	16	380*380	0.0004	*Dynamic Reduce(5)*
FlowerFusion	60	16	380*380	0.0004	*Dynamic Reduce(5)*

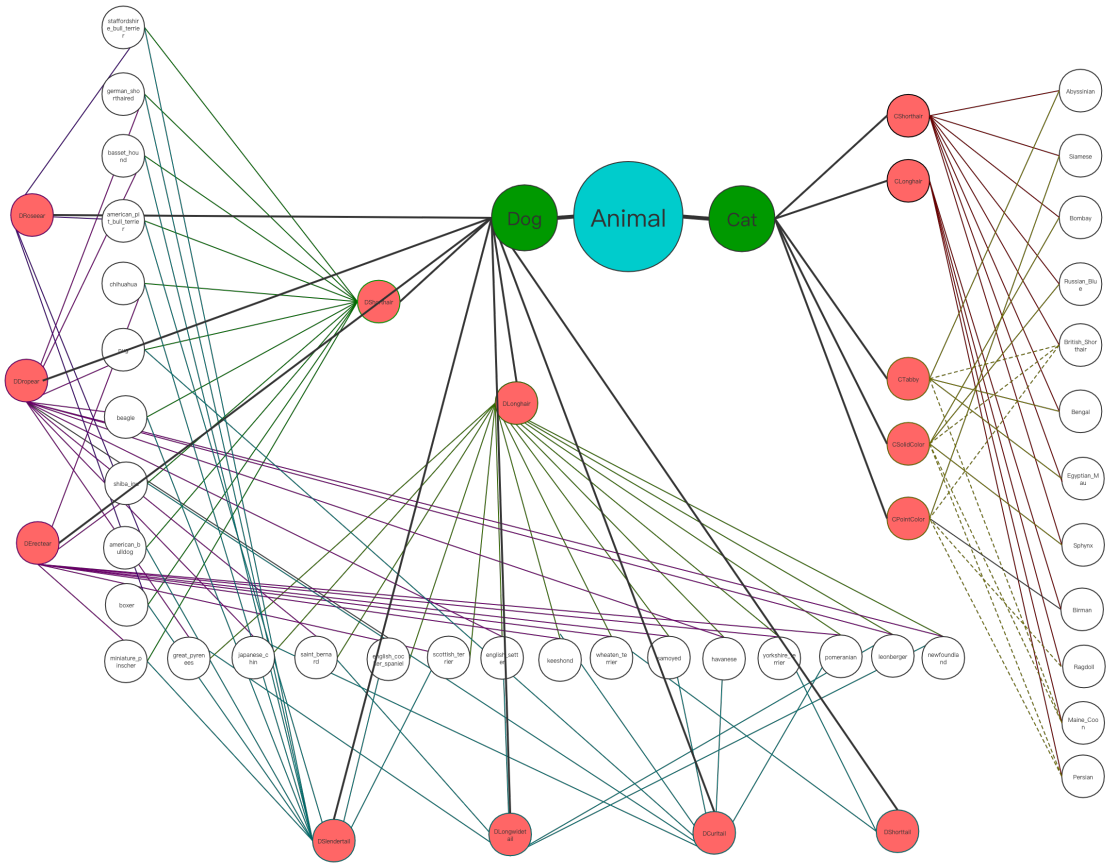

Figure 6. Complete label graph for *Pet datasets*. The red and green color denote augmented nodes at different hierarchies, and white color indicates label nodes. The dotted lines indicate nondeterministic paths.