INF368 Selected Topics in Machine Learning
Exercise 3: Siamese Networks and Zero Shot Learning

Håvard Vågstøl

May 19, 2019

1 Abstract

The exercise covers vector embedding of image data. Plankton images from the ZooScanNet data set are embedded to a 64 dimensional space using the provided neural network implementation. Training and validation progress is illustrated, together with the development of cluster sizes and distance between centroid distances.

A visualization of the clustering is presented using t-SNE plots of model output when embedding the validation data following each iteration.

Finally, the embeddings are classified by supervised (SVM) and unsupervised (k-means) machine learning algorithms to evaluate performance of such an approach.

Code for the exercise will be available at `http://www.github.com/hvagstol/inf368/ex3`

2 Training

2.1 Framework

The provided basis from `https://github.com/ketil-malde/plankton-siamese` was used. This first uses ImageMagick® to rescale images, and the uses the *prepare-data.sh* script to split the data into training, validation and test sets.

The model is then created using the *create_model.py* script. Most importantly, the triplet loss function used during training is defined here. The triplet loss calculates a loss based on distances between an anchor image (positive), a positive example and a negative example. The loss is smaller when the distance between anchor and positive is smaller, and when distance between anchor and negative is bigger. Distance here would mean the euclidian distance between the embeddings of the three images.

2.2 Loss

The *std_triplet_loss* function was used during training for this exercise. This loss function may converge prematurely if the negative-anchor distance increases too fast compared to the positive-anchor distance. The function has a lower bound at zero, and if it is driven here by the negative term any changes during further training will not be able to improve on the network weights.

2.3 Data Generators

The supplied code for data generators was used, which picks random positive and negative samples from the relevant folders and use them for training and validation. Further work on this project would be likely to include data augmentation practices such as presented in exercise two, but focus here was kept on the presented tasks for this exercise.

2.4 Reports

The provided framework produces confusion matrices showing how the data in the validation set clusters around the centroids, and to which extent this is accurate based on ground truth. For interpreting the heatmap illustrations below, we note that dark colors/high values on the diagonal are indicative of high accuracy for the class in that row. High values/dark colors off the diagonal are indicative of confusion between classes.

We note that there is a weak tendency from the first to the second plot that data aggregate on the diagonal, but that some confusion points are actually stronger in the second plot. Significant change is evident between the second plot @50 epochs and the third plot @100 epochs.

The confusion plots are generated from the *Heatmaps* Jupyter notebook, available in the code repository. Summary plots, training/validation loss, and centroid distance/cluster size plots are generated from the *Plotting* notebook.

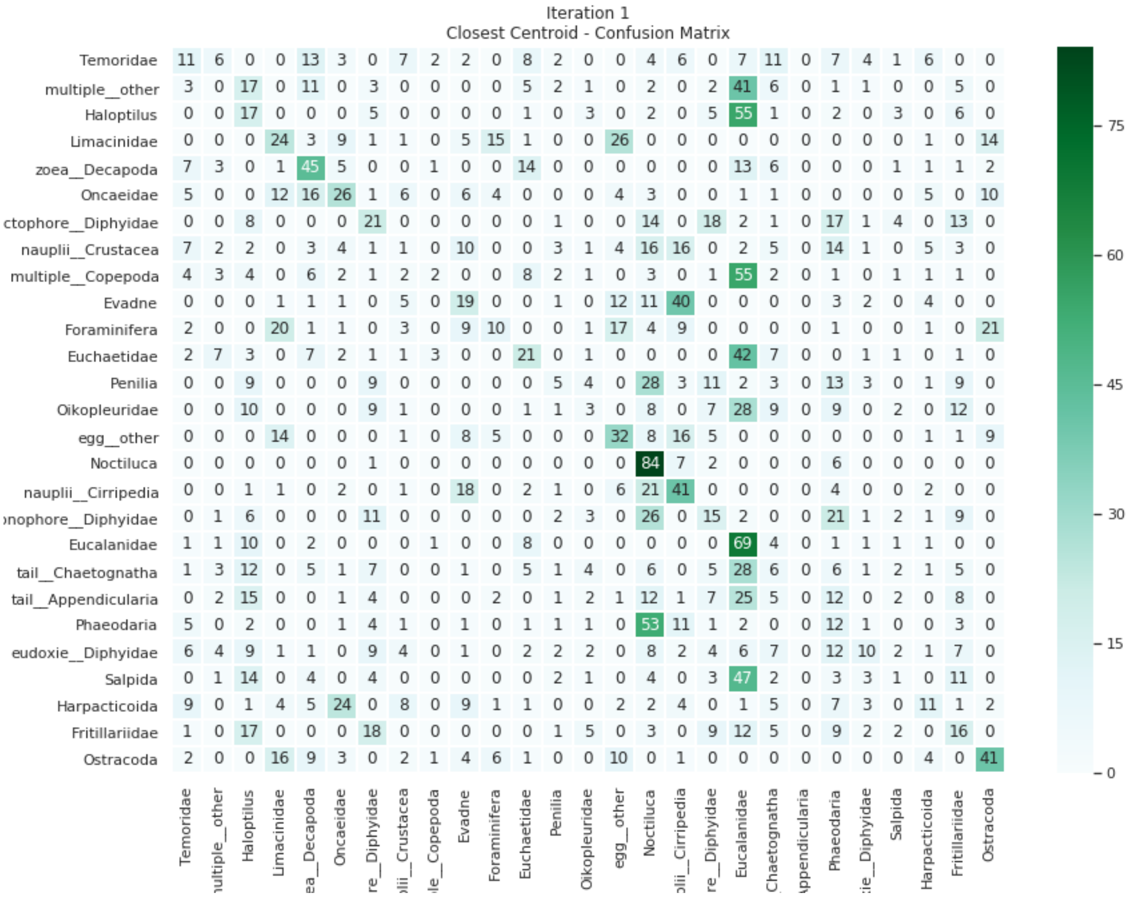
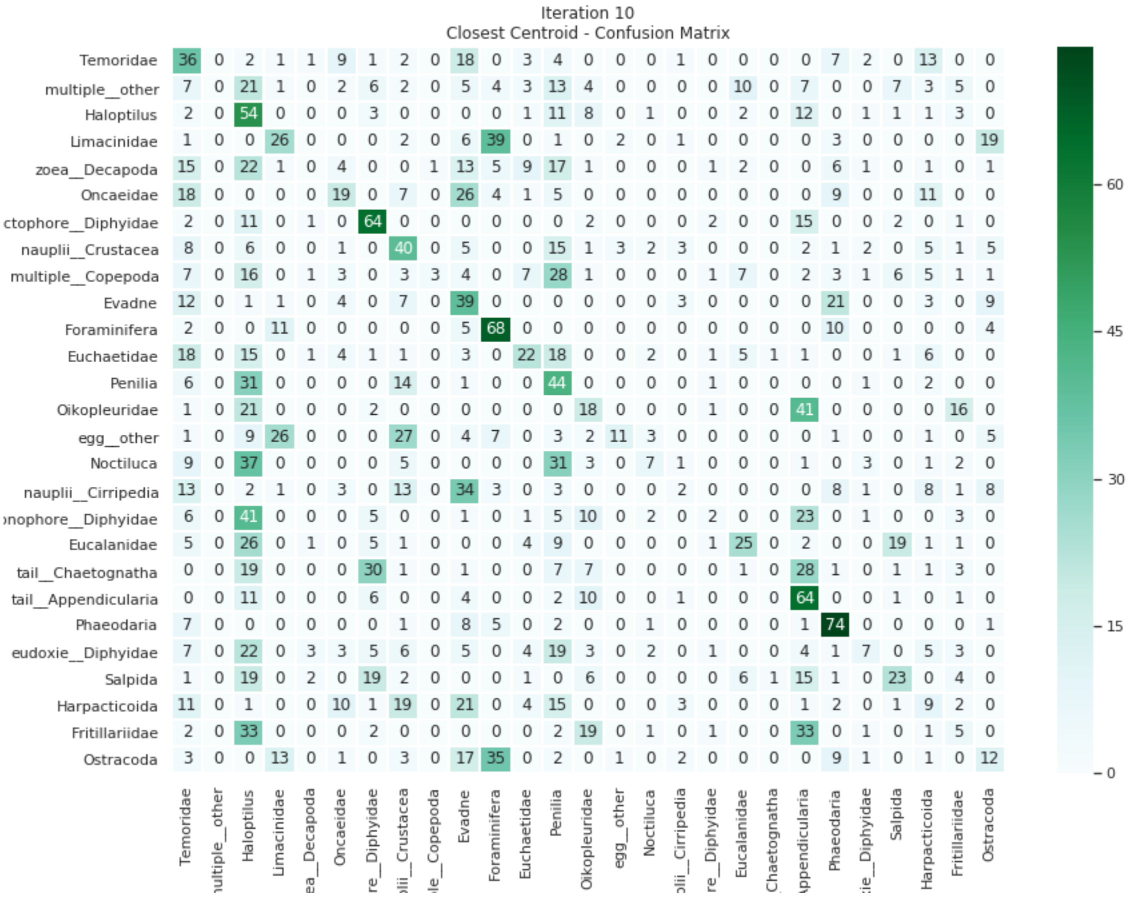
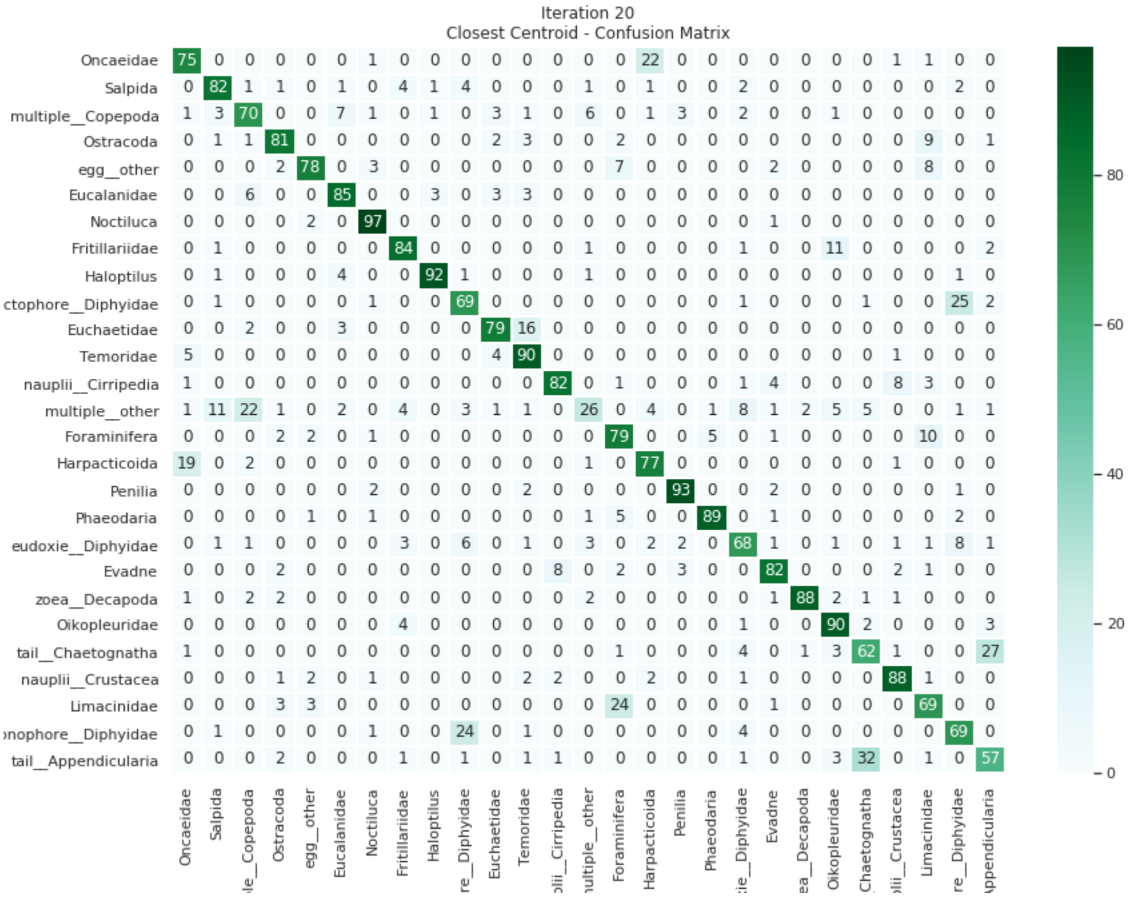


Figure 1: Confusion @ 5 epochs. True class in rows, predicted in columns.



	Temoridae	ıultiple__other	Haloptilus	Limacinidae	ea__Decapoda	Oncaeidae	re__Diphyidae	ɔlii__Crustacea	le__Copepoda	Evadne	Foraminifera	Euchaetidae	Penilia	Oikopleuridae	egg__other	Noctiluca	ɔlii__Cirripedia	re__Diphyidae	Eucalanidae	Chaetognatha	Appendicularia	Phaeodaria	ie__Diphyidae	Salpida	Harpacticoida	Fritillariidae	Ostracoda
Temoridae	36	0	2	1	1	9	1	2	0	18	0	3	4	0	0	0	1	0	0	0	0	7	2	0	13	0	0
multiple__other	7	0	21	1	0	2	6	2	0	5	4	3	13	4	0	0	0	0	10	0	7	0	0	7	3	5	0
Haloptilus	2	0	54	0	0	0	3	0	0	0	0	1	11	8	0	1	0	0	2	0	12	0	1	1	1	3	0
Limacinidae	1	0	0	26	0	0	0	2	0	6	39	0	1	0	2	0	1	0	0	0	0	3	0	0	0	0	19
zoea__Decapoda	15	0	22	1	0	4	0	0	1	13	5	9	17	1	0	0	0	1	2	0	0	6	1	0	1	0	1
Oncaeidae	18	0	0	0	0	19	0	7	0	26	4	1	5	0	0	0	0	0	0	0	0	9	0	0	11	0	0
ctophore__Diphyidae	2	0	11	0	1	0	64	0	0	0	0	0	0	2	0	0	0	2	0	0	15	0	0	2	0	1	0
nauplii__Crustacea	8	0	6	0	0	1	0	40	0	5	0	0	15	1	3	2	3	0	0	0	2	1	2	0	5	1	5
multiple__Copepoda	7	0	16	0	1	3	0	3	3	4	0	7	28	1	0	0	0	1	7	0	2	3	1	6	5	1	1
Evadne	12	0	1	1	0	4	0	7	0	39	0	0	0	0	0	0	3	0	0	0	0	21	0	0	3	0	9
Foraminifera	2	0	0	11	0	0	0	0	0	5	68	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	4
Euchaetidae	18	0	15	0	1	4	1	1	0	3	0	22	18	0	0	2	0	1	5	1	1	0	0	1	6	0	0
Penilia	6	0	31	0	0	0	0	14	0	1	0	0	44	0	0	0	0	1	0	0	0	0	1	0	2	0	0
Oikopleuridae	1	0	21	0	0	0	2	0	0	0	0	0	0	18	0	0	0	1	0	0	41	0	0	0	0	16	0
egg__other	1	0	9	26	0	0	0	27	0	4	7	0	3	2	11	3	0	0	0	0	0	1	0	0	1	0	5
Noctiluca	9	0	37	0	0	0	0	5	0	0	0	0	31	3	0	7	1	0	0	0	1	0	3	0	1	2	0
nauplii__Cirripedia	13	0	2	1	0	3	0	13	0	34	3	0	3	0	0	0	2	0	0	0	0	8	1	0	8	1	8
ɔnophore__Diphyidae	6	0	41	0	0	0	5	0	0	1	0	1	5	10	0	2	0	2	0	0	23	0	1	0	0	3	0
Eucalanidae	5	0	26	0	1	0	5	1	0	0	0	4	9	0	0	0	0	1	25	0	2	0	0	19	1	1	0
tail__Chaetognatha	0	0	19	0	0	0	30	1	0	1	0	0	7	7	0	0	0	0	1	0	28	1	0	1	1	3	0
tail__Appendicularia	0	0	11	0	0	0	6	0	0	4	0	0	2	10	0	0	1	0	0	0	64	0	0	1	0	1	0
Phaeodaria	7	0	0	0	0	0	0	1	0	8	5	0	2	0	0	1	0	0	0	0	1	74	0	0	0	0	1
eudoxie__Diphyidae	7	0	22	0	3	3	5	6	0	5	0	4	19	3	0	2	0	1	0	0	4	1	7	0	5	3	0
Salpida	1	0	19	0	2	0	19	2	0	0	0	1	0	6	0	0	0	0	6	1	15	1	0	23	0	4	0
Harpacticoida	11	0	1	0	0	10	1	19	0	21	0	4	15	0	0	0	3	0	0	0	1	2	0	1	9	2	0
Fritillariidae	2	0	33	0	0	0	2	0	0	0	0	0	2	19	0	1	0	1	0	0	33	0	1	0	1	5	0
Ostracoda	3	0	0	13	0	1	0	3	0	17	35	0	2	0	1	0	2	0	0	0	0	9	1	0	1	0	12

Figure 2: Confusion @ 50 epochs. True class in rows, predicted in columns.



Iteration 20
Closest Centroid - Confusion Matrix

	Oncaeidae	Salpida	le__Copepoda	Ostracoda	egg__other	Eucalanidae	Noctiluca	Fritillariidae	Haloptilus	re__Diphyidae	Euchaetidae	Temoridae	lii__Cirripedia	ultiple__other	Foraminifera	Harpacticoida	Penilia	Phaeodaria	ie__Diphyidae	Evadne	ea__Decapoda	Oikopleuridae	Chaetognatha	lii__Crustacea	Limacinidae	re__Diphyidae	ppendicularia
Oncaeidae	75	0	0	0	0	0	1	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	1	1	0	0
Salpida	0	82	1	1	0	1	0	4	1	4	0	0	0	1	0	1	0	0	2	0	0	0	0	0	0	2	0
multiple__Copepoda	1	3	70	0	0	7	1	0	1	0	3	1	0	6	0	1	3	0	2	0	0	1	0	0	0	0	0
Ostracoda	0	1	1	81	0	0	0	0	0	0	2	3	0	0	2	0	0	0	0	0	0	0	0	0	9	0	1
egg__other	0	0	0	2	78	0	3	0	0	0	0	0	0	0	7	0	0	0	0	2	0	0	0	0	8	0	0
Eucalanidae	0	0	6	0	0	85	0	0	3	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Noctiluca	0	0	0	0	2	0	97	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Fritillariidae	0	1	0	0	0	0	0	84	0	0	0	0	0	1	0	0	0	0	1	0	0	11	0	0	0	0	2
Haloptilus	0	1	0	0	0	4	0	0	92	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
ctophore__Diphyidae	0	1	0	0	0	0	1	0	0	69	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	25	2
Euchaetidae	0	0	2	0	0	3	0	0	0	0	79	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Temoridae	5	0	0	0	0	0	0	0	0	0	4	90	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
nauplii__Cirripedia	1	0	0	0	0	0	0	0	0	0	0	0	82	0	1	0	0	0	1	4	0	0	0	8	3	0	0
multiple__other	1	11	22	1	0	2	0	4	0	3	1	1	0	26	0	4	0	1	8	1	2	5	5	0	0	1	1
Foraminifera	0	0	0	2	2	0	1	0	0	0	0	0	0	0	79	0	0	5	0	1	0	0	0	0	10	0	0
Harpacticoida	19	0	2	0	0	0	0	0	0	0	0	0	0	1	0	77	0	0	0	0	0	0	0	1	0	0	0
Penilia	0	0	0	0	0	0	2	0	0	0	0	2	0	0	0	0	93	0	0	2	0	0	0	0	0	1	0
Phaeodaria	0	0	0	0	1	0	1	0	0	0	0	0	0	1	5	0	0	89	0	1	0	0	0	0	0	2	0
eudoxie__Diphyidae	0	1	1	0	0	0	0	3	0	6	0	1	0	3	0	2	2	0	68	1	0	1	0	1	1	8	1
Evadne	0	0	0	2	0	0	0	0	0	0	0	0	8	0	2	0	3	0	0	82	0	0	0	2	1	0	0
zoea__Decapoda	1	0	2	2	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	88	2	1	1	0	0	0
Oikopleuridae	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	1	0	0	90	2	0	0	0	3
tail__Chaetognatha	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	4	0	1	3	62	1	0	0	27
nauplii__Crustacea	0	0	0	1	2	0	1	0	0	0	0	2	2	0	0	2	0	0	1	0	0	0	0	88	1	0	0
Limacinidae	0	0	0	3	3	0	0	0	0	0	0	0	0	0	24	0	0	0	0	1	0	0	0	0	69	0	0
nophore__Diphyidae	0	1	0	0	0	0	1	0	0	24	0	1	0	0	0	0	0	0	4	0	0	0	0	0	0	69	0
tail__Appendicularia	0	0	0	2	0	0	0	1	0	1	0	1	1	0	0	0	0	0	1	0	0	3	32	0	1	0	57

Figure 3: Confusion @ 100 epochs. True class in rows, predicted in columns.

The training also provides periodic summaries showing the intra class centroid distances after each iteration as well as cluster sizes. As the centroid distance to itself is always zero, the diagonal of these plots is instead used to describe the cluster sizes.

Regarding the summaries, we want big distances (dark colors off the diagonal) and small cluster sizes (light colors on the diagonal). It's hard to discern a strong development in either between the two first plots. A quick look at the time progression of the mean cluster size and centroid distance data confirms this.

When we look at the summary heatmap of centroid distances and cluster sizes for the final epoch, the trend seems to be towards more distance, while sizes do not seem to change all that much. Looking at the progression of the two there seems to be little development after about 75 epochs where the mean size seems to stabilize, while the mean distance may show a weak positive trend from 50-100 epochs though variance is high.

All in all there seems to be little purpose in continuing training further than the 100 epochs here. There is a noticeable shift in the training and validation loss at 50 epochs where training was restarted with the original training rate - leading to the conclusion that perhaps the learning rate decay could be less agressive throughout.

Finally we look at the training and validation losses as training progresses. Training loss decreases at a steady rate over the first 50 epochs used, while validation loss is more variable. For the last 50 epochs used, the validation loss also seems to plateau.

3 Visualization

Using the models generated during training, t-SNE visualizations were done for the validation data. At 50 epochs, The t-SNE visualization indicates some improvements in clustering of the data that is not clearly evident from the plots from the training section. At 100 epochs clustering is more evident, albeit with confusion between some classes being likely. Note in particular classes 24/25, classes 18/23 and 19/20. Though class numbers are used in the plots for clarity, these can easily be cross referenced in the confusion matrix plots below. The attached GIF animation illustrates how this develops over time.

The code for loading the models, running the vector embedding and finally creating and visualizing the t-SNE embedding can be found in the TSNE notebook in the GitHub repository referred above.

4 Classification

For the classification, we base the following on embeddings of the images rather than the actual images. The embeddings correspond to the vectors returned by the final model from above.

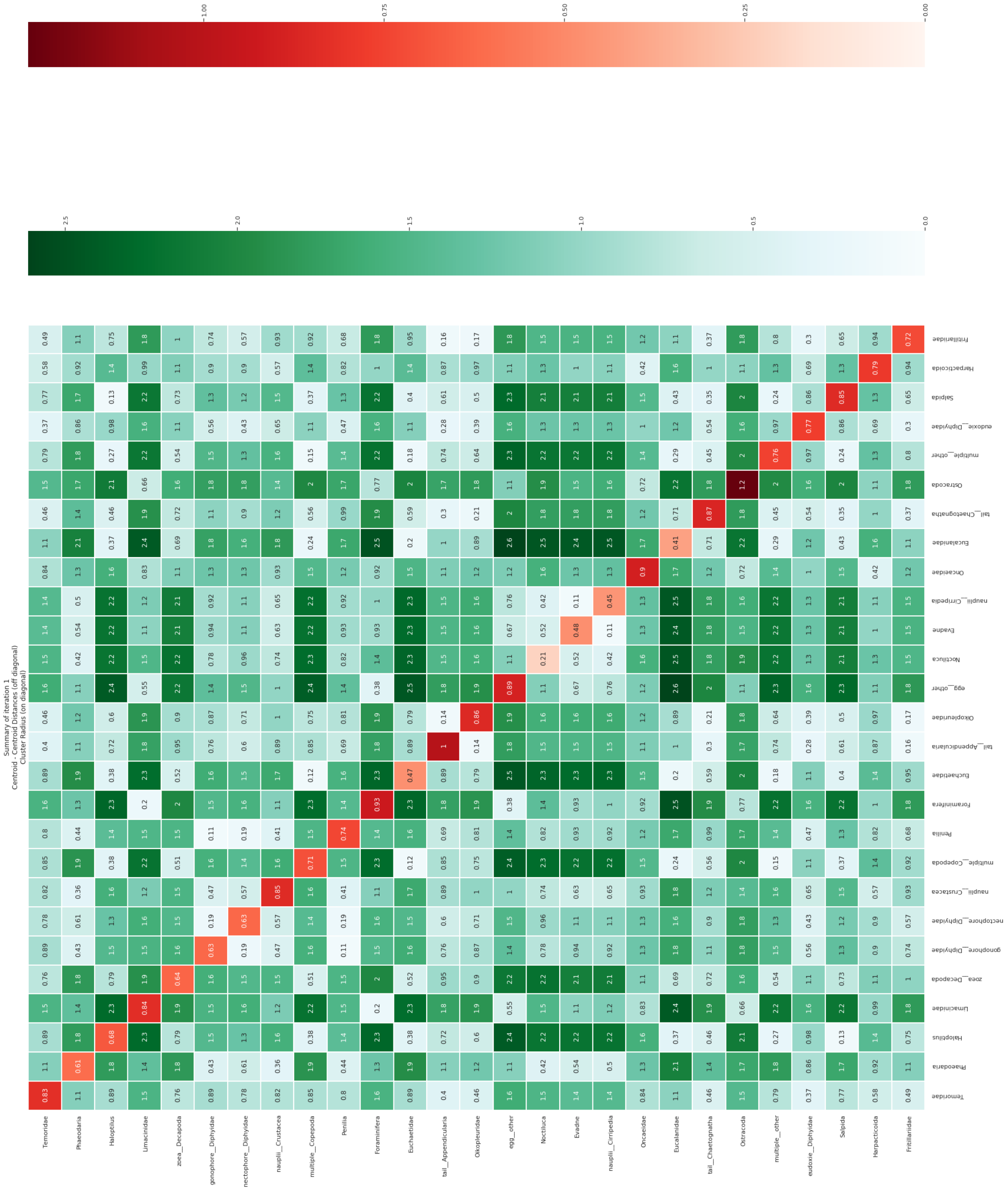


Figure 4: Centroid distances @ 5 epochs. Cluster sizes on diagonal.

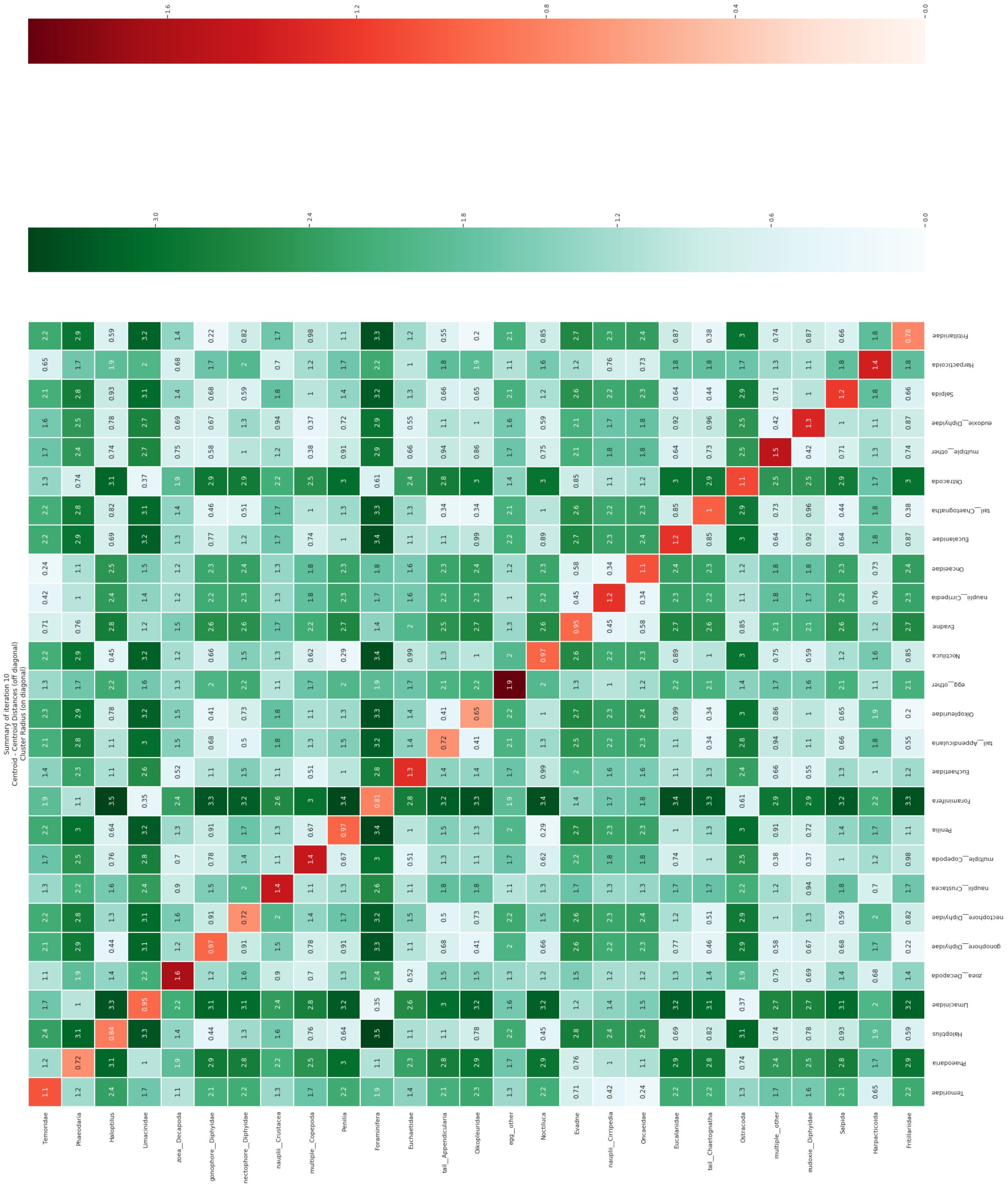


Figure 5: Centroid distances @ 50 epochs. Cluster sizes on diagonal.

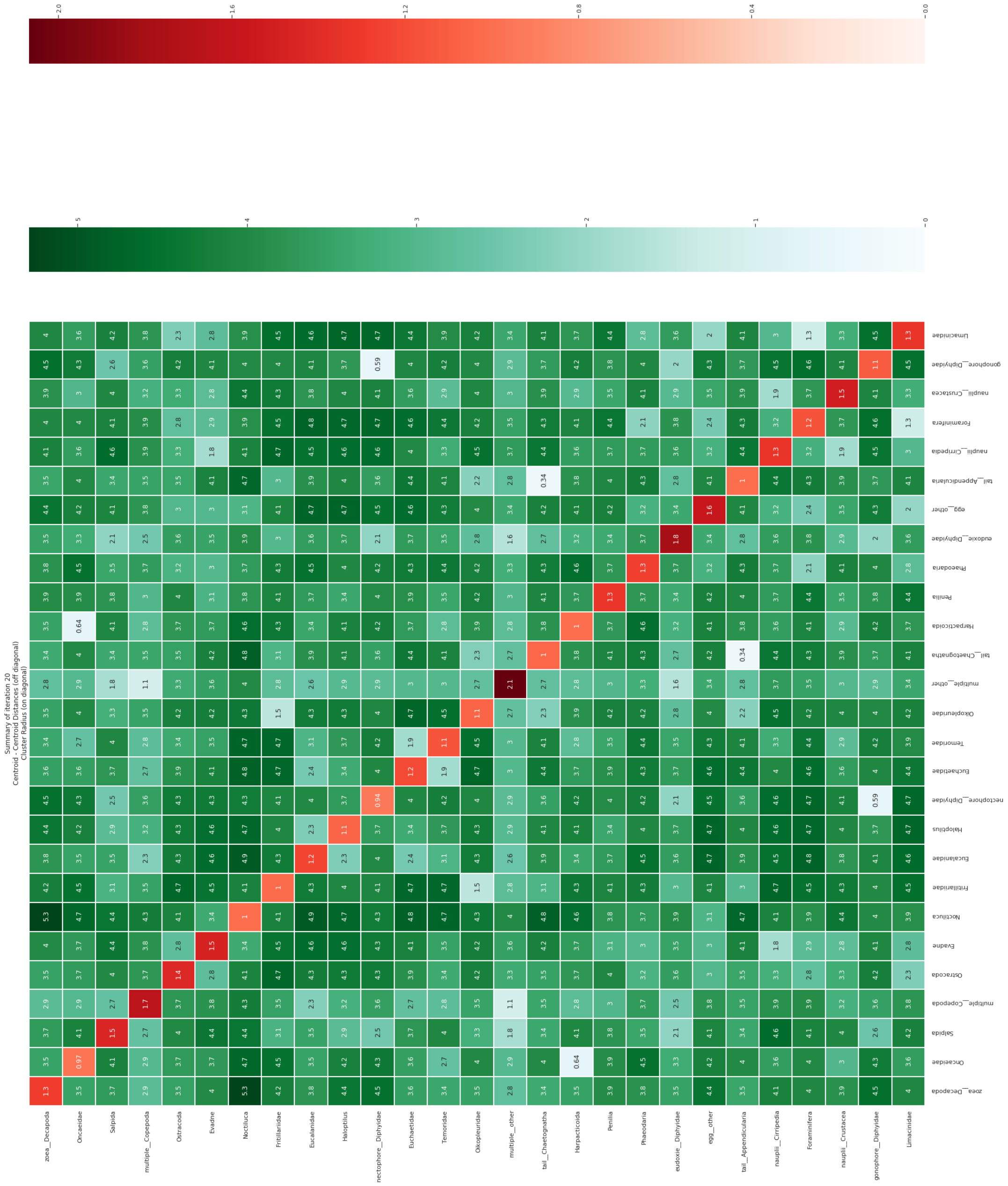


Figure 6: Centroid distances @ 100 epochs. Cluster sizes on diagonal.

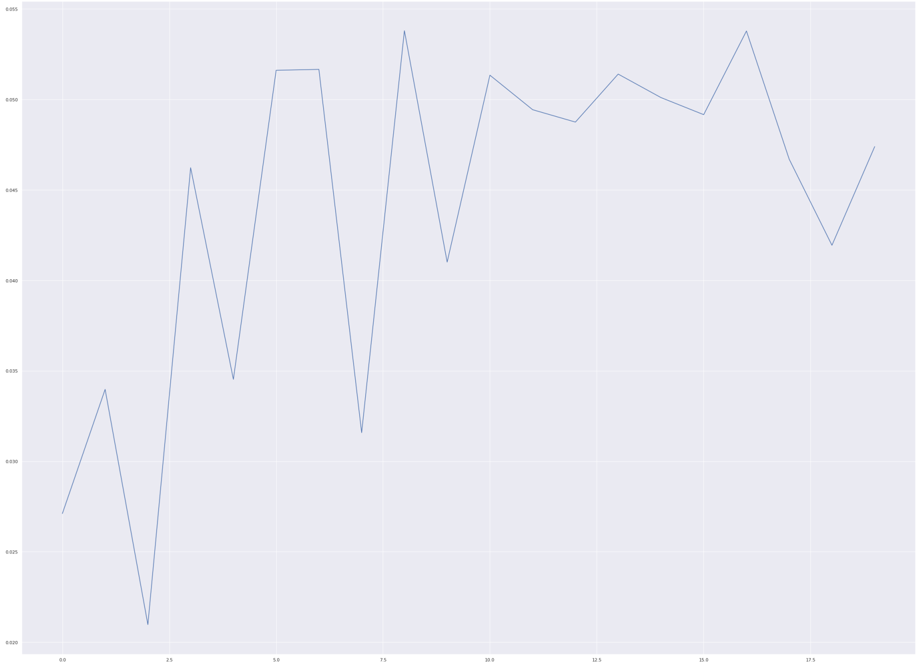


Figure 7: Mean cluster size progression over 100 epochs.

Two separate classification schemes were set up. First among these was a support vector machine with a 3rd degree polynomial kernel. The classifier was fit to the validation data, and used to predict the class of the test data.

The classifiers are available in the *Classifier* notebook in the code repository.

For an unsupervised approach, k-means was chosen, using the MiniBatchK-Means implementation from scikit-learn, with 27 clusters corresponding to the classes of our data. A challenge with evaluating the unsupervised learning is that the class of the data is not known, so it's not immediately apparent which cluster correspons to which class. To overcome this problem, the clustering algorithm was first fit to the validation data, then the test data was run through the prediction in a class-wise manner. The majority class of this prediction was then assumed to be predictive of the correct mapping between cluster number and actual class. We are thus able to produce a confusion matrix for the unsupervised case as well.

While the accuracy of the k-means clustering is slightly worse than that of the SVM classification, with 0.63 vs 0.75, it actually does a decent job for most of the classes. Some noteable exceptions occur, and it is of particular interest to note the confusion between the two tail classes - as this pops up in all our visualizations.

For an example of classes which are confounded in the classifiers, we can have a look at some samples from the classes *tail__ Appendicularia* and *tail__ Chaetognatha*.

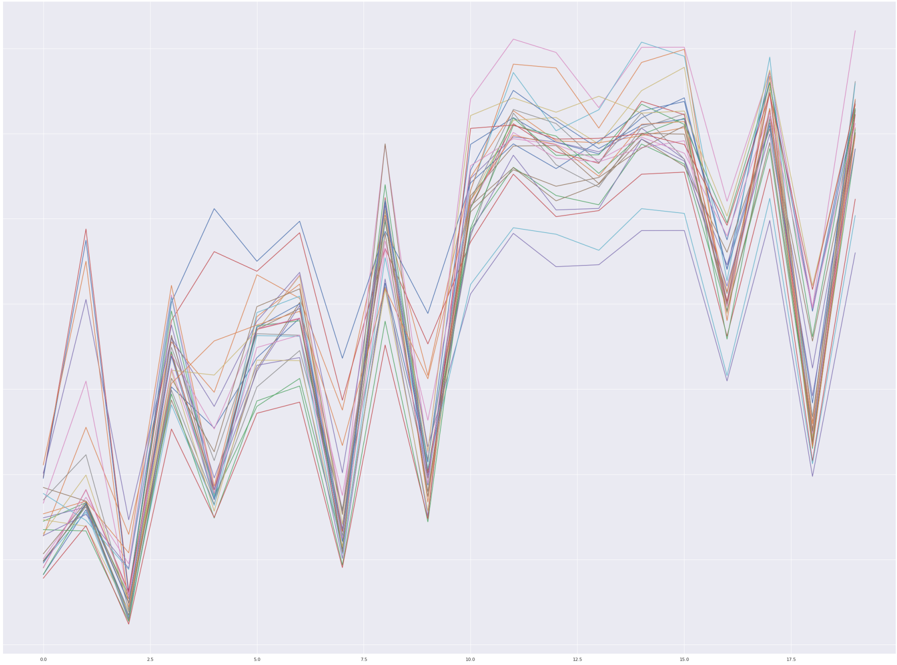


Figure 8: Mean centroid distances progression over 100 epochs.

Clearly these are both tails, and visual inspection shows them to be similar in appearance as well. It is interesting to note that this is clearly reflected in both the t-SNE plot above, where classes 24 and 25 are quite close, and the confusion matrices of unsupervised or supervised classification run on the vector embeddings of the images.

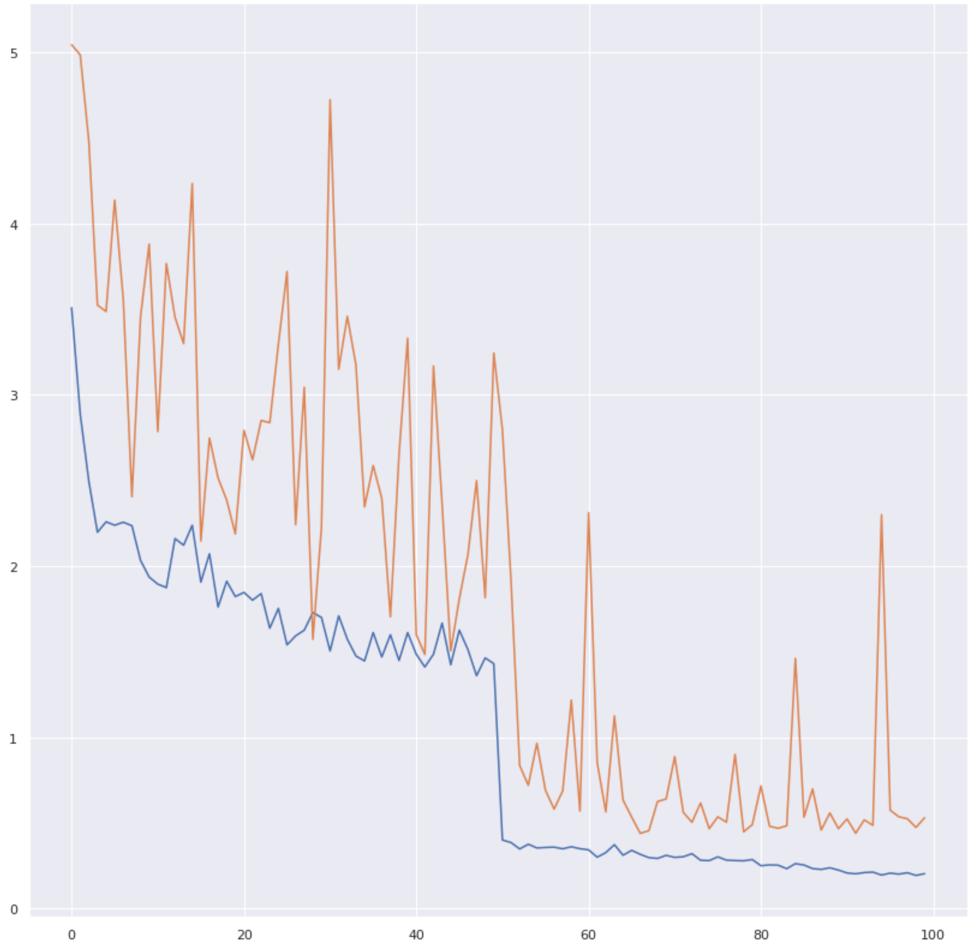


Figure 9: Training (blue) and validation (orange) losses over 100 epochs.

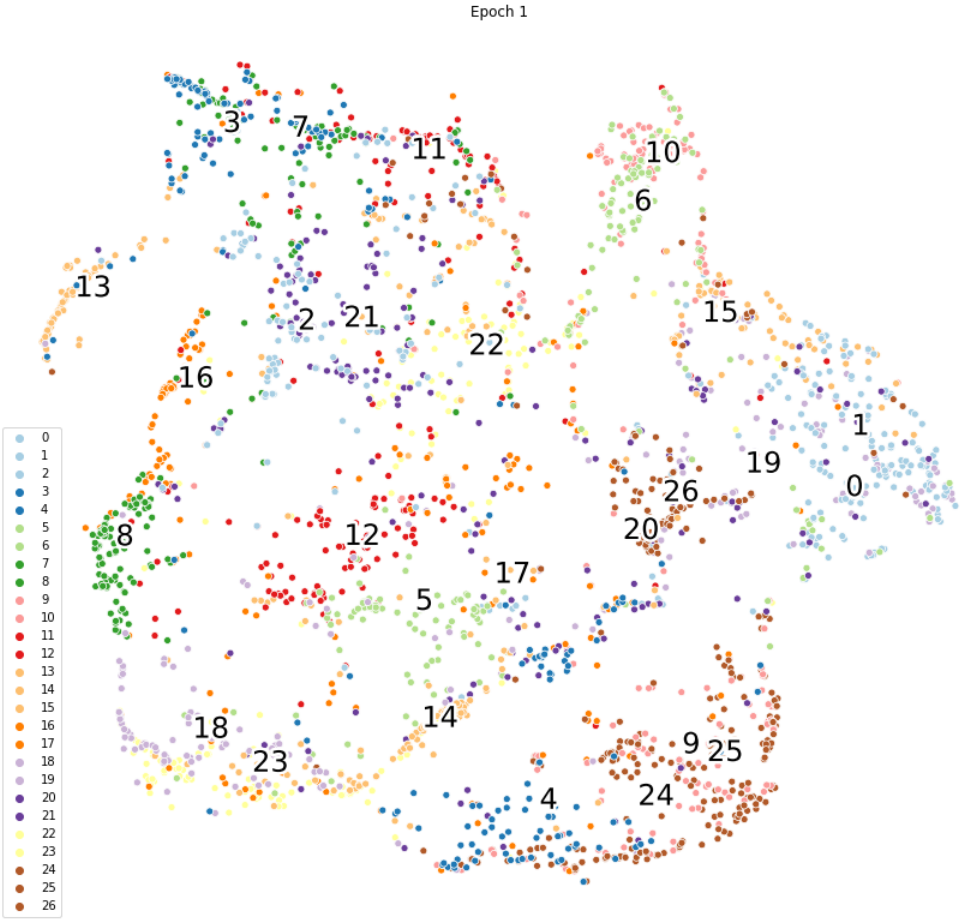


Figure 10: t-SNE visualization after 5 epochs

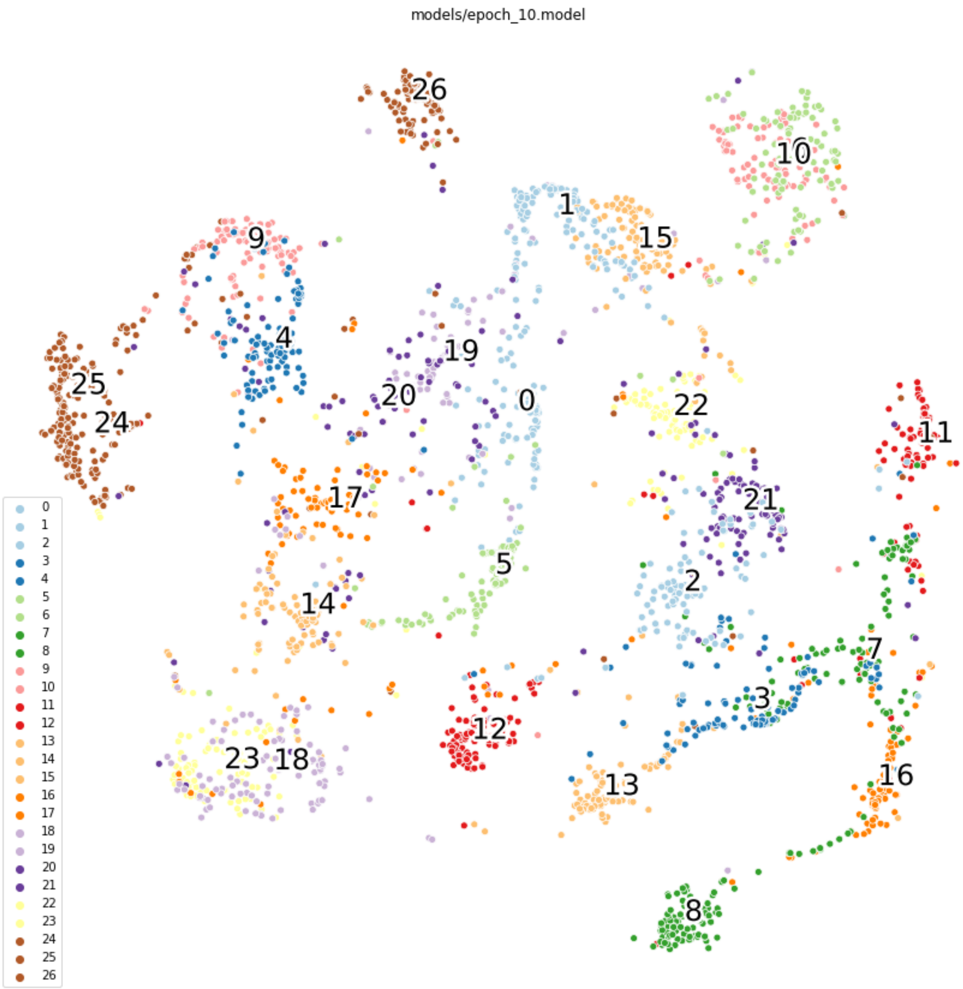


Figure 11: t-SNE visualization after 50 epochs

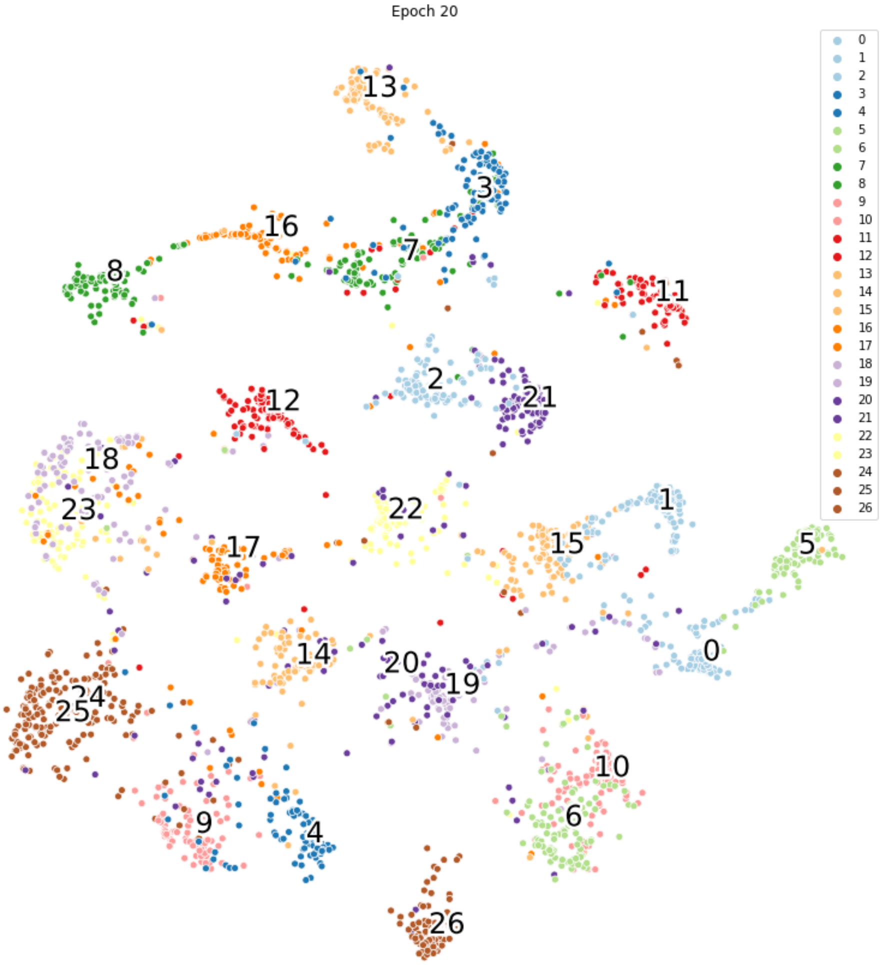
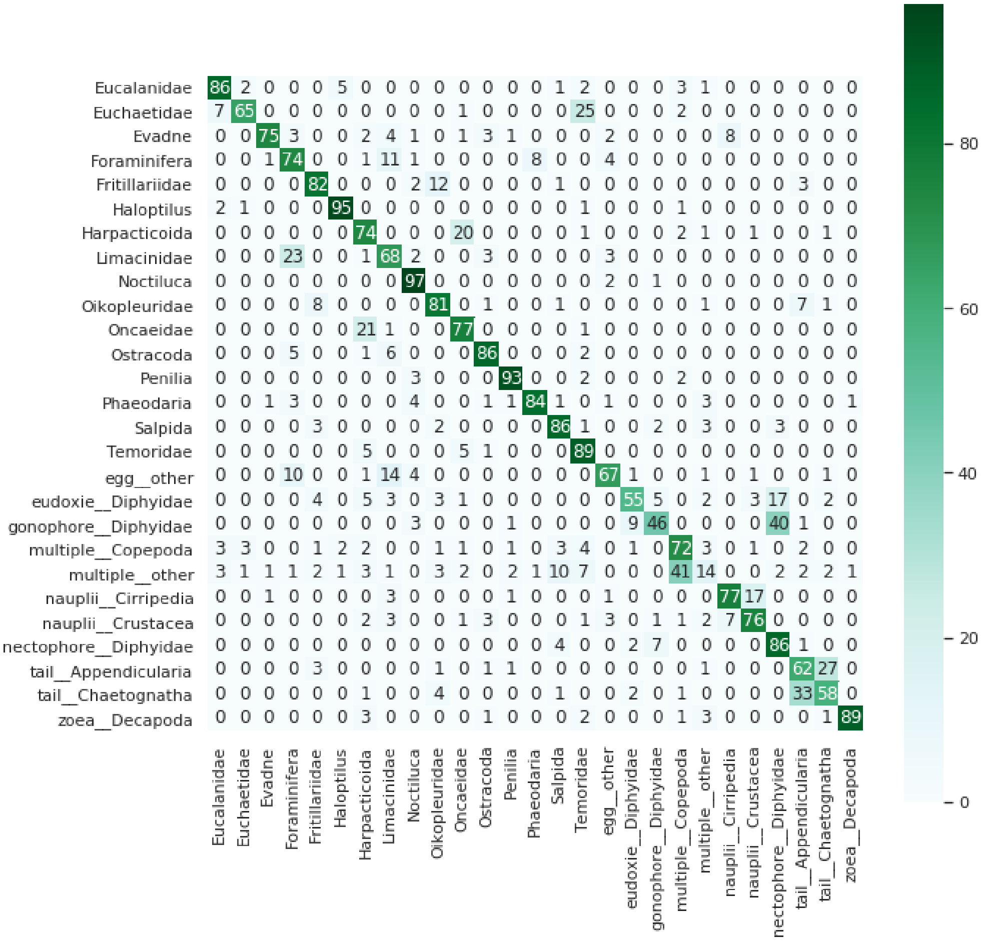
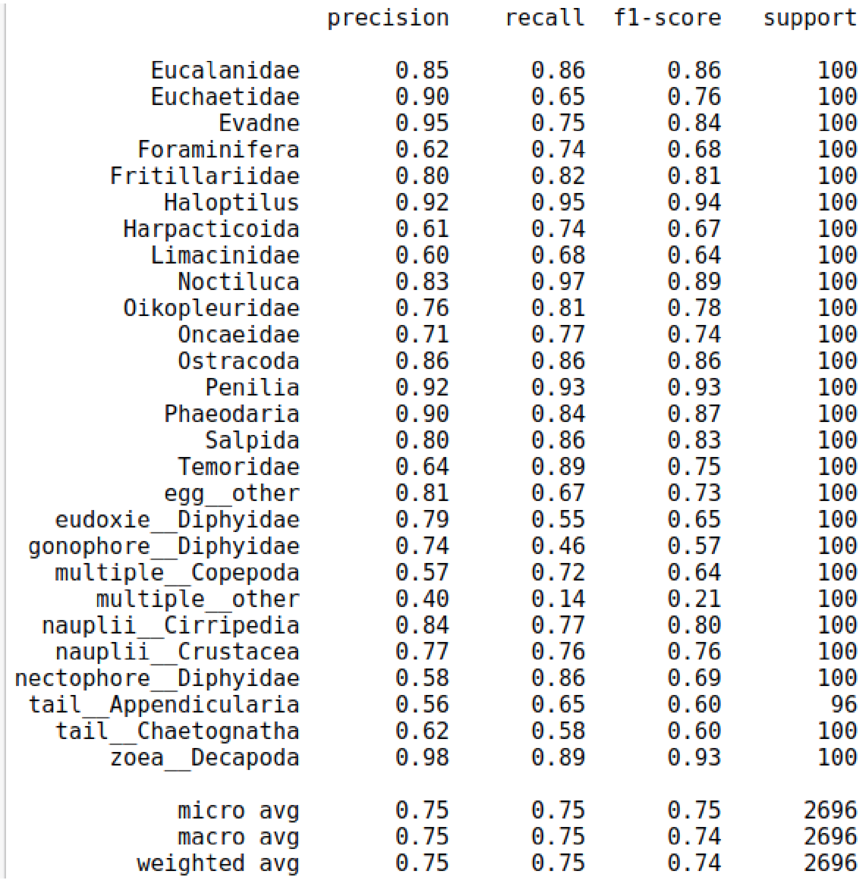


Figure 12: t-SNE visualization after 100 epochs



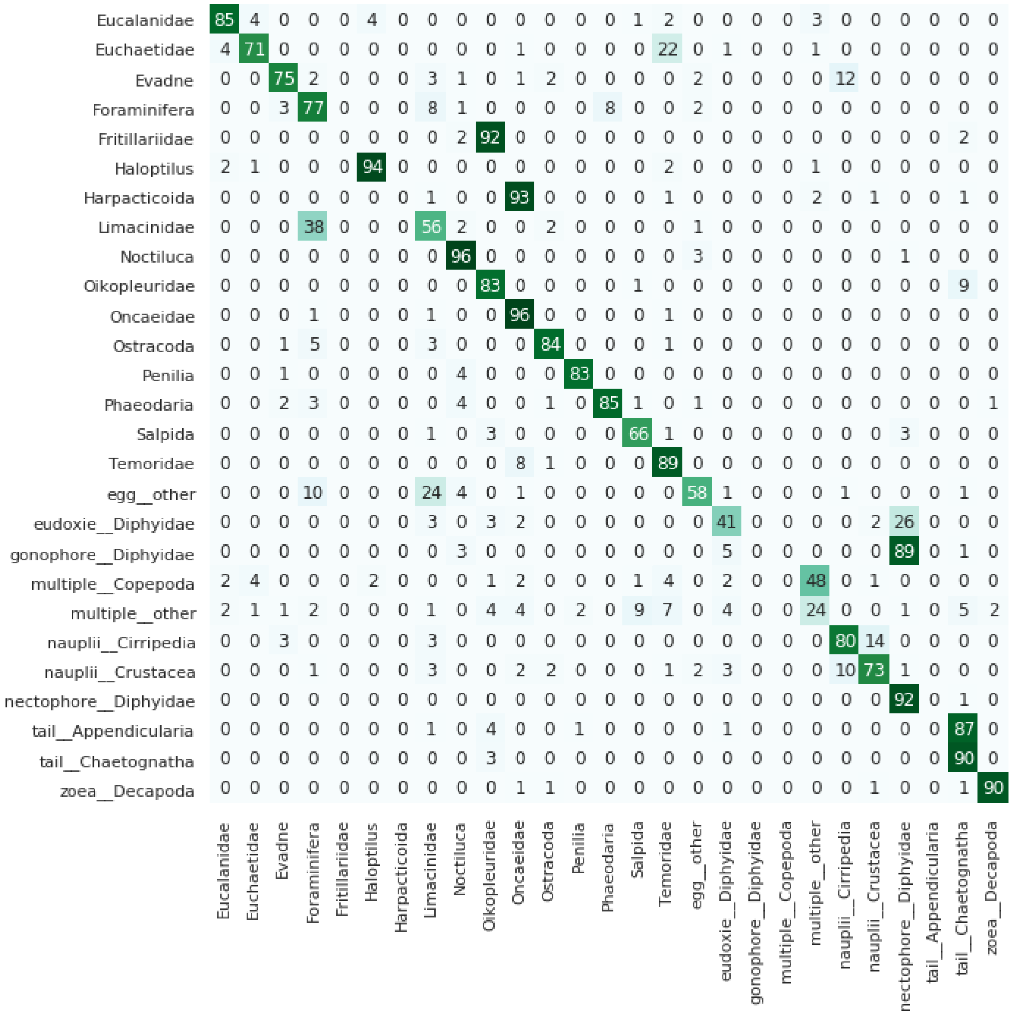
	Eucalanidae	Euchaetidae	Evadne	Foraminifera	Fritillariidae	Haloptilus	Harpacticoida	Limacinidae	Noctiluca	Oikopleuridae	Oncaeidae	Ostracoda	Penilia	Phaeodaria	Salpida	Temoridae	egg__other	eudoxie__Diphyidae	gonophore__Diphyidae	multiple__Copepoda	multiple__other	nauplii__Cirripedia	nauplii__Crustacea	nectophore__Diphyidae	tail__Appendicularia	tail__Chaetognatha	zoea__Decapoda
Eucalanidae	86	2	0	0	0	5	0	0	0	0	0	0	0	0	1	2	0	0	0	3	1	0	0	0	0	0	0
Euchaetidae	7	65	0	0	0	0	0	0	0	0	1	0	0	0	0	25	0	0	0	2	0	0	0	0	0	0	0
Evadne	0	0	75	3	0	0	2	4	1	0	1	3	1	0	0	0	2	0	0	0	0	8	0	0	0	0	0
Foraminifera	0	0	1	74	0	0	1	11	1	0	0	0	0	8	0	0	4	0	0	0	0	0	0	0	0	0	0
Fritillariidae	0	0	0	0	82	0	0	0	2	12	0	0	0	0	1	0	0	0	0	0	0	0	0	0	3	0	0
Haloptilus	2	1	0	0	0	95	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
Harpacticoida	0	0	0	0	0	0	74	0	0	0	20	0	0	0	0	1	0	0	0	2	1	0	1	0	0	1	0
Limacinidae	0	0	0	23	0	0	1	68	2	0	0	3	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0
Noctiluca	0	0	0	0	0	0	0	0	97	0	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0
Oikopleuridae	0	0	0	0	8	0	0	0	0	81	0	1	0	0	1	0	0	0	0	0	1	0	0	0	7	1	0
Oncaeidae	0	0	0	0	0	0	21	1	0	0	77	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Ostracoda	0	0	0	5	0	0	1	6	0	0	0	86	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
Penilia	0	0	0	0	0	0	0	0	3	0	0	0	93	0	0	2	0	0	0	2	0	0	0	0	0	0	0
Phaeodaria	0	0	1	3	0	0	0	0	4	0	0	1	1	84	1	0	1	0	0	0	3	0	0	0	0	0	1
Salpida	0	0	0	0	3	0	0	0	0	2	0	0	0	0	86	1	0	0	2	0	3	0	0	3	0	0	0
Temoridae	0	0	0	0	0	0	5	0	0	0	5	1	0	0	0	89	0	0	0	0	0	0	0	0	0	0	0
egg__other	0	0	0	10	0	0	1	14	4	0	0	0	0	0	0	0	67	1	0	0	1	0	1	0	0	1	0
eudoxie__Diphyidae	0	0	0	0	4	0	5	3	0	3	1	0	0	0	0	0	0	55	5	0	2	0	3	17	0	2	0
gonophore__Diphyidae	0	0	0	0	0	0	0	0	3	0	0	0	1	0	0	0	0	9	46	0	0	0	0	40	1	0	0
multiple__Copepoda	3	3	0	0	1	2	2	0	0	1	1	0	1	0	3	4	0	1	0	72	3	0	1	0	2	0	0
multiple__other	3	1	1	1	2	1	3	1	0	3	2	0	2	1	10	7	0	0	0	41	14	0	0	2	2	2	1
nauplii__Cirripedia	0	0	1	0	0	0	0	3	0	0	0	0	1	0	0	0	1	0	0	0	0	77	17	0	0	0	0
nauplii__Crustacea	0	0	0	0	0	0	2	3	0	0	1	3	0	0	0	1	3	0	1	1	2	7	76	0	0	0	0
nectophore__Diphyidae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	2	7	0	0	0	0	86	1	0	0
tail__Appendicularia	0	0	0	0	3	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	1	0	0	0	62	27	0
tail__Chaetognatha	0	0	0	0	0	0	1	0	0	4	0	0	0	0	1	0	0	2	0	1	0	0	0	0	33	58	0
zoea__Decapoda	0	0	0	0	0	0	3	0	0	0	0	1	0	0	0	2	0	0	0	1	3	0	0	0	0	1	89

Figure 13: SVM Confusion Matrix. True class in rows, predicted in cols.



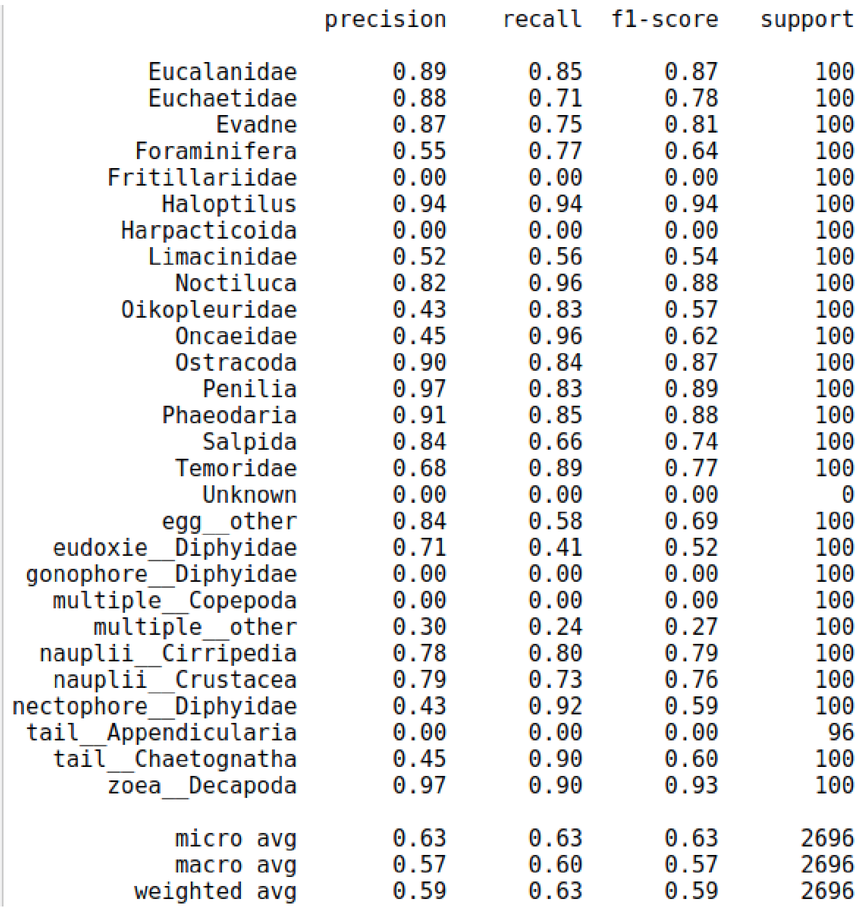
```
                      precision    recall  f1-score   support

           Eucalanidae       0.85      0.86      0.86       100
          Euchaetidae       0.90      0.65      0.76       100
               Evadne       0.95      0.75      0.84       100
         Foraminifera       0.62      0.74      0.68       100
       Fritillariidae       0.80      0.82      0.81       100
           Haloptilus       0.92      0.95      0.94       100
        Harpacticoida       0.61      0.74      0.67       100
          Limacinidae       0.60      0.68      0.64       100
            Noctiluca       0.83      0.97      0.89       100
        Oikopleuridae       0.76      0.81      0.78       100
            Oncaeidae       0.71      0.77      0.74       100
            Ostracoda       0.86      0.86      0.86       100
              Penilia       0.92      0.93      0.93       100
           Phaeodaria       0.90      0.84      0.87       100
              Salpida       0.80      0.86      0.83       100
            Temoridae       0.64      0.89      0.75       100
           egg__other       0.81      0.67      0.73       100
    eudoxie__Diphyidae       0.79      0.55      0.65       100
  gonophore__Diphyidae       0.74      0.46      0.57       100
    multiple__Copepoda       0.57      0.72      0.64       100
       multiple__other       0.40      0.14      0.21       100
   nauplii__Cirripedia       0.84      0.77      0.80       100
    nauplii__Crustacea       0.77      0.76      0.76       100
nectophore__Diphyidae       0.58      0.86      0.69       100
  tail__Appendicularia       0.56      0.65      0.60        96
    tail__Chaetognatha       0.62      0.58      0.60       100
        zoea__Decapoda       0.98      0.89      0.93       100

            micro avg       0.75      0.75      0.75      2696
            macro avg       0.75      0.75      0.74      2696
         weighted avg       0.75      0.75      0.74      2696
```

Figure 14: SVM Classification Report



	Eucalanidae	Euchaetidae	Evadne	Foraminifera	Fritillariidae	Haloptilus	Harpacticoida	Limacinidae	Noctiluca	Oikopleuridae	Oncaeidae	Ostracoda	Penilia	Phaeodaria	Salpida	Temoridae	egg__other	eudoxie__Diphyidae	gonophore__Diphyidae	multiple__Copepoda	multiple__other	nauplii__Cirripedia	nauplii__Crustacea	nectophore__Diphyidae	tail__Appendicularia	tail__Chaetognatha	zoea__Decapoda
Eucalanidae	85	4	0	0	0	4	0	0	0	0	0	0	0	0	1	2	0	0	0	0	3	0	0	0	0	0	0
Euchaetidae	4	71	0	0	0	0	0	0	0	0	1	0	0	0	0	22	0	1	0	0	1	0	0	0	0	0	0
Evadne	0	0	75	2	0	0	0	3	1	0	1	2	0	0	0	0	2	0	0	0	0	12	0	0	0	0	0
Foraminifera	0	0	3	77	0	0	0	8	1	0	0	0	0	8	0	0	2	0	0	0	0	0	0	0	0	0	0
Fritillariidae	0	0	0	0	0	0	0	0	2	92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
Haloptilus	2	1	0	0	0	94	0	0	0	0	0	0	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0
Harpacticoida	0	0	0	0	0	0	0	1	0	0	93	0	0	0	0	1	0	0	0	0	2	0	1	0	0	1	0
Limacinidae	0	0	0	38	0	0	0	56	2	0	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Noctiluca	0	0	0	0	0	0	0	0	96	0	0	0	0	0	0	0	3	0	0	0	0	0	0	1	0	0	0
Oikopleuridae	0	0	0	0	0	0	0	0	0	83	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	9	0
Oncaeidae	0	0	0	1	0	0	0	1	0	0	96	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Ostracoda	0	0	1	5	0	0	0	3	0	0	0	84	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Penilia	0	0	1	0	0	0	0	0	4	0	0	0	83	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Phaeodaria	0	0	2	3	0	0	0	0	4	0	0	1	0	85	1	0	1	0	0	0	0	0	0	0	0	0	1
Salpida	0	0	0	0	0	0	0	1	0	3	0	0	0	0	66	1	0	0	0	0	0	0	0	3	0	0	0
Temoridae	0	0	0	0	0	0	0	0	0	0	8	1	0	0	0	89	0	0	0	0	0	0	0	0	0	0	0
egg__other	0	0	0	10	0	0	0	24	4	0	1	0	0	0	0	0	58	1	0	0	0	1	0	0	0	1	0
eudoxie__Diphyidae	0	0	0	0	0	0	0	3	0	3	2	0	0	0	0	0	0	41	0	0	0	0	2	26	0	0	0
gonophore__Diphyidae	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	5	0	0	0	0	0	89	0	1	0
multiple__Copepoda	2	4	0	0	0	2	0	0	0	1	2	0	0	0	1	4	0	2	0	0	48	0	1	0	0	0	0
multiple__other	2	1	1	2	0	0	0	1	0	4	4	0	2	0	9	7	0	4	0	0	24	0	0	1	0	5	2
nauplii__Cirripedia	0	0	3	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	80	14	0	0	0	0
nauplii__Crustacea	0	0	0	1	0	0	0	3	0	0	2	2	0	0	0	1	2	3	0	0	0	10	73	1	0	0	0
nectophore__Diphyidae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	92	0	1	0
tail__Appendicularia	0	0	0	0	0	0	0	1	0	4	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	87	0
tail__Chaetognatha	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	90	0
zoea__Decapoda	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	90

Figure 15: 27-means clustering Confusion Matrix. True class in rows, predicted in cols.



```
                         precision    recall  f1-score   support

            Eucalanidae       0.89      0.85      0.87       100
            Euchaetidae       0.88      0.71      0.78       100
                 Evadne       0.87      0.75      0.81       100
           Foraminifera       0.55      0.77      0.64       100
          Fritillariidae      0.00      0.00      0.00       100
             Haloptilus       0.94      0.94      0.94       100
          Harpacticoida       0.00      0.00      0.00       100
            Limacinidae       0.52      0.56      0.54       100
              Noctiluca       0.82      0.96      0.88       100
          Oikopleuridae       0.43      0.83      0.57       100
              Oncaeidae       0.45      0.96      0.62       100
              Ostracoda       0.90      0.84      0.87       100
                Penilia       0.97      0.83      0.89       100
             Phaeodaria       0.91      0.85      0.88       100
                Salpida       0.84      0.66      0.74       100
              Temoridae       0.68      0.89      0.77       100
                Unknown       0.00      0.00      0.00         0
             egg__other       0.84      0.58      0.69       100
     eudoxie__Diphyidae       0.71      0.41      0.52       100
   gonophore__Diphyidae       0.00      0.00      0.00       100
     multiple__Copepoda       0.00      0.00      0.00       100
        multiple__other       0.30      0.24      0.27       100
     nauplii__Cirripedia      0.78      0.80      0.79       100
     nauplii__Crustacea       0.79      0.73      0.76       100
  nectophore__Diphyidae       0.43      0.92      0.59       100
   tail__Appendicularia       0.00      0.00      0.00        96
     tail__Chaetognatha       0.45      0.90      0.60       100
         zoea__Decapoda       0.97      0.90      0.93       100

              micro avg       0.63      0.63      0.63      2696
              macro avg       0.57      0.60      0.57      2696
           weighted avg       0.59      0.63      0.59      2696
```

Figure 16: k-means Classification Report

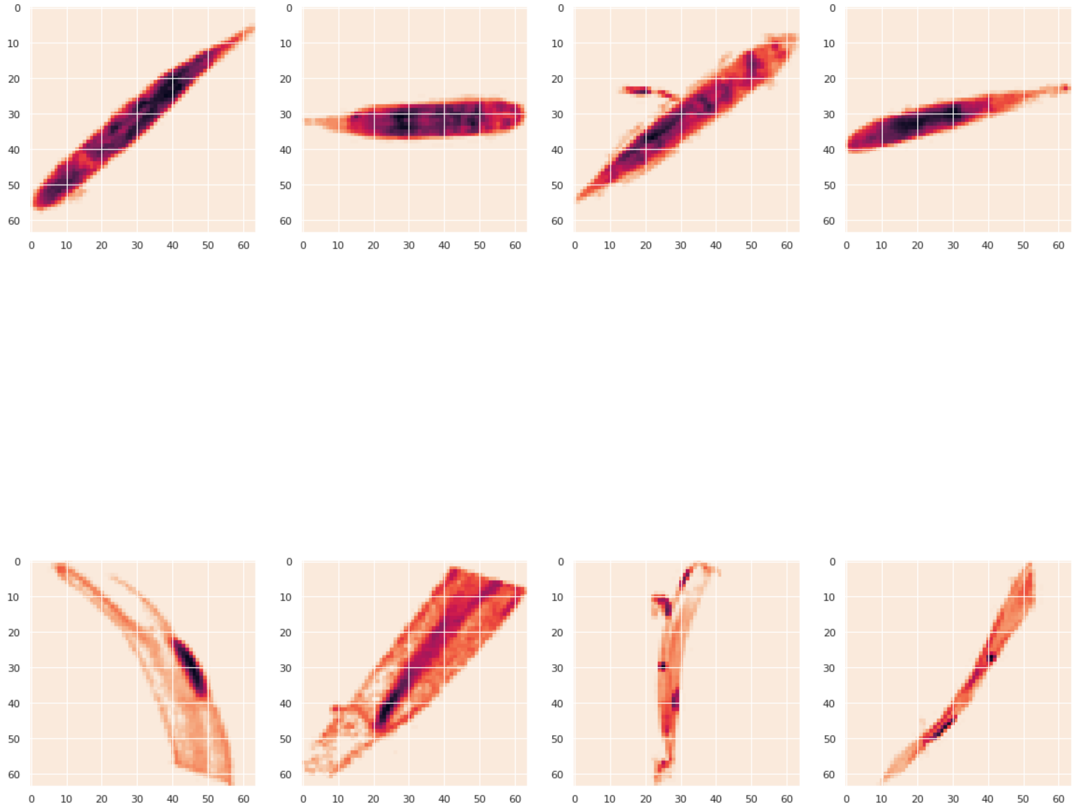


Figure 17: Similarity between the classes tail-Appendicularia (above) and tail-Chaetognatha (below)