Multi-loss Function to Improve the Text Detection and Segmentation

Team name: AA_Vision

Team members:

1. Dr. Anuj Abraham, Senior Researcher TII, Abu Dhabi, UAE.
2. Dr. Shitala Prasad Scientist II, A*STAR, Singapore.

Problem Statement: "Slidin' videos": Slide Transition Detection and Title Extraction in Lecture Videos
https://challenge.aiforgood.itu.int/match/matchitem/74

Introduction

In this current task, we need to address two different task detection and segmentation of titles for which we have a joint learning concept. The problem statement/motivation can be defined as below:

YouTube's "Video Chapter" feature segments a video into sections marked by timestamps so that the user can easily navigate to the part of the video which is of most interest as shown in Figure 1. This can be done by clicking or pressing the chapter marker, or by selecting the timestamp in the video description.

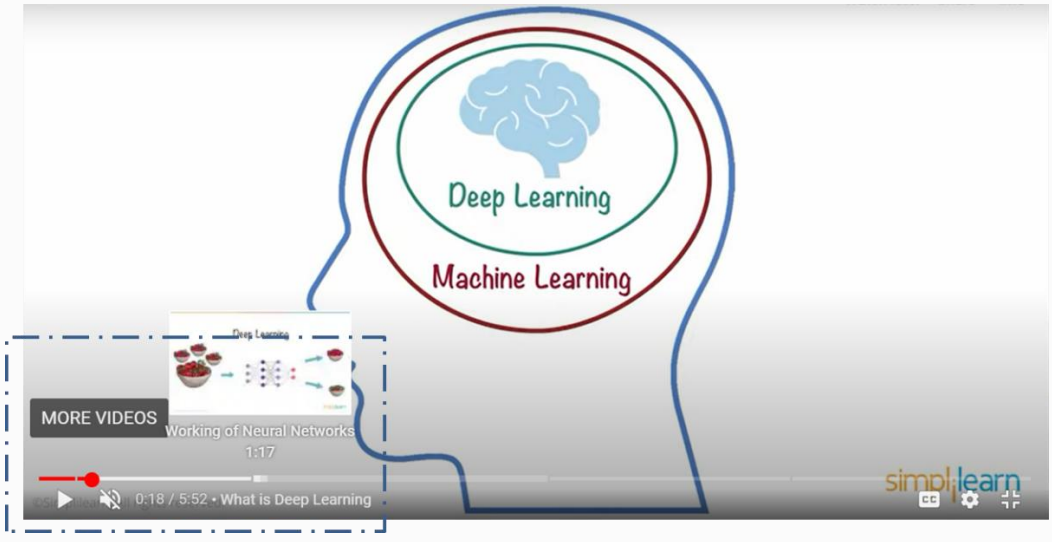


Figure 1: YouTube's "Video Chapter" with timestamps.

One of the important factors in deep feature learning is their loss function design which highly influences the network performance. In this work, we proposed a

classroom (CR) learning approach to obtain high-level discriminative features without any extra load made to the network architecture. [1]

Problem definition

In this problem statement on "*Slide Transition Detection and Title Extraction in Lecture Videos* ", the challenge for us is to create the best AI model which annotates slide transitions by:

- Identifying starting and ending frames of each slide shown in the video
- Extracting (apparent) titles of each slide
- All videos were recorded at 25FPS

Recordings of 100 "AI for Good" (https://aiforgood.itu.int/) webinars were sourced to assemble a diverse collection of more than 140 video presentations made by members of the scientific community, entrepreneurs, and standardization experts.

Dataset:

The dataset was collected by the challenge, and we downloaded them and used them for proposing a solution to improve the baseline performance. The details for the same is as below:

1. Video files covering the presentation from when speaker started screenshare right to the moment when it was turned off. Video files vary in duration (from several minutes to several hours) and resolution (from 1600x1200 to 3840x2160).
2. A ground truth data set with 2500+ slide transitions showing the starting and ending frame of each slide including (apparent) titles.

Problem detail

While this is a slide annotation problem, the dataset contains some complex features. Cases of presenters demonstrating real-world footage amid the slideshow or minimizing PowerPoint and opening another program should be treated as non-slide content.

To distinguish slide content from everything else in predictions we utilize the "is_slide" column. It is set to "0" (zero) for any video fragment that is not a slide, and to "1" (one) otherwise. Non-slide content can be identified in videos through tracking pixels refresh ratio (a typical slide fragment will have simple and discrete visual changes unlike real-world footage) or through an advanced image recognition model trained specifically for this task. Existing evaluation metric will focus on how accurate slide content/slide

transitions were predicted in the video without taking in consideration other types of content.

The representation of frames extracted from a screenshare at 25 frame per second is illustrated in Figure 2. Similarly, Figure 3 shows the excel file listing of all ground truth slides available in the video.

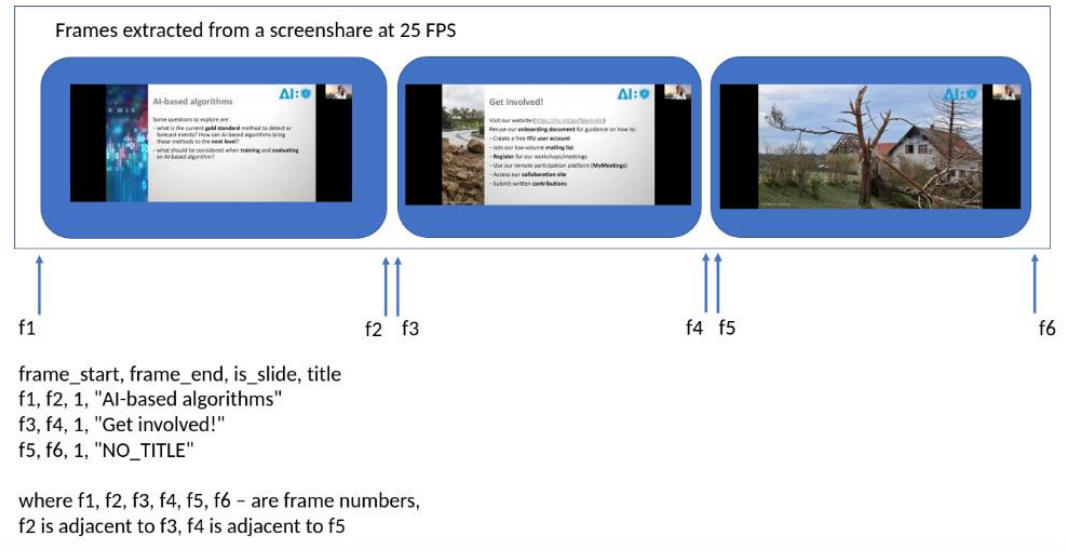


Figure 2: Expected annotation workflow.

A1 starting_frame1

	A	B	C	D	E	F	G	H	I	J	K	L
1	starting_frame1	starting_frame2	ending_frame1	ending_frame2	title1	title2	title3	title4	bonus_title1	bonus_title2	is_hybrid	
2	25	25	97	97	CAN TECHNOLOGY SCALE TO FEED THE WORLD ?						0	
3	98	102	949	977	CAN TECHNOLOGY SCALE TO FEED THE WORLD ?						0	
4	950	978	1332	1345	HPE-STUDENTS-FARMERS						0	
5	1333	1346	1765	1782	HPE-STUDENTS-FARMERS						0	
6	1766	1783	2721	2730	HPE-STUDENTS-FARMERS						0	
7	2722	2731	4268	4279	RESULTS						0	

Figure 3: "Ground_truth.csv"

The structure information in the excel file named "*groundtruth.csv*" is as follows:

a) starting_frame1, starting_frame2: A range of possible starting frame numbers of the slide

- If slide appearance is animated, this range will match the duration of the animation
- Predicting any frame within this range as a starting point of a slide will be correct
- If slide appearance is not animated "starting_frame_1" will be equal to "starting_frame2"

b) ending_frame1, ending_frame2: A range of possible ending frames of the slide.

c) title1, title2, title3, title4: One or more optional titles of the slide

- If only "title1" is specified, a solution should predict it as a slide title
- If "title2"/"title3"/"title4" are specified, a solution can predict any of the specified titles

d) bonus_title1, bonus_title2: Bonus point answers

- If "bonus_title1" is specified, a solution will be granted double points for predicting this title
- If "bonus_title1" and "bonus_title2" are specified, a solution will be granted double points for predicting both titles

e) is_hybrid: "hybrid" slides

- Some slides have video elements and are labelled as "hybrid" (is_hybrid=1)
- These slides are of optional prediction: if a hybrid slide will be part of a prediction output it will not be considered a mistake
- No points will be given for predicting hybrid slides, but you can use them to enhance your training

Motivation

Prasad et al. [1] introduced a multi-loss view invariant stochastic prototype embedding to minimize and improve the recognition accuracy of novel objects at different viewpoints by using a progressive multi-view learning approach.

In the above-mentioned work, the authors have attempted a new learning strategy for object recognition and in the future work addressed to apply the concept to advanced computer vision tasks such as segmentation and prediction. Therefore, in our proposed work we inherit the concept of multi-loss classroom learning strategy for detection and segmentation tasks. The additional benefit of such methods is that we do not need to modify the base model architecture in the baseline work proposed by the challenge *AI for Good ITU*. Hence, our main task and contribution is in the development of an optimization method that minimizes the loss errors using the same architecture without any extensive computational cost. Our methodology is focused on modifying the learning strategy used based on *Multi-loss Function to Improve the Text Detection and Segmentation*.

Training refinements leads to improve accuracy: The parameters that we can focus is on Learning Rate Decay, Label Smoothing, Knowledge Distillation, Mixup Training, Transfer learning to see if they benefit from any downstream learning improvements to improve accuracy. Figure 4 shows the test accuracy per epoch illustration for multi -loss based on seen and unseen categories obtained for object recognition. [2]

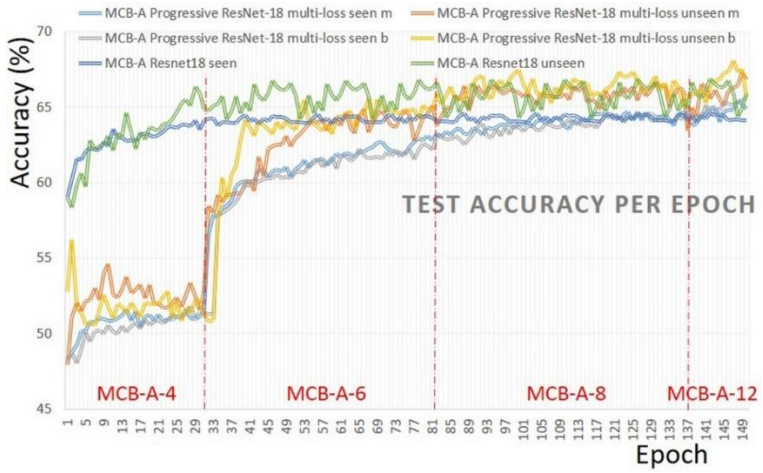


Figure 4: Test accuracy per epoch [2].

It is observed from Figure 4, that after few epochs of learning the view angles are increased for all objects. This results in a better embedding without vanishing gradient issue.

Proposed Solution

The proposed solution is based on the modifying the learning strategy and development of multi-loss function to improve the text detection and segmentation. The basic block diagram of the methodology used in our work is illustrated in Figure 5. We do not modify the deep feature extractor block in this work. In literature, it is seen that if we use more complex architectures, model will be better, but results in extensive computational cost.

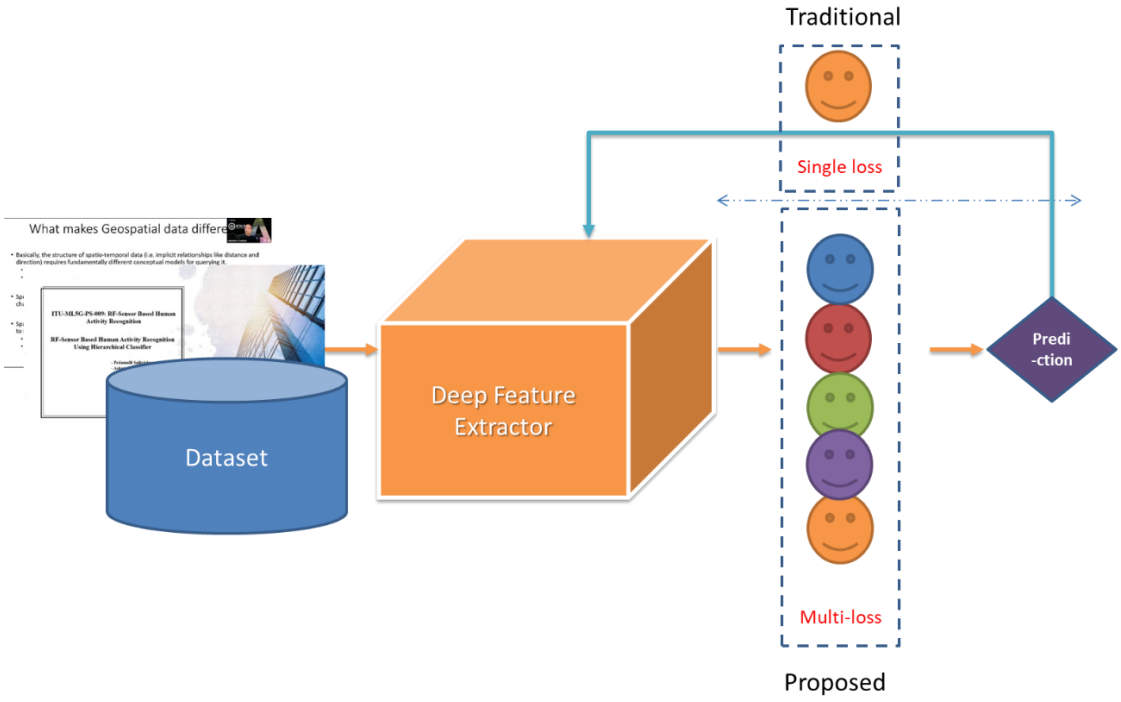


Figure 5: Proposed solution based on learning strategy.

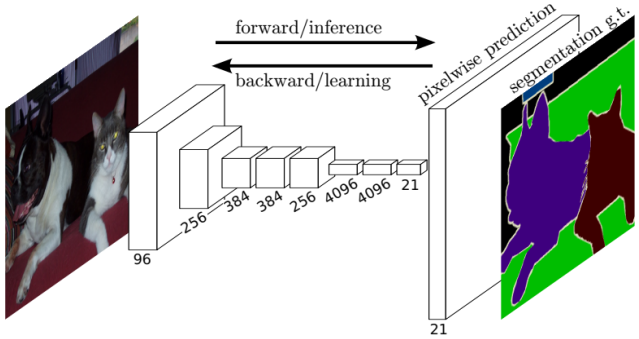
ResNet Architecture:

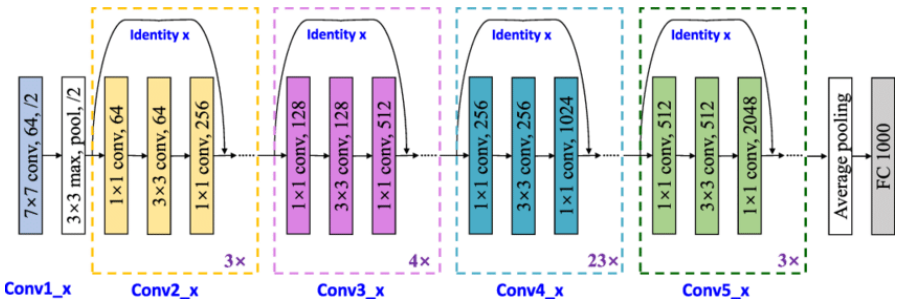
ResNet, short for Residual Networks is a classic neural network used as a backbone for many Computer Vision (CV) tasks. This model was the winner of the ImageNet challenge in 2015.

1. ResNet allowed us to train extremely deep neural networks with 150+layers successfully
2. Prior to ResNet training very deep neural networks was difficult due to the problem of vanishing gradients.
3. RestNet is basically a pre-trained CNN model.
4. It helps to train the model with computer vision task.

Set the define path and circumstances with loops or symbols to get the value. *To solve the vanishing gradient problem, Residual neural networks are introduced - (source selective internet….)*

The most used architectures are Resnet 18, Resnet 50, and Resnet 101. Reset 18 has around 11 million trainable parameters. It will have only 2 pooling layers (one at the beginning and another at the end). It follows 3x3 CONV layers. It also follows a batch normalization process at every layer where the inputs will be normalized for each batch, if every batch will resemble the normal distribution characteristics of the entire sample.





https://github.com/msminhas93/DeepLabv3FineTuning/blob/bcdc3dfc79a5b75bc30c52b32315661c0a4da17e/model.py#L6

Figure 6: VGG convolution layer and DeepLabv3 class with custom head for prediction.

For simpler datasets, where the classes are easily distinguishable, or where the features are easier to identify and classify, a Resnet 50 will be a better fit; it's smaller, faster to train and easier to use and deploy. In our work, we use **createDeepLabv3 ResNet 101**, as the backbone network.

Loss function:

The loss function is the function that computes the distance between the current output of the algorithm and the expected output. For each prediction that we make, our loss function will simply measure the absolute difference between our prediction and the actual value. There is different loss function types classification seen in the state of the art as follows:

1. *Regression Types*: Mean Square Error (MSE) / Quadratic Loss / L2 Loss, Mean Absolute Error (MAE) / Mean Square Logarithm Error (MSLE), L1 Loss, Huber Loss / Smooth Mean Absolute Error (combination of MSE and MAE loss functions), Log-Cosh Loss (logarithm of the hyperbolic cosine), Quantile Loss, Mean Bias Error, Likelihood loss.
2. *Classification types*: Binary Cross-Entropy Loss / Log Loss, Hinge Loss.

We used MAE, MSE, MSLE and variations of Huber loss functions for our experiments.

Parameter settings:

- Trainable parameters: 60996202
- Training time: 8.6 minutes / epoch X 10 @ Nvidia GeForce GTX 1080Ti GPU x 2, RAM-11GB, Batch size: 4 images, Train model weight size: 235MB
- Training set: 2528, Validation set: 506,
- Total: 2528 ground truth slides

Dataset Preparation - Steps involved:

Since this challenge has video as the input source, we needed to extract frames from these video files for feeding deep model. Therefore, for dataset preparation we used following steps:

1. The videos frames per second (fps) is 25
2. Use video files to extract images, 25 frames per second video
3. That is, if the video is of X length, the total number of frames will be X(minute)*60(second)*25(fps)
4. Once the images are extracted, they are categorized into training and validation sets
5. There are three types of images: no title, same title and new title slides

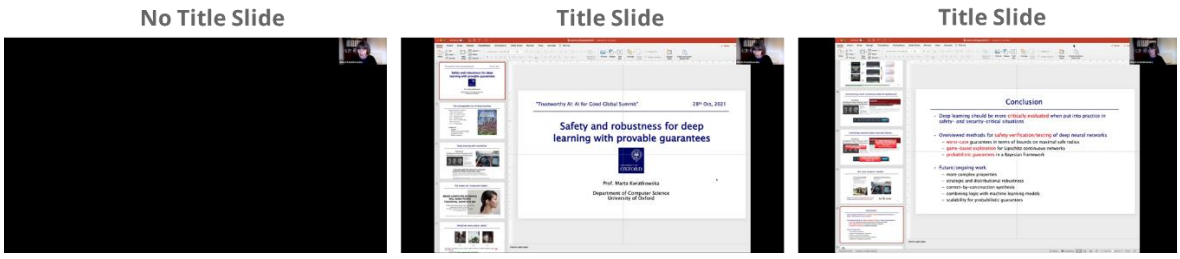


Figure 7: Title and No Title slide data extraction.

Dataset Preparation - Creating training and validation sets

- ❑ From the ground truth CSV file, we extracted the number of frames with the same titles by using starting and ending frame numbers
- ❑ Split the computed frame number by 80%-20%
- ❑ Using the frame number, split the dataset

groundtruth.csv

	A	B	C	D	E	F	
	starting_frame1	starting_frame2	ending_frame1	ending_frame2	title1	title2	title3
	35	35	219	219	Safety and robustness for deep learning with pro		
	220	220	954	954	Safety and robustness for deep learning with pro		
	955	955	3217	3217	The unstoppable rise of deep learning		
	3218	3218	3258	3258	Should we worry about safety?		
	3259	3259	3273	3273	The stakes are rising even higher...		
	3274	3274	3288	3288	Deep learning with everything		
	3289	3289	3314	3314	The unstoppable rise of deep learning		
	3315	3315	3354	3354	Deep learning with everything		

```
train:  147  and validation:  37 out of  184  total frames
train:  587  and validation:  147 out of  734  total frames
train:  1810  and validation:  452 out of  2262  total frames
train:  32  and validation:  8 out of  40  total frames
train:  11  and validation:  3 out of  14  total frames
train:  11  and validation:  3 out of  14  total frames
train:  20  and validation:  5 out of  25  total frames
train:  31  and validation:  8 out of  39  total frames
train:  46  and validation:  12 out of  58  total frames
train:  84  and validation:  21 out of  105  total frames
train:  132  and validation:  33 out of  165  total frames
train:  26  and validation:  6 out of  32  total frames
train:  156  and validation:  39 out of  195  total frames
train:  45  and validation:  11 out of  56  total frames
train:  22  and validation:  6 out of  28  total frames
train:  254  and validation:  64 out of  318  total frames
train:  215  and validation:  54 out of  269  total frames
train:  105  and validation:  26 out of  131  total frames
train:  46  and validation:  12 out of  58  total frames
train:  459  and validation:  115 out of  574  total frames
train:  516  and validation:  129 out of  645  total frames
```

Figure 8: Illustration of training and validation sets of datasets in '*dataset.part01'*.

Then main task is to detect the title from the slides. We have from the baseline of the ground truth dataset with titles. Hence, we have extracted frames from these video files for feeding deep model with title and no title slides as shown in Figure 7 and Figure 8.

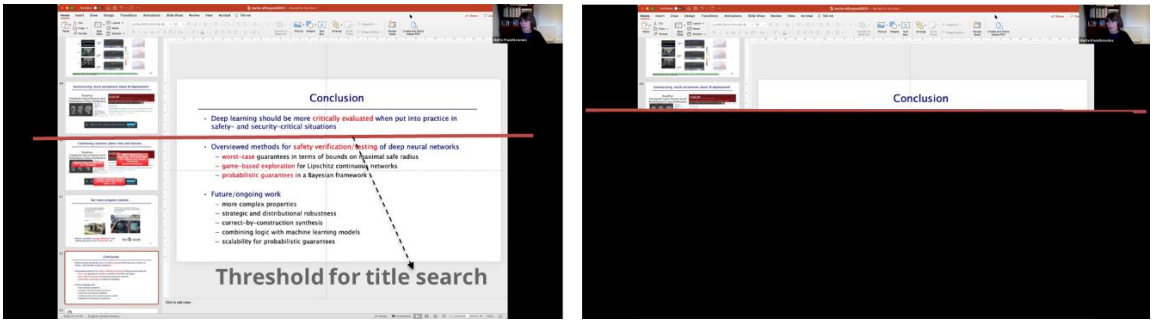


Figure 9: Slide masking based on the threshold for title search.

On an average we assume that titles are seen on the top of the slides. So, from each slide frame we assume that 1/3rd of the portion represents the probability of title search and used for training. Whereas the remaining 2/3rd represents the no title slide. In order to distinguish them we create a mask (white=1 or black=0) with a threshold for title search line represented in red color as shown in Figure 9. Rather than annotation, we create mask which are easy logic to implement.

Key contribution and advantages

- ❑ Model used ResNet101 (createDeepLabv3)
- ❑ No architectural changes
- ❑ Negligible model computation cost (equivalent to the original)
- ❑ Multi-loss training strategy converges the network much faster
- ❑ Gives a significant boost in the performance by ~5%

Results

Environment - Software and Hardware requirements

- Linux ubuntu 18.04 LTS
- Python - 3.9.13
- PyTorch
- Sklearn
- Cuda – 11.6
- Nvidia GeForce GTX 1080Ti GPU x 2, RAM-11GB

Comparison

Table 1: Comparative study of baseline vs proposed multi-loss (4).

Metrics	Baseline	Proposed Multi-loss (4)
Epochs	10	10
Train Loss	0.00279	0.00192
Train F1	0.853	0.911
Train AUROC	0.993	0.990
Test Loss	0.0241	0.0200
Test F1	0.764	**0.812**
Test AUROC	0.925	**0.949**

Ablation study

We compared our methods with several settings which are shown below. The first experiment is to show that the multi-loss concept actually helps the network boost the performance without much computational overhead. Figure 10 shows a bar graph comparison with several loss functions of the students in the classroom of learning. As we see that the multi-loss with four students outperforms the baseline method with roughly 5%, which is really a significant number. That means there are different losses that can recognize different title captions at different times of feature representations. This learning strategy actually reduces the computational cost and instead of several epochs, our model converges much faster. The cost cut is roughly 37.2%, as shown in Figure 10.

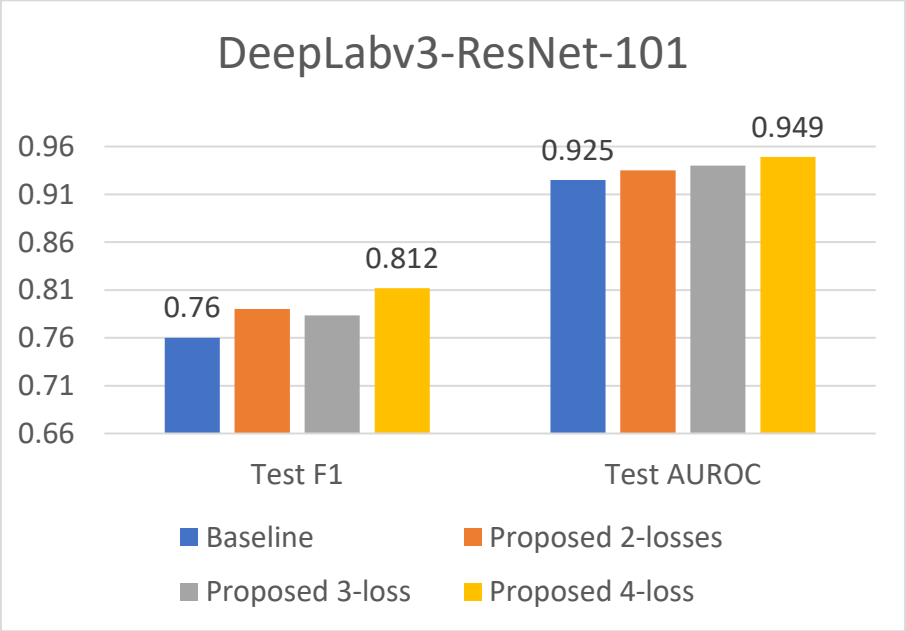


Figure 10: Simulation results of Test F1 and AUROC.

- In statistical analysis of binary classification, the F-score or F-measure or F1 score is a measure of a test's accuracy. The evaluation is based on F1-score that combines the precision and recall into a single metric by taking their harmonic mean defined by:

$$F_1 = 2 * \frac{precision * recall}{precision + recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \frac{1}{2}(\mathrm{FP} + \mathrm{FN})}$$

Numerical calculation for boost in performance of F1= (0.812-0.76)x100=5.2%

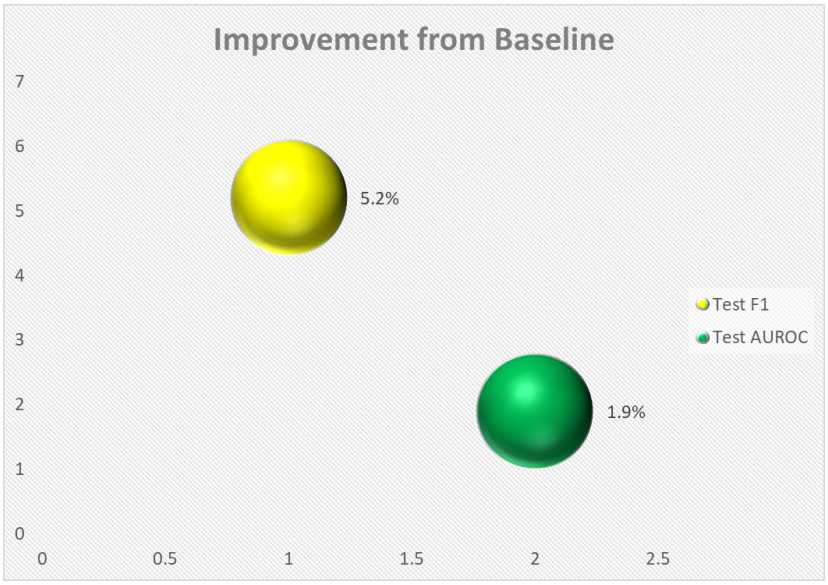


Figure 11: Improvement from baseline work.

- AUROC is the cost difference of accuracy between the ROC curve for the baseline and proposed: Basically, area under the curve.

  Numerical calculation for boost in performance of AUROC= (0.949-0.925)x100=1.9%

From the ablation study in terms of computation, we observe that not all loss is suitable for this challenge task. We had tried for a 5-6-7 loss, but not much improvement was seen. Also, there is no point in simply using all loss functions without knowing their purposes. We observed that the best computational results were obtained at 4-loss functions.

Therefore, the proposed 4-loss classroom-based learning strategy uses a computation cost of 86/137*100=62.8%, resulting in saving of 37.2% of computation cost compared to the baseline.

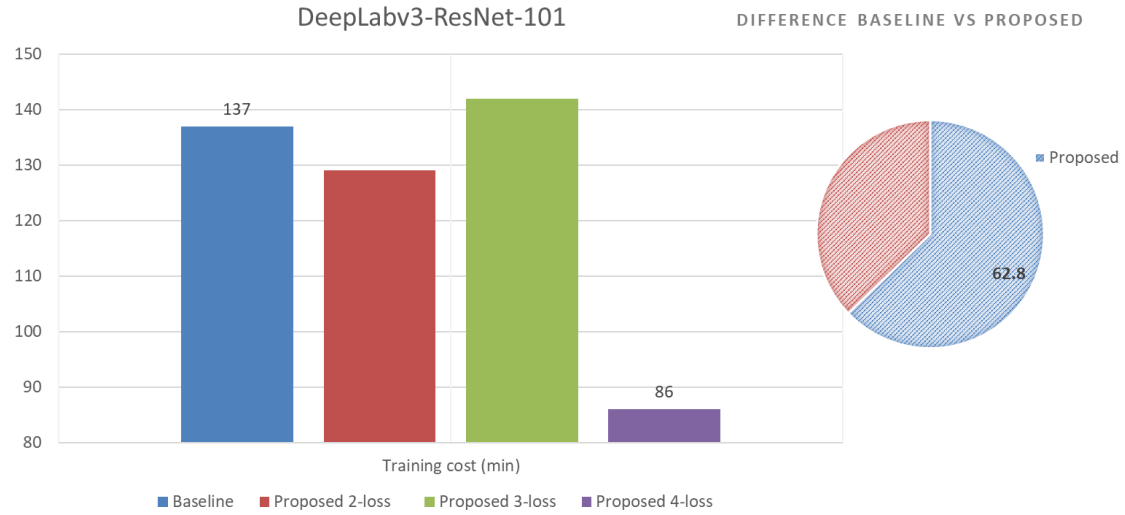


Figure 12: Simulation results in terms of computational cost.

Conclusion

In this report, we proposed a multi-loss-based progressive multi-view learning approach for segmentation and prediction. The proposed method boosts the network learning capability and converges must faster with a significant performance improvement. Among these, the proposed 4-loss uses a computation cost of 62.8%, resulting in saving 37.2% of the computation cost from the baseline work. Also, to show the efficacy of the proposed methods there is a significant improvement observed in the F1 score and AUROC with values of 5.2% and 1.9% respectively.

In the deep digital world, only creating a complex network architecture does not work for all types of datasets or tasks. Sometime there has to be a proper thinking if learning strategy along with the parameters tuning. We summarized few take-ins as below:

- Not very important to improve the network architectural to improve the model performance
- Training strategy is important to optimize the learning
- Multi-loss training strategy converges the network
- Significant boost in the performance by simply involving several loss functions for same task
- Gradient calculation is optimized for detection and segmentation

In the future work, we would like to further optimize the learning curve with a minimal number of learning parameters.

References

[1]. Shitala Prasad, Tingting Chai, Jiahui Li, Zhaoxin Zhang, CR Loss: Improving Biometric Using ClassRoom Learning Approach, *The Computer Journal*, 2022, bxac134, https://doi.org/10.1093/comjnl/bxac134

[2] Prasad, S., Li, Y., Lin, D., Dong, S. and Nwe, M.T.L., 2021. A Progressive Multi-View Learning Approach for Multi-Loss Optimization in 3D Object Recognition. *IEEE Signal Processing Letters*, 29, pp.707-711.

[3] Convolutional-Block-Attention Dual Path Networks for Slide Transition Detection in Lecture Videos. In: Zhai, G., Zhou, J., Yang, H., An, P., Yang, X. (eds) Digital TV and Wireless Multimedia Communication. *IFTC 2019. Communications in Computer and Information Science*, vol 1181. Springer, Singapore, 2019.

[4] SPaSe - Multi-Label Page Segmentation for Presentation Slides; Monica Haurilet, Ziad Al-Halah, Rainer Stiefelhagen; Winter Conference on *Applications of Computer Vision*.

Contact

To contact us, please send an email to: anuj1986aei@gmail.com or shitala@ieee.org