Slide Title Extraction for Zoom-format Lecture Videos through MASK R-CNN

Sam Gorman
sgorman@stanford.edu

Seyi Olujide
oluseyi@stanford.edu

Kaili Wang
kkwang22@stanford.edu

Abstract

Given an online lecture containing slides on Zoom, we identify the major sections of the lecture based on the title. This output may be used to generate a linked "table of contents" for any online video presentation, enabling students to directly access key parts of a lecture recording. We employed a MASK R-CNN to achieve an 88% IoU at a threshold of 50.

1. Introduction

In the COVID-19 crisis, where millions of students are now primarily learning in an online context, there is profound need for the ability to empower learners to access content asynchronously and filter by what they most need. More technically, we approach this as entirely an image problem, where the images are frames from the lecture video. This is because slideshows themselves are sequences of static images, where there is little depth in the image as there would be from an image taken from a camera. The input to our algorithm is an image, representing a frame from a lecture video. We then use a MASK R-CNN to output bounding box coordinates of where a title is present, and an integer confidence value for this prediction.

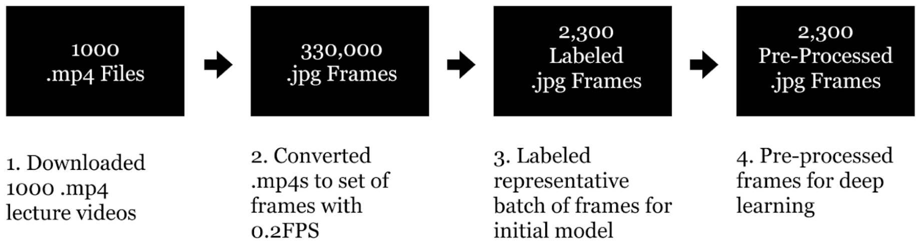
2. Related Work

There exists a substantive body of research focused on improving online learning. Imran et al used lecture audio and slides to create word clouds to indicate what each segment of the video was about [1]. Quite a few others made advancements on slide transition detection between 2014 and 2016 [2][3]. At first, we expected to rely on detection of slide transitions in order to identify whether a new section has been reached. However, a literature review led us to decide to focus on detecting titles rather than having an individual segment of the process dedicated to detecting slide transitions. A previous technology that has been developed and patented achieved a somewhat similar challenge as ours; it extracted metadata (titles, authors, addresses, and more) from documents of the similar format [4]. This project first applied OCR to the entire document, and then used deep learning to identify which text lies in which category (with some NLP processing). Additionally, Yang et Al employed OCR to generate lecture outlines of slide presentations, but this OCR-driven approach could not meaningfully distinguish between titles and plaintext. [5] This led us to adopt an approach driven by deep-learning, with no OCR present.

3. Dataset and Features

We constructed a novel dataset, which we have titled “Online Learning Repository”, consisting of 330,000 .jpgs of online lecture recording frames. All frames are extracted from a repository of 1000 videos from the Advanced Placement Youtube Channel. This channel was chosen for strong consistency of format: each uploaded video is a recorded Zoom lecture in which a professor screenshares a presentation. Further, the channel has high variance in subject matter, which we categorized and tagged under government, arts, english, history & social science, math & computer science, sciences, and world languages.

Data pipeline overview:


1000
.mp4 Files
330,000
.jpg Frames
2,300
Labeled
.jpg Frames
2,300
Pre-Processed
.jpg Frames
1. Downloaded 1000 .mp4 lecture videos
2. Converted .mp4s to set of frames with 0.2FPS
3. Labeled representative batch of frames for initial model
4. Pre-processed frames for deep learning

As displayed in step 3, a representative batch of 2,300 frames were chosen by randomly selecting accompanying frames for at least one video across each subject category. These frames were then manually labeled in LabelBox by drawing bounding boxes where a title was present. MASK

R-CNN handles rescaling and transforms of images. Finally, we split our data into a size-600 test set and a size-1700 training set, and exported these sets as JSON input for our model.

4. Methods

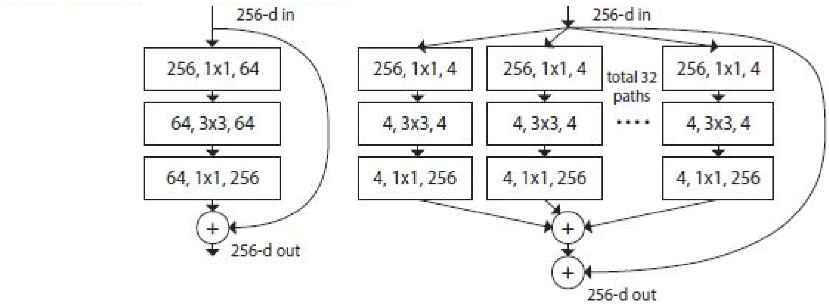
Data manipulation and models were built with TensorFlow[8], Pytorch[10] and Facebook Detectron2[9].

4.1 Baseline SSD MobileNet V2

We first employed SSD MobileNetV2 as an off-the-shelf baseline as part of TensorFlow's Object Detection Library. This model achieved an accuracy of 60% on our test set, leading us to shift to the more computationally expensive but more accurate MASK R-CNN architecture.

4.2 Transfer learning with MASK R-CNN

We then applied transfer learning with our novel dataset on a pre-trained state-of-the-art MASK-R-CNN implementation in Facebook Detectron2.The model, titled ResNeXt-101-32x8d is trained with Caffe2 and uses a ResNet + FPN backbone with standard convolution. This model efficiently detects objects in an image while generating a high-quality segmentation mask for each instance. MASK R-CNN is an extension of the Faster R-CNN architecture, including an additional mast head to perform pixel wise segmentation on each object and extract each object separately irrespective of background.



Left: A block of ResNet [14]. Right: A block of ResNeXt with cardinality = 32, with roughly the same complexity. A layer is shown as (# in channels, filter size, # out channels).

[6]

The model employs a multi-task loss function as enumerated below, where *L* is the total loss function is , *Lcls* is the classification loss *, Lbox* is the loss of the bounding box, *and Lmask* is the average binary cross-entropy loss [7].

$$L = L_{cls} + L_{box} + L_{mask}$$

Figure 1. Multi-task Loss Function for Mask R-CNN

$$-\frac{1}{m^2}\sum_{1\leq i,j\leq m}\left[y_{ij}\log\hat{y}^{k}_{ij}+(1-y_{ij})\log(1-\hat{y}^{k}_{ij})\right]$$

Figure 2. Average binary cross-entropy loss function for *Lmask*

5. Experiments / Results / Conclusions

We decreased the learning rate from its default to 0.0005, because our dataset was relatively small and thus computational expense was not a huge concern. Otherwise, we mostly used the default hyperparameters as in ResNeXt-101-32x8d, our transferred learning model.

5.1 Accuracy

We achieved a result of 0.88 IoU at threshold 50 on our test set. In other words, our model correctly identified when *(area of overlap / area of union > 0.5*) in 88% of the test set. To the extent the authors are aware, this results represents the best-in-class result on detection of titles from lecture slides.

Qualitatively speaking, our model performed well on standard slides, where the title is most a few words long (see Appendix for successful examples). It performed well on the rare instances of slides with titles in the lower half of the frame (Ex. 2). It even performed well when it was a non-screenshared image--in other words, an image from an actual camera (Ex. 4). This was unexpected, because we assumed that any camera tilt might skew the results. This gives us hope that our model would perform well in real in-person lectures, where the video is taken from the back of the room.

5.2 Loss Function

We observed a loss function approaching 0.4 after 300 steps, which is an acceptable level of loss relative to both the high IoU we observed and benchmarks set in prior literature.

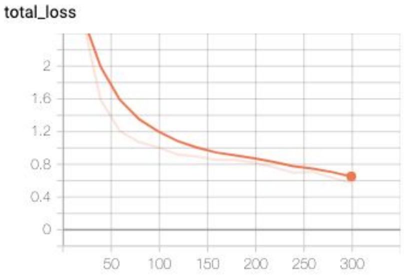


Figure 3: Loss Function Output of MASK R-CNN

5.3 Failures of the Model

Our model, however, sometimes was not able to identify the titles in some ambiguous slides, such as those listed in the Appendix (examples 5-9). We believe that most of these errors may be attributed to the limited size and variance within the training set.

5.3.1 Possible Cases of Overfitting

Since all of our data came from online AP videos on Youtube, there was undoubtedly overfitting to those presentation formats. Additionally, after conducting error analysis, we noticed that almost all of the multi-line titles were not correctly identified; only the first line of the title was boxed (see Ex. 10 in Appendix). This tells us that there must have been overfitting for one-line titles.

When we ran our model on our training data, there was a result of 0.836 IoU. We admittedly are not exactly sure why there was a worse outcome on the training data than on the unseen test data.

6. Conclusions

In summary, our model generated positive results of 0.88 IoU on our test set, using transfer learning on a novel dataset with a MASK R-CNN architecture.

With more time and team members, we see an opportunity to create a user-facing product that deploys our model via a web application. Given an mp4 video lecture, our application may output a table of contents containing titles of each lecture slides, and timestamps to jump to that location in a video.

Further, it may be beneficial to expand the size of our labeled dataset and reduce potential overfitting, as while we have access to 330,000+ frames, human labeling constraints led us to employ transfer learning with a relatively small dataset of 2,300. Further, we seek to expand our sources of data as well, going beyond the Advanced Placement channel we extracted frames from to account for higher diversity of data.

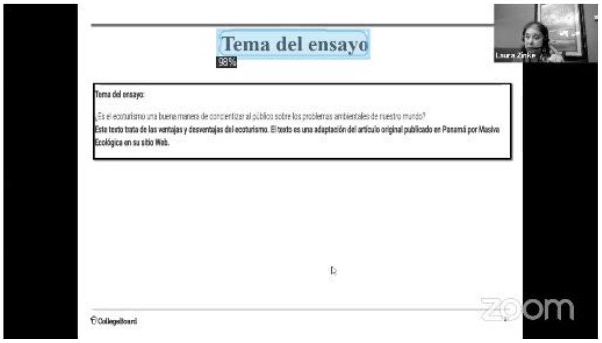
We plan to release our growing labeled dataset publicly to support future research efforts.

7. Contributions

Kaili Wang contributed to implementing our MASK R-CNN and determining which model to use. Seyi Olujide contributed to gathering raw data and ensuring proper storage and access methods to the data. Sam Gorman contributed to preprocessing of the data and construction of the novel datatset. All authors contributed equal amounts of work.

8. Appendix

Examples of Good Results

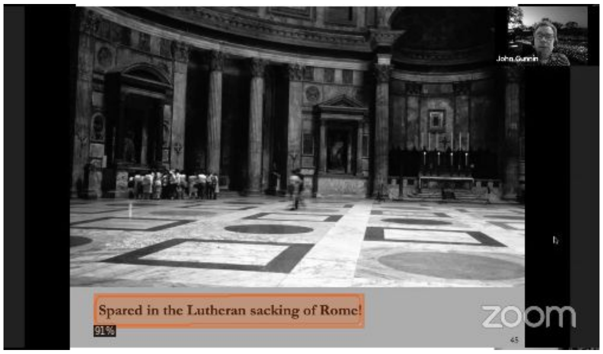


Tema del ensayo

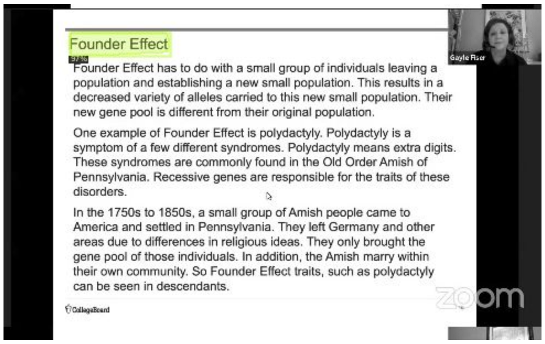
Tema del ensayo:

¿Es el ecoturismo una buena manera de concientizar al público sobre los problemas ambientales de nuestro mundo?
Este texto trata de las ventajas y desventajas del ecoturismo. El texto es una adaptación del artículo original publicado en Panamá por Masivo Ecológica en su sitio Web.

Ex. 1: Correct prediction with 98% confidence



Ex. 2: Correct prediction of caption in lower half, with 91% confidence



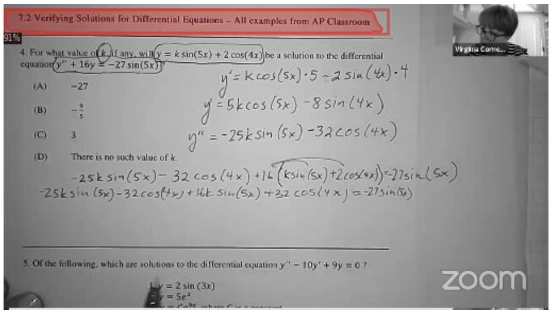
Founder Effect

Founder Effect has to do with a small group of individuals leaving a population and establishing a new small population. This results in a decreased variety of alleles carried to this new small population. Their new gene pool is different from their original population.

One example of Founder Effect is polydactyly. Polydactyly is a symptom of a few different syndromes. Polydactyly means extra digits. These syndromes are commonly found in the Old Order Amish of Pennsylvania. Recessive genes are responsible for the traits of these disorders.

In the 1750s to 1850s, a small group of Amish people came to America and settled in Pennsylvania. They left Germany and other areas due to differences in religious ideas. They only brought the gene pool of those individuals. In addition, the Amish marry within their own community. So Founder Effect traits, such as polydactyly can be seen in descendants.

Ex. 3: Correct prediction with 97% accuracy



7.2 Verifying Solutions for Differential Equations – All examples from AP Classroom

4. For what value of k, if any, will $y = k\sin(5x) + 2\cos(4x)$ be a solution to the differential equation $y'' + 16y = -27\sin(5x)$?

(A) -27

(B) $-\frac{9}{5}$

(C) 3

(D) There is no such value of k.

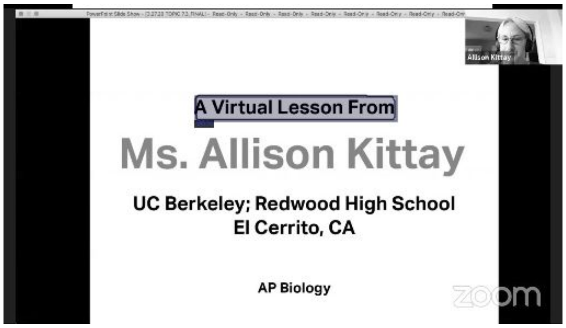
$$y' = k\cos(5x)\cdot 5 - 2\sin(4x)\cdot 4$$
$$y' = 5k\cos(5x) - 8\sin(4x)$$
$$y'' = -25k\sin(5x) - 32\cos(4x)$$
$$-25k\sin(5x) - 32\cos(4x) + 16(k\sin(5x) + 2\cos(4x)) = -27\sin(5x)$$
$$-25k\sin(5x) - 32\cos(4x) + 16k\sin(5x) + 32\cos(4x) = -27\sin(5x)$$

5. Of the following, which are solutions to the differential equation $y'' - 10y' + 9y = 0$?

I. $y = 2\sin(3x)$
$y = 5e^{x}$

Ex. 4: Correct prediction in a camera mode, not screen-share mode, with 91% prediction.

Examples of Ambiguous Results



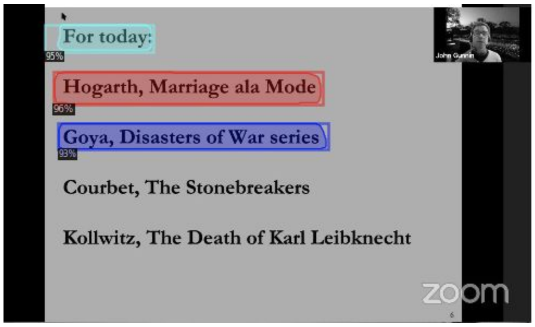
A Virtual Lesson From

Ms. Allison Kittay

UC Berkeley; Redwood High School
El Cerrito, CA

AP Biology

Ex. 5. Ambiguous slide. Title probably should be "A Virtual Lesson From Ms. Allison Kittay"



For today:

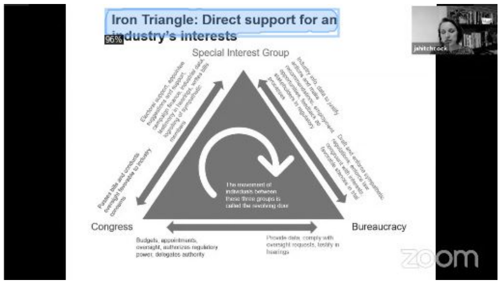
Hogarth, Marriage ala Mode

Goya, Disasters of War series

Courbet, The Stonebreakers

Kollwitz, The Death of Karl Leibknecht

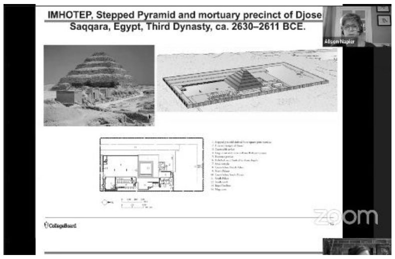
Ex. 6. Ambiguous slide. Title probably should be "For today:" Assigned title status to correct title with 95% confidence, but assigned title status to incorrect answer with 96% confidence


96%
jahitchcock
Special Interest Group
Congress
Bureaucracy
Budgets, appointments, oversight, authorizes regulatory power, delegates authority
Provide data, comply with oversight requests, testify in hearings
The movement of individuals between these three groups is called the revolving door
zoom

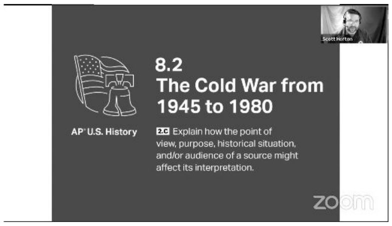
Iron Triangle: Direct support for an industry's interests

Ex. 7: Prediction only included part of the second line, which is not ideal.

Examples of Erroneous Results


IMHOTEP, Stepped Pyramid and mortuary precinct of Djose
Saqqara, Egypt, Third Dynasty, ca. 2630–2611 BCE.

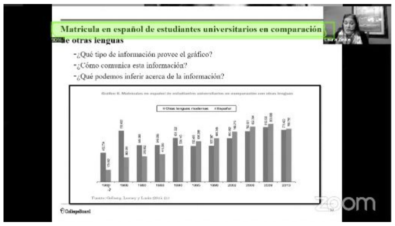
Ex. 8: Did not identify title


8.2
The Cold War from 1945 to 1980

AP U.S. History

Explain how the point of view, purpose, historical situation, and/or audience of a source might affect its interpretation.

Ex. 9: Did not identify title


Matricula en español de estudiantes universitarios en comparación de otras lenguas

-¿Qué tipo de información provee el gráfico?
-¿Cómo comunica esta información?
-¿Qué podemos inferir acerca de la información?

Ex. 10: Only identifies first line of title

References

[1] Imran, Ali Shariq, et al. “Automatic Annotation of Lecture Videos for Multimedia Driven Pedagogical Platforms.” *Knowledge Management & E-Learning*, vol. 8, no. 4, 2016, p. 550. *Advanced Technologies & Aerospace Collection*, search-proquest-com.stanford.idm.oclc.org/docview/1955087720?accountid=14026. Accessed 10 May 2020.

[2] Jeong, Hyun & Kim, Tak-Eun & Kim, Hyeon Gyu & Kim, Myoung Ho. (2014). Automatic detection of slide transitions in lecture videos. Multimedia Tools and Applications. 74. 10.1007/s11042-014-1990-6.

[3] Zhang, Xiangrong, et al. "Automated segmentation of MOOC lectures towards customized learning." *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*. IEEE, 2016.

[4] Shmueli, Oded, et al. *Automatic Extraction of Metadata Using a Neural Network*. US6044375A, 28 Mar. 2000, https://patents.google.com/patent/US6044375A/en?q=Automatic+extraction+metadata+neural+network&oq=Automatic+extraction+of+metadata+using+a+neural+network.

[5] Yang, Haojin et al. “Automated Extraction of Lecture Outlines from Lecture Videos - A Hybrid Solution for Lecture Video Indexing.” *CSEDU* (2012).

[6] S. Xie, R. Girshick, P. Dollár, Z. Tu and K. He, "Aggregated Residual Transformations for Deep Neural Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 5987-5995, doi: 10.1109/CVPR.2017.634.

[7] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, 2017, pp. 2980-2988, doi: 10.1109/ICCV.2017.322.

[8] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis,Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow,Andrew Harp, Geoffrey Irving, Michael Isard, Rafal Jozefowicz, Yangqing Jia,Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Mike Schuster, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Jonathon Shlens,Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker,Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas,Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke,Yuan Yu, and Xiaoqiang Zheng.TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[9] Yuxin Wu and Alexander Kirillov and Francisco Massa and Wan-Yen Lo and Ross Girshick, Detectron2,https://github.com/facebookresearch/detectron2,2019

[10] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d extquotesingle Alch&#39;e-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 8024–8035). Curran Associates, Inc. Retrieved from http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf