Abstract

Note-taking is a **universal activity** among students because of its benefits to the learning process. This research focuses on **end-to-end generation** of formatted **summaries** of lecture videos.

Our automated **multimodal** approach will

- **decrease** the **time** required to create notes
- **increase quiz scores** and content knowledge
- **enable faster learning** through enhanced previewing

The project is broken into **three main components:** the slide classifier, summarization models, and end-to-end-process.

The system begins by **extracting important keyframes** using the slide classifier, a deep **CNN**. Then, unique slides are determined using a combination of **clustering** and **keypoint matching**. The structure of these unique slides is analyzed and converted to a formatted transcript that includes **figures** present on the slides.

The **audio** is transcribed using one of several program, including Vosk, Wav2Vec, & Sphinx.

We approach the process of **combining** and **summarizing** these **transcripts** in several ways including as **keyword-based sentence extraction** and **temporal audio-slide-transcript association** problems.

For the **summarization** stage, **state-of-the-art** models are used, including novel models specifically for this project, which are collectively called "**TransformerSum**" and advance the state-of-the-art in **long** and **resource-limited summarization**. **Extractive** and **abstractive** approaches are used in conjunction to summarize the **long-form** content extracted from the **lectures**.

Results

Slide Classifier

Model	WER	MER	WIL	LS TC WER	Processing Time
DeepSpeech (chunking)	43.01	41.82	59.04	5.97	≈4 hours
DeepSpeech	44.44	42.98	59.99	5.97	≈20 hours
Google STT ("default" model)	34.43	33.14	49.05	12.23	≈20 minutes
Vosk small-0.3	35.43	33.64	50.84	15.34	≈8.5 hours
Vosk daanzu-20200905	33.67	31.87	48.28	7.08	≈5.5 hours
Vosk aspire-0.2	41.38	38.44	56.35	13.64	≈19 hours
Vosk aspire-0.2 (chunking)	41.45	38.65	56.56	13.64	≈19 hours

Speech-To-Text

Model	Dataset	Accuracy	Accuracy (train)	F-score
Final-general	train-test	82.62	98.58	87.44
Three-category	train-test-three	89.97	99.72	93.82
Squished-image	train-test	83.86	97.16	88.16
	train-test-three	87.21	100	91.57

TransformerSum

R1/R2/RL-Sum	CNN/DM	WikiHow	ArXiv/PubMed
distilbert-base-uncased	42.71/19.91/39.18	30.69/08.65/28.58	34.93/12.21/31.00
distilroberta-base	42.87/20.02/39.31	31.07/08.96/28.95	34.70/12.16/30.82
bert-base-uncased	42.78/19.83/39.18	30.68/08.67/28.59	34.80/12.26/30.92
roberta-base	43.24/20.36/39.65	31.26/09.09/29.14	34.81/12.26/30.91
mobilebert-uncased	42.01/19.31/38.53	30.72/08.78/28.59	33.97/11.74/30.19
BertSumExt	43.25/20.24/39.63	None	None
BertSumExt-large	43.85/20.34/39.90	None	None
MatchSum bert-base	44.22/20.62/40.38	31.85/08.98/29.58	None
MatchSum roberta-base	44.41/20.86/40.55	None	None
PEGASUS-base	41.79/18.81/38.93	36.58/15.64/30.01	37.39/12.66/23.87
PEGASUS-large HN	44.17/21.47/41.11	41.35/18.51/33.42	44.88/18.37/26.58
PEGASUS-large C4	43.90/21.20/40.76	43.06/19.71/34.80	45.10/18.59/26.75

Summary

Slide Classifier
- Identify important frames in slide presentations

Summarization Models
- Improve state-of-the-art in text summarization
- Novel approaches

End-To-End
- Multimodal approach
- Convert lecture videos to notes

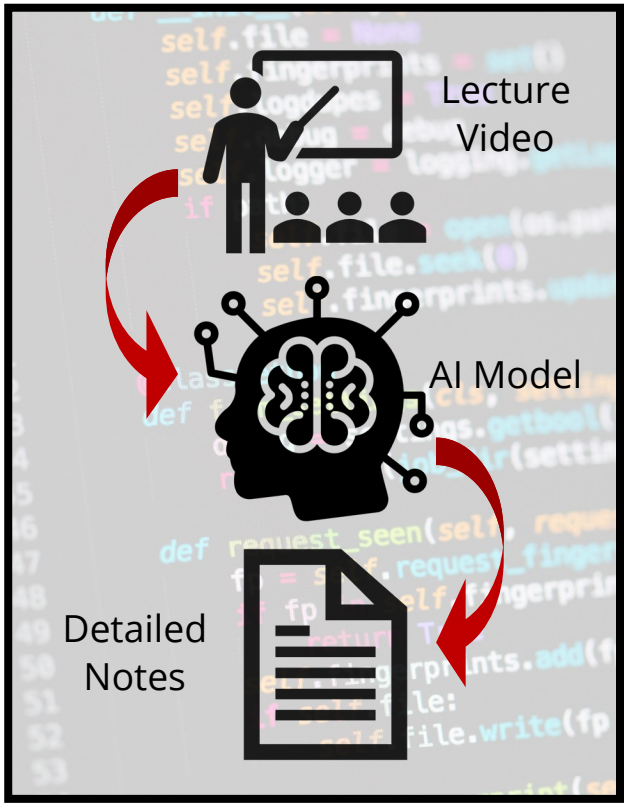
Online Web Service
- Allows anyone to automatically create notes based on their own lectures

References

1. Ballard, D. H., & Brown, C. M. (1982). Computer Vision. Englewood Cliffs, NJ: Prentice-Hall.
2. Huang T. S. (1996). Computer Vision: Evolution and Promise. CERN School of Computing, Geneva: 21–25. Retrieved from https://cds.cern.ch/record/400313/files/p21.pdf
3. Papert, Seymour. (1966). The Summer Vision Project.
4. Szeliski, R. (2011). Computer vision: Algorithms and applications. London: Springer.
5. "Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper." VNI Global Fixed and Mobile Internet Traffic Forecasts. Cisco, 27 February 2019, https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html
6. Artificial intelligence. (n.d.). In Collins Dictionary. Retrieved May 11, 2019, from https://www.collinsdictionary.com/dictionary/english/artificial-intelligence
7. Artificial intelligence. (n.d.). In Merriam Webster. Retrieved May 11, 2019, from https://www.merriam-webster.com/dictionary/artificial intelligence
8. Copeland, B. J. (2019, May 9). Artificial intelligence. Retrieved May 11, 2019, from https://www.britannica.com/technology/artificial-intelligence
9. Machine learning. (n.d.). In Collins Dictionary. Retrieved May 11, 2019, from https://www.collinsdictionary.com/dictionary/english/machine-learning
10. Machine learning. (n.d.). In Merriam Webster. Retrieved May 11, 2019, from https://www.merriam-webster.com/dictionary/machine learning
11. Hosch, W. L. (2016, September 01). Machine learning. Retrieved May 11, 2019, from https://www.britannica.com/technology/machine-learning
12. Neural network. (n.d.). In Collins Dictionary. Retrieved May 11, 2019, from https://www.collinsdictionary.com/dictionary/english/neural-network
13. Zwass, V. (2018, July 27). Neural network. Retrieved May 11, 2019, from https://www.britannica.com/technology/neural-network
14. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90. doi:10.1145/3065386
15. Cleeremans, A., Servan-Schreiber, D., & McClelland, J. L. (1989). Finite State Automata and Simple Recurrent Networks. Neural Computation, 1(3), 372-381. doi:neco.1989.1.3.372
16. Pearlmutter, B. A. (1989). Learning State Space Trajectories in Recurrent Neural Networks. Neural Computation, 1(2), 263-269. doi:neco.1989.1.2.263
17. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780. Doi:neco.1997.9.8.1735

Lecture2Notes

Summarizing Lecture Videos by Classifying Slides and Analyzing Text using Machine Learning



Hayden Housen
www.haydenhousen.com
GitHub: @HHousen
Pawling High School

Introduction

Computer Vision (CV)

- **Definition:** **Interdisciplinary** scientific field that attempts to give computers the **ability** to **understand** digital **images** and **videos**.[1]
- **Goal of CV:** Create **computational models** of functions & abilities **associated** with **human visual system**.[2]

Face Recognition | Emotion Analysis | Sports: draw lines on field | Medical Imaging | Movie Special Effects | Self Driving Cars

Natural Language Processing

- **Definition:** Field that gives computers the **ability to read**, **understand** and **derive** meaning from **human languages**.[28]
- **Goal of CV:** **Read**, decipher, and **understand** of the **human languages** in a manner that is **valuable**.[29]

Spam filters | News tracking | Voice Assistants | Chatbots | Translators | Grammar Checking

Note Taking

- Note taking is almost a **universal activity** among **students**.[30]
- Students' notes are generally **incomplete**, and thus **not adequate** for reviewing the material.[31]
- Students prefer **guided notes**[32] and course final exam performance **higher** for guided notes.[33]

Influential Paper

"BERT Text Summarization Lectures"[23] – Most related to research since it applies **deep learning** to **lecture summarization**. Approach is **limited** because:

- It **only** tests the **BERT** model
- Does **not consider** the information on the **slides**
- Does **not fine-tune BERT** and merely uses it for embeddings
- Has **no STT** component

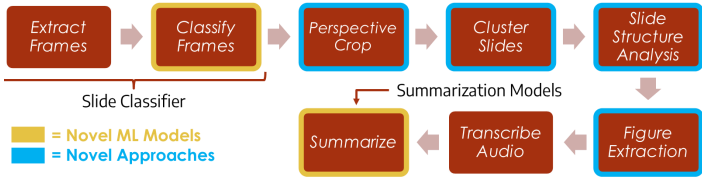
Methodology

	Gap in Research	Purpose
1	There is **no ML model** to **classify slides** in lecture videos or **up-to-date dataset** of **lecture videos**	Train a ML model to classify slides and create a **dataset of lecture videos** that contains many **different topics and styles**.
2	No robust combination-based summarization model has been applied to lecture videos.	Merge various **combination approaches** (keywords, concatenation) with **extractive** and **abstractive summarization** models.
3	No **end-to-end process** to summarize slide presentations	Use various **algorithms** (OCR, clustering, ASR) to **fuse** all **components** together
4	Both a **web app** for public use and an API for developers **do not exist**.	Launch model into **production** though a **user friendly web app**

Slide Classification Model

- **Classify** frames of input video into **9 classes**
- Allows system to **identify images** containing slides for **additional processing**
- Means system can **ignore useless frames** (containing only the presenter or audience)
- **Final Dataset:** Contains **15599 images** extracted from **78 videos** manually classified into **9 categories**

End-to-End Approach



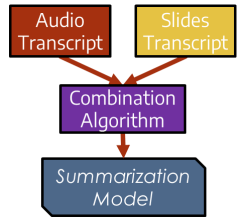
Feature Matching

- **Iterates** through the "screen capture" and "video camera" slides in **chronological** order.
- **Goal:** find "video camera" within "screen capture"
- If number of **matches** is above a **threshold** then the **slides** are considered **identical**.

Speech-To-Text

- Tested performance of **STT** models on **43 lecture videos** from slide classifier **dataset**.
- **Chunking by Voice Activity:** Uses WebRTC Voice Activity Detector – **increases** the **speed** of **STT** by **reducing** the amount of **audio without speech**.

Combination Algorithm

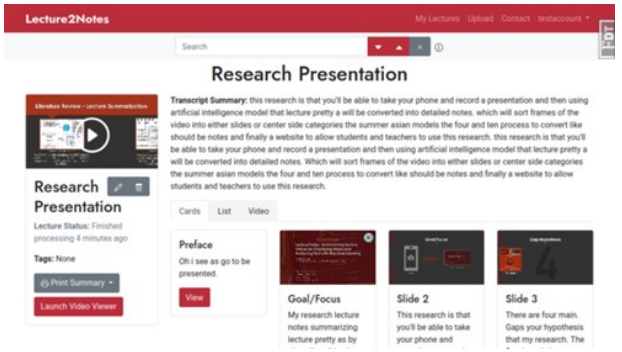


- **only_asr** – only uses the audio transcript (deletes the slide transcript)
- **only_slides** – reverse of only_asr
- **concat** – appends audio transcript to slide transcript
- **full_sents** – audio transcript appended to only the complete sentences of the slide transcript
- **keyword_based** (most advanced) – selects a certain percentage of sentences from the audio transcript based on keywords found in the slides transcript

Summarization Algorithm

- Extractive Summarization
  - **Cluster** – groups lecture transcript into categories & summarizes each using "generic"
  - **Generic** (non-neural) – uses algorithms from *sumy* package: lsa, luhn, lex_rank, text_rank, edmundson, random
- **Abstractive Summarization:** PreSumm or BART or PEGASUS or TransformerSum

Website



A portion of the website showing the summary of a lecture.