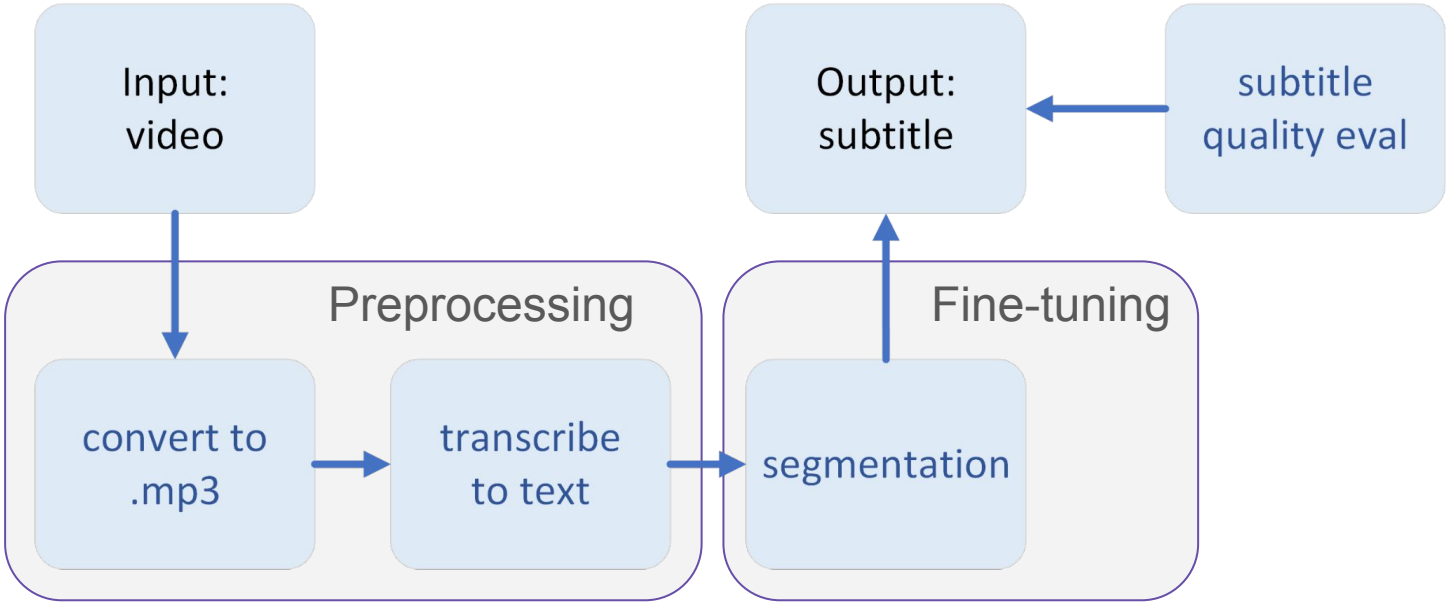
Deep Learning Solution for Precise Subtitle Segmentation

Lilin Jin, Anush Veeranala, Xinyu Zhang

Intro

Motivation: subtitles are crucial for accessibility, academic conferences rely heavily on volunteers for manual transcription.

Pipeline:



Video Preprocessing

Convert Video to Audio: MoviePy

- We need mp3 files for timestamp matching

Transcribe Audio to Text: Whisper by OpenAI

- This allows for precise segmentation of text

Model Anatomy

Words - Punctuations - Segmentation tags

- so I want to talk about diffusion models So this is I guess rounding out the generative image parts
- so I want to talk about diffusion models. So this is, I guess, rounding out the generative image parts.
- so I want to talk<eol> about diffusion models. <eob> So this is,<eob>I guess,<eol> rounding out the generative image parts. <eob>

BOSTON UNIVERSITY

Diffusion Models

DL4DS – Spring 2024

Based on Rocca, 2022, "Understanding Diffusion Probabilistic Models (DPMs)", Towards Data Science

so I want to talk
about diffusion models.

Prince, *Understanding Deep Learning*,

BOSTON UNIVERSITY

Diffusion Models

DL4DS – Spring 2024

Based on Rocca, 2022, "Understanding Diffusion Probabilistic Models (DPMs)", Towards Data Science

so I want to talk |EOL|
about diffusion models. |EOB|

Prince, *Understanding Deep Learning*,

Token classification

- so I want to talk<eol> about diffusion models. <eob> So this is,<eob>I guess,<eol> rounding out the generative image parts. <eob>

So	Class1
I	Class1
want	Class1
to	Class1
talk	EOL
about	Class2
diffusion	Class2
models	EOB

Segmentation

Dataset: MuST-Cinema

- Texts marked with "<eol>" (End of Line) and "<eob>" (End of Block) for subtitle segmentation
- Example:

> Thank you so much, Chris. <eob> And it's truly a great honor <eol> to have the opportunity <eob> to come to this stage twice; <eol> I'm extremely grateful. <eob>

Pre-trained Model: xlmr-multilingual-sentence-segmentation

- Finetuned from FacebookAI/xlm-roberta-base
- Purpose: Identifies sentence boundaries in text; not for subtitle segmentation
- Example:

> Thank you so much, Chris . And it's truly a great honor to have the opportunity to come to this stage twice; I'm extremely grateful .

Segmentation

Model Input			Thank you so much, Chris. And it's truly a great honor																										
Ground True			Thank you so much, Chris. <eob> And it's truly a great honor <eol>																										
Pre-trained Model's Tokenizer	Tokens	<s>	_Thank	_you	_so	_much	,	_Chris	.	_<	e	ob	>	_And	_it	'	s	_truly	_a	_great	_honor	_<	e	ol	>	...	</s>
	Token IDs	0	25689	398	221	5045	4	31745	5	4426	13	3522	2740	3493	442	25	7	87607	10	6782	20338	4426	13	929	2740	...	2
Target		0	0	0	0	0	0	0	2					0	0	0	0	0	0	0	1					...	0
Pre-trained Model Output		0	0	0	0	0	0	0	1					0	0	0	0	0	0	0	0					...	0

1. Change the Classifier Layer

```
model.classifier =
torch.nn.Linear(model.classifie
r.in_features, 3)
model.num_labels = 3
```

2. Adjust Class Weights in Cross-Entropy Loss

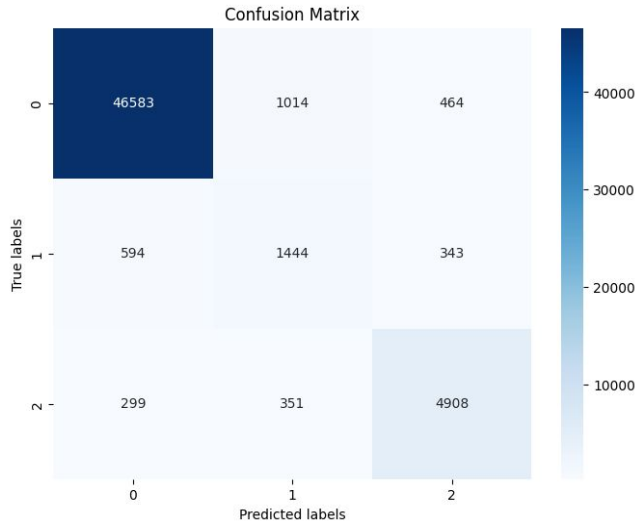
```
weights =
torch.tensor([1.0, 2.5, 2]).to(device)
criterion =
nn.CrossEntropyLoss(weight=weights,
ignore_index=-1)
```

Segmentation

Other Training Details:

- Batch Size: 16
- Padding: Fixed length of 330 (max length 300); Masks used to ignore padding
- Epochs: 4
- Optimizer: AdamW with a learning rate of 2e-5 and weight decay of 0.0001
- Learning Rate Scheduler: Linear decrease, no warmup
- Optimization Steps: Every 2 steps, gradients clipped at 4.0

Segmentation



F1 Score on Test Dataset Using Macro-Average: **0.801**

Subtitle File Generation: Aeneas

Using Dynamic Time Warping (DTW) algorithm to synchronize text and audio.

Input text (.txt)	**Sync map (.json)**	**Output subtitle file (.srt)** (most popular format for subtitle)

OK,<eob>
so I want to talk<eol>
about diffusion models.
<eob>
So this is,<eob>
I guess,<eol> rounding out the generative image parts. <eob>
	{"fragments": [{
	"begin": "0.000","children": [],
	"end": "2.920","id": "f000001",
	"language": "eng",
	"lines": ["**OK,<eob>**"]},
	{
	"begin": "2.920","children": [],
	"end": "6.640","id": "f000002",
	"language": "eng",
	"lines": ["**so I want to talk<eol> about diffusion models. <eob>**"]},
		1
		00:00:00,000 --> 00:00:02,920
		OK,
		
		2
		00:00:02,920 --> 00:00:06,640
		so I want to talk
		about diffusion models.

Evaluation

Human evaluation (Let us know from the demo :))

F1 Score: 0.544

Demo Video

Link: https://mymedia.bu.edu/media/t/1_0hmhg3kg/340650712

Further improvements

A more comprehensive **subtitle evaluation scheme**: Current scoring methods all rely on ground truth.

Content summarize function: restriction of the text window of open source LLMs

Questions/Concerns?