

Masked Language Modeling YouTube Captioning Errors?

Emily McMilin

Masked Language Modeling YouTube Captioning Errors?

Emily McMilin



Motivation

Motivation: stupid or..?

The fastai book - draft

These draft notebooks cover an introduction to deep learning, [fastai](#), and [PyTorch](#). fastai is a layered API for deep learning; for more information, see [the fastai paper](#). Everything in this repo is copyright Jeremy Howard and Sylvain Gugger, 2020 onwards.

These notebooks will be used for [a course we're teaching](#) in San Francisco from March 2020, and will be available as a MOOC from around July 2020. In addition, our plan is that these notebooks will form the basis of [this book](#), which you can pre-order. It will not have the same GPL restrictions that are on this draft.

The code in the notebooks and python `.py` files is covered by the GPL v3 license; see the LICENSE file for details.

The remainder (including all markdown cells in the notebooks and other prose) is not licensed for any redistribution or change of format or medium, other than making copies of the notebooks or forking this repo for your own private use. No commercial or broadcast use is allowed. We are making these materials freely available to help you learn deep learning, so please respect our copyright and these restrictions.

If you see someone hosting a copy of these materials somewhere else, please let them know that their actions are not allowed, and may lead to legal action. Moreover, they would be hurting the community, because we're not likely to release additional materials in this way if people ignore our copyright.

This is an early draft. If you get stuck running notebooks, please search the [fastai-v2 forum](#) for answers, and ask for help there if needed. Please don't use GitHub issues for problems running the notebooks.

If you make any pull requests to this repo, then you are assigning copyright of that work to Jeremy Howard and Sylvain Gugger. (Additionally, if you are making small edits to spelling or text, please specify the name of the file and very brief description, with the original text, so that we can merge it quickly and know which corrections have already been made. Thank you.)

read this thing for free as stupid or
notebooks but that is not as convenient

2:31 / 1:22:30 • Introduction

Motivation: stupid or...?



What you don't need, to do deep learning

Myth (don't need)	Truth
Lots of math	Just high school math is sufficient
Lots of data	We've seen record-breaking results with <50 items of data
Lots of expensive computers	You can get what you need for state of the art work for free

learn deep learning or are you too
stupid or you don't have enough vast

Motivation: This MLM needs a label

The fastai book - draft

These draft notebooks cover an introduction to deep learning, [fastai](#), and [PyTorch](#). fastai is a layered API for deep learning; for more information, see [the fastai paper](#). Everything in this repo is copyright Jeremy Howard and Sylvain Gugger, 2020 onwards.

These notebooks will be used for [a course we're teaching](#) in San Francisco from March 2020, and will be available as a MOOC from around July 2020. In addition, our plan is that these notebooks will form the basis of [this book](#), which you can pre-order. It will not have the same GPL restrictions that are on this draft.

The code in the notebooks and python `.py` files is covered by the GPL v3 license; see the LICENSE file for details.

The remainder (including all markdown cells in the notebooks and other prose) is not licensed for any redistribution or change of format or medium, other than making copies of the notebooks or forking this repo for your own private use. No commercial or broadcast use is allowed. We are making these materials freely available to help you learn deep learning, so please respect our copyright and these restrictions.

If you see someone hosting a copy of these materials somewhere else, please let them know that their actions are not allowed, and may lead to legal action. Moreover, they would be hurting the community, because we're not likely to release additional materials in this way if people ignore our copyright.

This is an early draft. If you get stuck running notebooks, please search the [fastai-v2 forum](#) for answers, and ask for help there if needed. Please don't use GitHub issues for problems running the notebooks.

If you make any pull requests to this repo, then you are assigning copyright of that work to Jeremy Howard and Sylvain Gugger. (Additionally, if you are making small edits to spelling or text, please specify the name of the file and very brief description of what you changed.) We do this to make it easier for us to review and merge changes. If you want to know which corrections have already been made, thank you.)

read this thing for free as stupid or notebooks but that is not as convenient

Playback speed Normal >

Subtitles/CC (9) English (auto-generated) >

Quality Auto 720p >

2:31 / 1:22:30 • Introduction

Motivation: Label provided!

The screenshot shows a browser window with multiple tabs open. The active tab is titled "fastai/fastbook: Draft of the fac" and displays the content of the fastai book draft. The page title is "The fastai book - draft". It contains text about the draft notebooks, their use in a course, and copyright information. A video player is overlaid on the page, showing a man speaking into a microphone. A subtitle menu is open, listing various languages: Off, Arabic, Bulgarian, Chinese (Simplified), English (selected), Japanese, Spanish, Tamil, and Turkish.

The fastai book - draft

These draft notebooks cover an introduction to deep learning, [fastai](#), and [PyTorch](#). fastai is a layered API for deep learning; for more information, see [the fastai paper](#). Everything in this repo is copyright Jeremy Howard and Sylvain Gugger, 2020 onwards.

These notebooks will be used for a [course we're teaching](#) in San Francisco from March 2020, and will be available as a MOOC from around July 2020. In addition, our plan is that these notebooks will form the basis of [this book](#), which you can pre-order. It will not have the same GPL restrictions that are on this draft.

The code in the notebooks and python `.py` files is covered by the GPL v3 license; see the LICENSE file for details.

The remainder (including all markdown cells in the notebooks and other prose) is not licensed for any redistribution or change of format or medium, other than making copies of the notebooks or forking this repo for your own private use. No commercial or broadcast use is allowed. We are making these materials freely available to help you learn deep learning, so please respect our copyright and these restrictions.

If you see someone hosting a copy of these materials somewhere else, please let them know that their actions are not allowed, and may lead to legal action. Moreover, they would be hurting the community, because we're not likely to release additional materials in this way if people ignore our copyright.

This is an early draft. If you get stuck running notebooks, please search the [fastai-v2 forum](#) for answers, and ask for help there if needed. Please don't use GitHub issues for problems running the notebooks.

If you make any pull requests to this repo, then you are assigning copyright of that work to Jeremy Howard and Sylvain Gugger. (Additionally, if you are making small edits to spelling or text, please specify the name of the file and very brief description of what you changed.)

free as Jupyter notebooks, but that is not as convenient as reading it on a Kindle or

Subtitles/CC Options

- Off
- Arabic
- Bulgarian
- Chinese (Simplified)
- English
- Japanese
- Spanish
- Tamil
- Turkish

2:31 / 1:22:30 • Introduction

Dataset

Dataset: Corrections from my fav creators

main ▾ youtube_captions_corrections / data / transcripts / en / labeled_transcripts /	
em Adding postprocessing script that produces single sequence with label... ...	
..	
3Blue1Brown.json	adding datasets
Alfredo_Canziani.json	adding datasets
Aurélien_Géron.json	Adding postprocessing script that produces single sequence with label...
DeepMind.json	adding datasets
Jeremy_Howard.json	adding datasets
Luis_Serrano.json	Adding postprocessing script that produces single sequence with label...
Pieter_Abbeel.json	adding datasets
TED.json	adding datasets
Veritasium.json	adding datasets
Weights_&_Biases.json	Adding postprocessing script that produces single sequence with label...
minutephysics.json	adding datasets
nature_video.json	adding datasets

https://github.com/2dot71mily/youtube_captions_corrections

Dataset: Diff tools!

The screenshot shows a file browser interface with the following path: `main > youtube_captions_corrections / data / transcripts / en / labeled_transcripts /`. The main content area displays a list of files and folders:

- `em` Adding postprocessing script that produces single sequence with label... `...`
- `..`
- `3Blue1Brown.json` adding datasets

`class difflib.Differ`

This is a class for comparing sequences of lines of text, and producing human-readable differences or deltas. Differ uses `SequenceMatcher` both to compare sequences of lines, and to compare sequences of characters within similar (near-matching) lines.

Each line of a `Differ` delta begins with a two-letter code:

Code	Meaning
' - '	line unique to sequence 1
' + '	line unique to sequence 2
' ' '	line common to both sequences

Dataset: Jeremy Howard example

Video metadata

```
In [4]: transcripts[['video_titles', 'playlist_ids', 'channel_ids']]
```

Out[4]:

	video_titles	playlist_ids	channel_ids
5L3Ao5KuCC4	Lesson 3 - Deep Learning for Coders (2020)	PLfYUBJiXbdRL3FMB3GoWHRI8ieU6FhfM	UCX7Y2qWriXpqocG97SFW2OQ
BvHmRx14HQ8	Lesson 2 - Deep Learning for Coders (2020)	PLfYUBJiXbdRL3FMB3GoWHRI8ieU6FhfM	UCX7Y2qWriXpqocG97SFW2OQ
CzdWqFTmn0Y	Intro to Machine Learning: Lesson 1	PLfYUBJiXbdSyktd8A_x0JNd6IxDcZE96	UCX7Y2qWriXpqocG97SFW2OQ
IPBSB1HLNLo	Lesson 1: Deep Learning 2018	PLfYUBJiXbdS2UQRzyrxmyVHoGW0gmLSM	UCX7Y2qWriXpqocG97SFW2OQ
JNxcznsrRb8	Lesson 2: Deep Learning 2018	PLfYUBJiXbdS2UQRzyrxmyVHoGW0gmLSM	UCX7Y2qWriXpqocG97SFW2OQ
VEG5xT5gAHc	Lesson 7 - Deep Learning for Coders (2020)	PLfYUBJiXbdRL3FMB3GoWHRI8ieU6FhfM	UCX7Y2qWriXpqocG97SFW2OQ
WjnwwGjZcM	Lesson 8 - Deep Learning for Coders (2020)	PLfYUBJiXbdRL3FMB3GoWHRI8ieU6FhfM	UCX7Y2qWriXpqocG97SFW2OQ
XfoYk_Z5Akl	Lesson 1: Deep Learning 2019 - Image classific...	PLfYUBJiXbdSIJb-Qd3pw0cqCbkGeS0xn	UCX7Y2qWriXpqocG97SFW2OQ
_QUEXsHfsA0	Lesson 1 - Deep Learning for Coders (2020)	PLfYUBJiXbdRL3FMB3GoWHRI8ieU6FhfM	UCX7Y2qWriXpqocG97SFW2OQ
cX30jxMNBUw	Lesson 6 - Deep Learning for Coders (2020)	PLfYUBJiXbdRL3FMB3GoWHRI8ieU6FhfM	UCX7Y2qWriXpqocG97SFW2OQ
ccMHJeQU4Qw	Lesson 2: Deep Learning 2019 - Data cleaning a...	PLfYUBJiXbdSIJb-Qd3pw0cqCbkGeS0xn	UCX7Y2qWriXpqocG97SFW2OQ
krIVOb23EH8	Lesson 5 - Deep Learning for Coders (2020)	PLfYUBJiXbdRL3FMB3GoWHRI8ieU6FhfM	UCX7Y2qWriXpqocG97SFW2OQ
p50s63nPq9I	Lesson 4 - Deep Learning for Coders (2020)	PLfYUBJiXbdRL3FMB3GoWHRI8ieU6FhfM	UCX7Y2qWriXpqocG97SFW2OQ

Dataset: Jeremy Howard example

Transcript text and metadata

```
In [5]: transcripts[['autogen', 'manual', 'autogen_text', 'manual_text', 'diffs']]
```

Out[5]:

	autogen	manual	autogen_text	manual_text	diffs
5L3Ao5KuCC4	[{"text": "oh hello and welcome to lesson three of practical deep learning for everyone"}, {"text": "So hello, and welcome to Lesson 3 of Practical Deep Learning for Everyone"}]	[{"text": "So hello, and welcome to Lesson 3 of Practical Deep Learning for Everyone"}]	oh hello and welcome to lesson three of practical... So hello, and welcome to Lesson 3 of Practical...	So hello, and welcome to Lesson 3 of Practical...	[{"-": ["oh", "hello"], "+": ["So", "hello"]}]
BvHmRx14HQ8	[{"text": "so uh hello everybody and welcome back to practical deep learning for everyone"}, {"text": "So, hello everybody, and welcome back to Practical Deep Learning for Everyone"}]	[{"text": "So, hello everybody, and welcome back to Practical Deep Learning for Everyone"}]	so uh hello everybody and welcome back to practical... So, hello everybody, and welcome back to Practic...	So, hello everybody, and welcome back to Practic...	[{"-": ["so", "uh"], "+": ["So,"], "diffs": [{"-": "uh", "+": "So,"}], "-": "So,"}]
CzdWqFTmn0Y	[{"text": "Oh", "start": 0.0, "duration": 9.05}, {"text": "Okay, so let me introduce everybody to the course"}, {"text": "Okay, so let me introduce everybody to the course"}]	[{"text": "Okay, so let me introduce everybody to the course"}]	Oh good okay so let me introduce everybody to ... Okay, so let me introduce everybody to everybo...	Okay, so let me introduce everybody to everybo...	[{"-": ["Oh", "good"], "+": ["Okay,"], "-": "Okay,"}]
IPBSB1HLNLo	[{"text": "hi everybody welcome to practical deep learning for everyone"}, {"text": "Hi everybody welcome to practical deep learning for everyone"}]	[{"text": "Hi everybody welcome to practical deep learning for everyone"}]	hi everybody welcome to practical deep learnin... Hi everybody welcome to practical deep learnin...	Hi everybody welcome to practical deep learnin...	[{"-": ["hi"], "+": ["Hi"]}]
JNxcznsrRb8	[{"text": "okay so welcome back to deep learning for everyone"}, {"text": "Okay so welcome back to deep learning for everyone"}]	[{"text": "Okay so welcome back to deep learning for everyone"}]	okay so welcome back to deep learning lesson 2... Okay so welcome back to deep learning lesson 2...	Okay so welcome back to deep learning lesson 2...	[{"-": ["okay"], "+": ["Okay"]}]
VEG5xT5gAHc	[{"text": "hi everybody and welcome to lesson 7"}, {"text": "Hi everybody and welcome to lesson 7"}]	[{"text": "Hi everybody and welcome to lesson 7"}]	hi everybody and welcome to lesson 7 we're goin... Hi everybody and welcome to lesson 7! We're\ngoin...	Hi everybody and welcome to lesson 7! We're\ngoin...	[{"-": ["hi"], "+": ["Hi"]}]
WjnwwWeGjZcM	[{"text": "hi everybody and welcome to lesson 8"}, {"text": "Hi everybody and welcome to lesson 8"}]	[{"text": "Hi everybody and welcome to lesson 8"}]	hi everybody and welcome to lesson eight the last one Hi everybody and welcome to lesson eight,\nthe last one	Hi everybody and welcome to lesson eight,\nthe last one	[{"-": ["hi"], "+": ["Hi"]}]
XfoYk_Z5Akl	[{"text": "okay so welcome practical deep learning for everyone"}, {"text": "Okay so Welcome Practical Deep Learning for Everyone"}]	[{"text": "Okay so Welcome Practical Deep Learning for Everyone"}]	okay so welcome practical deep learning for co... Okay so Welcome Practical deep learning for co...	Okay so Welcome Practical deep learning for co...	[{"-": ["okay"], "+": ["Okay"]}]
_QUEXsHfsA0	[{"text": "so hello everybody and welcome to deep learning for everyone"}, {"text": "So hello everybody and welcome to Deep Learning for Everyone"}]	[{"text": "So hello everybody and welcome to Deep Learning for Everyone"}]	so hello everybody and welcome to deep learnin... So hello everybody and welcome to Deep Learnin...	So hello everybody and welcome to Deep Learnin...	[{"-": ["so"], "+": ["So"]}]
cX30jxMNBUw	[{"text": "hi everybody and welcome to lesson 6"}, {"text": "Hi everybody and welcome to Lesson 6"}]	[{"text": "Hi everybody and welcome to Lesson 6"}]	hi everybody and welcome to lesson six where we're going to be talking about Hi everybody and welcome to Lesson 6, where\newnwe're going to be talking about	Hi everybody and welcome to Lesson 6, where\newnwe're going to be talking about	[{"-": ["hi"], "+": ["Hi"]}]
ccMHJeQU4Qw	[{"text": "welcome to lesson two where we're going to be talking about"}, {"text": "Welcome to Lesson 2 where we're going to be talking about"}]	[{"text": "Welcome to Lesson 2 where we're going to be talking about"}]	welcome to lesson two where we're going to be ... Welcome to Lesson 2 where we're going to be ta...	Welcome to Lesson 2 where we're going to be ta...	[{"-": ["welcome"], "+": ["Welcome"]}]
krlVOb23EH8	[{"text": "welcome to lesson five and we'll be talking about"}, {"text": "Welcome to lesson five and we'll be talking about"}]	[{"text": "Welcome to lesson five and we'll be talking about"}]	welcome to lesson five and we'll be talking ab... Welcome to lesson five and we'll be talking\nabout	Welcome to lesson five and we'll be talking\naabout	[{"-": ["welcome"], "+": ["Welcome"]}]
p50s63nPq9I	[{"text": "welcome back and here is lesson four"}, {"text": "Welcome back and here is lesson 4"}]	[{"text": "Welcome back and here is lesson 4"}]	welcome back and here is lesson or which is wh... Welcome back and here is lesson 4 which is\nwhich is	Welcome back and here is lesson 4 which is\nwhich is	[{"-": ["welcome"], "+": ["Welcome"]}]

Dataset: Jeremy Howard example

Label data				
In [42]:	transcripts[['is_autogen_unique', 'autogen_seq', 'manual_seq', 'manual_addl_rep']]			
Out[42]:	is_autogen_unique	autogen_seq	manual_seq	manual_addl_rep
5L3Ao5KuCC4	[2, 2, 0, 0, 0, 2, 2, 0, 2, 2, 2, 0, 2, 2, 0, ...]	[oh, hello, , , , lesson, three, , practical, ...]	[So hello,, So hello,, , , , Lesson 3, Lesson ...]	[1, 1, 0, 0, 1, 1, 0, 2, 2, 2, 0, 1, 1, 0, ...]
BvHmRx14HQ8	[2, 2, 0, 2, 0, 0, 0, 0, 2, 2, 2, 0, 2, 2, 0, ...]	[so, uh, , everybody, , , , practical, deep,...]	[So,, So,, , everybody,, , , , Practical Dee...]	[1, 1, 0, 0, 0, 0, 0, 2, 2, 2, 0, 1, 1, 0, ...]
CzdWqFTmn0Y	[2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 2, ...]	[Oh, good, okay, , , , , else, , , so,...]	[Okay,, Okay,, Okay,, , , , , else,, , , ...]	[2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]
IPBSB1HLNL0	[2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]	[hi, , , , , , , , , course, , p...]	[Hi, , , , , , , , , course,, , ...]	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]
JNxcznsrRb8	[2, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, ...]	[okay, , , , , last, , , got, , , , ...]	[Okay, , , , , Last, , , Got, , , , ...]	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]
VEG5xT5gAHc	[2, 0, 0, 0, 0, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0, ...]	[hi, , , , , 7, we're, , , , , , , ...]	[Hi, , , , , 7! We're, 7! We're, , , , , ...]	[0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, ...]
WjnWWeGjZcM	[2, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, ...]	[hi, , , , , eight, , , , , course, ...]	[Hi, , , , , eight,, , , , , course...]	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]
XfoYk_Z5Akl	[2, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]	[okay, , welcome, practical, , , , , , ...]	[Okay, , Welcome Practical, Welcome Practical,...]	[0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]
_QUEXsHfsAO	[2, 0, 0, 0, 0, 0, 2, 2, 0, 2, 2, 2, 0, 0, ...]	[so, , , , , deep, learning, , coders, lesso...]	[So, , , , , Deep Learning, Deep Learning, , ...]	[0, 0, 0, 0, 0, 0, 1, 1, 0, 3, 3, 3, 3, 0, 0, ...]
cX30jxMNBUw	[2, 0, 0, 0, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0, ...]	[hi, , , , , lesson, six, , , , , , , ...]	[Hi, , , , , Lesson 6,, Lesson 6,, , , , , ...]	[0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, ...]
ccMHJeQU4Qw	[2, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]	[welcome, , lesson, two, , , , , , , ...]	[Welcome, , Lesson 2, Lesson 2, , , , , , ...]	[0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]
krlVOb23EH8	[2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]	[welcome, , , , , , , , , , , , , ...]	[Welcome, , , , , , , , , , , , , ...]	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]
p50s63nPq9I	[2, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, ...]	[welcome, , , , , or, , , , , , , , ...]	[Welcome, , , , , 4, , , , , , , , ...]	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]

From `label_diff_targets()`

```
BOTH_AGREE = 0
BOTH_DIFFER = 2
AUTOGEN_INSERT = 1
MANUAL_INSERT = -1
```

Dataset: Jeremy Howard example

```
In [41]: transcripts.loc['_QUEXsHfsA0']['diffs'][40:70]

Out[41]: ['it',
          'to',
          'you',
          'live',
          'from',
          'day',
          'number',
          'one',
          'of',
          'a',
          'complete',
          {'+': ['shutdown.', 'Oh,'], '-': ['shutdown', 'or']},
          'not',
          {'+': ['a', 'complete', 'shutdown', 'but', 'nearly', 'a'], '-': []},
          'complete',
          'shutdown',
          {'+': ['of'], '-': ['but', 'nearly', 'complete', 'shutdown', 'in']},
          'San',
          {'+': ['Francisco.', "We're"], '-': ['Francisco', "we're"]},
          'going',
          'to',
          'be',
          'recording',
          'it',
          'over',
          'the',
          {'+': [], '-': ['defect']},
          'next',
          'two',
          'months']
```

Dataset: Add your fav channels

The screenshot shows a Google Colab notebook titled "Adding_a_youtube_channel_to_yt_caption_corrections_dataset". The notebook interface includes a toolbar with "File", "Edit", "View", "Insert", "Runtime", "Tools", and "Help" buttons, and a status bar indicating "All changes saved". The main area contains two code cells and a configuration file.

Code Cells:

- [5] %%capture
!git clone https://github.com/2dot71mily/youtube_captions_corrections.git
!pip install youtube_transcript_api
!pip install google-api-python-client
- [6] !python youtube_captions_corrections/src/prepare_data.py
requesting data for /Users/emcmilin/youtube_captions_corrections/temp/data/transcripts/en/raw_transcripts/Jeremy_Howard.json
requesting data for /Users/emcmilin/youtube_captions_corrections/temp/data/channels/Jeremy_Howard.json
If below requested channel is correct:
<https://www.youtube.com/channel/UCX7Y2qWriXpqocG97SFW200>
please hit [Enter]
Else [ctrl+c] script and check for other `channel_id` options in:
/Users/emcmilin/youtube_captions_corrections/temp/data/channels/Jeremy_Howard.json

Configuration File:

```
config.py X
1 from pathlib import PurePath
2
3 ### Prefs ###
4 DEVELOPER_KEY = "REDACTED"
5 LANGUAGE = "en"
6 CHANNEL_NAME = "Jeremy Howard"
7 GET_CHANNEL_IDS_ONLY = False
8 GET_PLAYLIST_IDS_ONLY = False
9 GET_VIDEO_IDS_ONLY = False
10
11 SAVE_INTERVAL = 100
12 TRANSCRIPT_SAVE_INTERVAL = SAVE_INTERVAL
13
14 PRINT_TRANSCRIPT_API_ERR = False
15 USE_ONLY_POSTPROC_LABELS = True
16 USE_VIDEO_ID_AS_IDX = False
17
18 SPLIT_FILE_N_LINES = 3000
19
20 ### Labels ###
```

Dataset: Revisiting stupid or...?

The fastai book - draft

These draft notebooks cover an introduction to deep learning, [fastai](#), and [PyTorch](#). fastai is a layered API for deep learning; for more information, see [the fastai paper](#). Everything in this repo is copyright Jeremy Howard and Sylvain Gugger, 2020 onwards.

These notebooks will be used for [a course we're teaching](#) in San Francisco from March 2020, and will be available as a MOOC from around July 2020. In addition, our plan is that these notebooks will form the basis of [this book](#), which you can pre-order. It will not have the same GPL restrictions that are on this draft.

The code in the notebooks and python `.py` files is covered by the GPL v3 license; see the LICENSE file for details.

The remainder (including all markdown cells in the notebooks and other prose) is not licensed for any redistribution or change of format or medium, other than making copies of the notebooks or forking this repo for your own private use. No commercial or broadcast use is allowed. We are making these materials freely available to help you learn deep learning, so please respect our copyright and these restrictions.

If you see someone hosting a copy of these materials somewhere else, please let them know that their actions are not allowed, and may lead to legal action. Moreover, they would be hurting the community, because we're not likely to release additional materials in this way if people ignore our copyright.

This is an early draft. If you get stuck running notebooks, please search the [fastai-v2 forum](#) for answers, and ask for help there if needed. Please don't use GitHub issues for problems running the notebooks.

If you make any pull requests to this repo, then you are assigning copyright of that work to Jeremy Howard and Sylvain Gugger. (Additionally, if you are making small edits to spelling or text, please specify the name of the file and very brief description, with the original text in the commit message, so we can know which corrections have already been made. Thank you.)

read this thing for free as stupid or
notebooks but that is not as convenient

2:31 / 1:22:30 • Introduction

Dataset: Revisiting stupid

The screenshot shows a web browser window with multiple tabs open. The active tab displays the 'The fastai book - draft' page from <https://github.com/fastai/fastbook>. The page content discusses the introduction to deep learning, the fastai library, and the course it's used for. It emphasizes that the code is covered by GPL v3 and the remainder is free for Jupyter notebooks. A note about pull requests mentions Gugger and a warning about hosting the material elsewhere. The bottom of the page contains a large block of text with several words highlighted in yellow.

The highlighted text reads:

read this thing for free as stupid or notebooks but that is not as convenient

Below the video player controls, there is a progress bar at 2:31 / 1:22:30 and a timestamp of 2:31 / 1:22:30 • Introduction.

autogen_seq	common_to_both_seq	manual_seq	manual_addl_rep
0	you		0
1	can		0
2	read		0
3	this		0
4	thing		0
5	for		0
6	free		0
7	as		0
8	stupid	Jupyter notebooks,	2
9	or	Jupyter notebooks,	2
10	notebooks	Jupyter notebooks,	2
11	but		0
12	that		0
13	is		0
14	not		0
15	as		0
16	convenient		0

Dataset: Subset added to HuggingFace

Hugging Face Models Datasets Pricing Resources We're hiring! Log In Sign Up

Dataset: youtube_caption_corrections

Tasks: other-other-token-classification-of-text-errors slot-filling Task Categories: other sequence-modeling Languages: en Multilinguality: monolingual Size Categories: 10K<n<100K Licenses: mit

Language Creators: machine-generated Annotations Creators: expert-generated machine-generated Source Datasets: original

Dataset Structure

- Data Instances
- Data Fields
- Data Splits

Dataset Creation

- Curation Rationale
- Source Data
- Annotations
- Personal and Sensitive Information

Considerations for Using the Data

- Social Impact of Dataset
- Discussion of Biases
- Other Known Limitations

Additional Information

- Dataset Curators
- Licensing Information
- Citation Information
- Contributions

Dataset Card for YouTube Caption Corrections

Dataset Summary

This dataset is built from pairs of YouTube captions where both an auto-generated and a manually-corrected caption are available for a single specified language. It currently only in English, but scripts at repo support other languages. The motivation for creating it was from viewing errors in auto-generated captions at a recent virtual conference, with the hope that there could be some way to help correct those errors.

The dataset in the repo at https://github.com/2dot71mily/youtube_captions_corrections records in a non-destructive manner all the differences between an auto-generated and a manually-corrected caption for thousands of videos. The dataset here focuses on the subset of those differences which are mutual and have the same size in token length difference, which means it excludes token insertion or deletion differences between the two captions. Therefore dataset here remains a non-destructive representation of the original auto-generated captions, but excludes some of the differences that are found in the manually-corrected captions.

Update on GitHub Use in dataset library

Explore dataset Edit Model Tags

Homepage: github.com Repository: github.com Paper: N/A Leaderboard: N/A

Point of Contact: Emily McMilin

Models trained or fine-tuned on youtube_caption_corrections

None yet

Dataset: Labels added to HuggingFace

Hugging Face In Sign Up

Dataset: youtube_caption_corrections

Tasks: other-other-token-classification-of-text-errors slot-filling Task Categories: other

Language Creators: machine-generated Annotations Creators: expert-generated machine-generated

Dataset Structure

- Data Instances
- Data Fields
- Data Splits

Dataset Creation

- Curation Rationale
- Source Data
- Annotations
- Personal and Sensitive Information

Considerations for Using the Data

- Social Impact of Dataset
- Discussion of Biases
- Other Known Limitations

Additional Information

- Dataset Curators
- Licensing Information
- Citation Information
- Contributions

```
if auto_token.lower() == man_token.lower():
    new_labels[idx] = config.CASE_DIFF

elif auto_token.strip(punctuation) == man_token.strip(punctuation):
    new_labels[idx] = config.PUNCTUATION_DIFF

elif auto_token.lower().strip(punctuation) == man_token.lower().strip(
    punctuation):
    new_labels[idx] = config.CASE_AND_PUNCTUATION_DIFF

elif stemmer.stem(auto_token.lower()) == stemmer.stem(man_token.lower()):
    new_labels[idx] = config.STEM_BASED_DIFF

# E.g. `2` <-> `two` is a common diff
elif re.match("\d+", man_token) or re.match("\d+", auto_token):
    new_labels[idx] = config.DIGIT_DIFF

elif "".join(tokenizer.tokenize(auto_token)).lower().strip(
    punctuation) == "".join(tokenizer.tokenize(man_token)).lower().strip(punctuation):
    new_labels[idx] = configINTRAWORD_PUNC_DIFF

else:
    new_labels[idx] = config.UNKNOWN_TYPE_DIFF
```

Model: Token-Classification

Model: Getting dataset from HuggingFace

```
[10] from datasets import load_dataset

[11] if TESTING:
    yt_dataset = load_dataset("youtube_caption_corrections", split='train[:5%]')
else:
    yt_dataset = load_dataset("youtube_caption_corrections")['train']

Downloading: ██████████ 4.16k/? [00:06<00:00, 647B/s]

Downloading: ██████████ 2.86k/? [00:00<00:00, 55.2kB/s]
Using custom data configuration default

Downloading and preparing dataset youtube_caption_corrections/default (download:
Downloading: ██████████ 59.9M/? [00:11<00:00, 5.16MB/s]

Downloading: ██████████ 54.3M/? [00:05<00:00, 9.82MB/s]

Downloading: ██████████ 51.5M/? [00:08<00:00, 6.36MB/s]

Downloading: ██████████ 56.9M/? [00:01<00:00, 40.4MB/s]

Dataset youtube_caption_corrections downloaded and prepared to /root/.cache/hugg

[12] yt_dataset

Dataset({
    features: ['video_ids', 'default_seq', 'correction_seq', 'diff_type'],
    num_rows: 10769
})
```

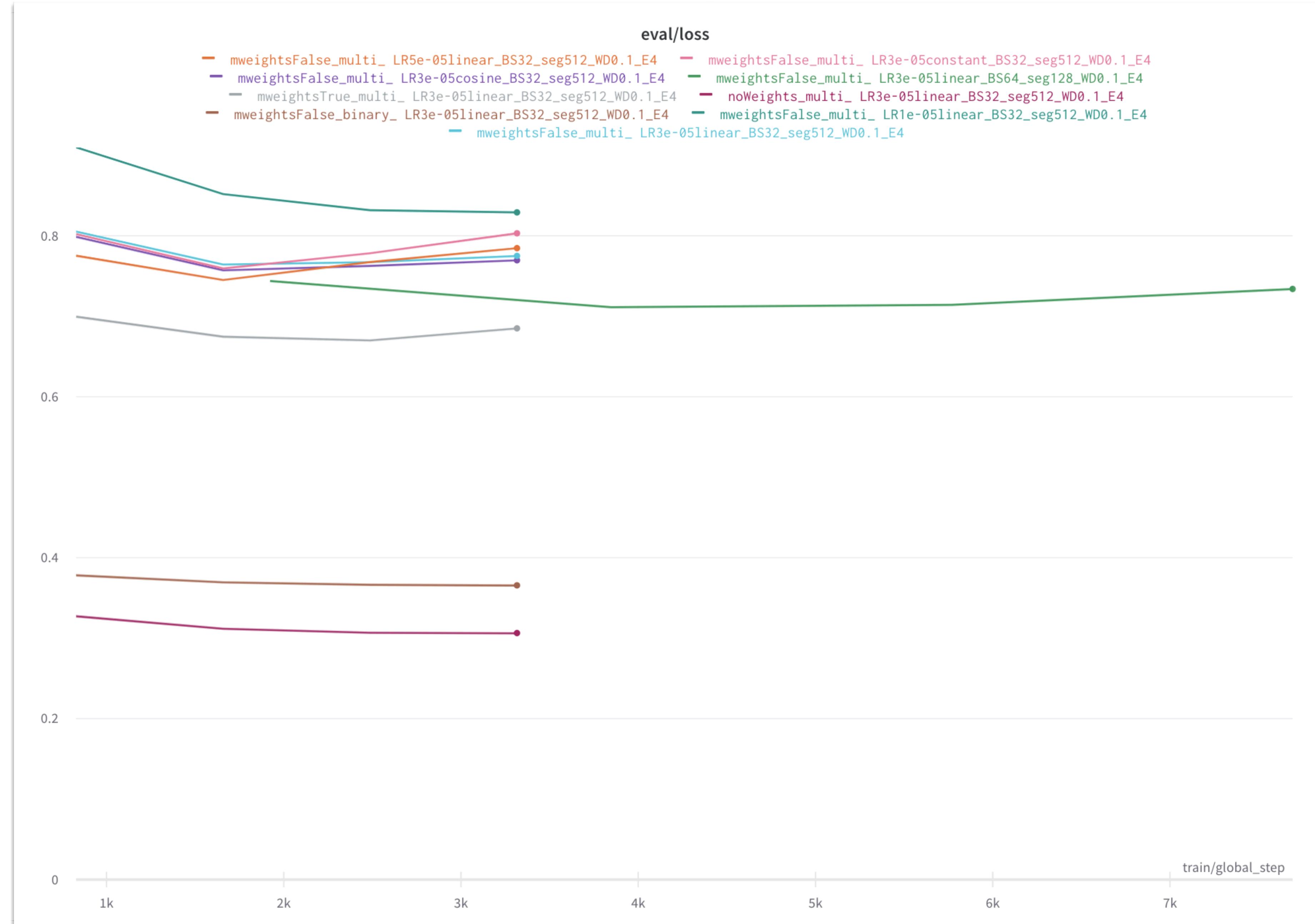
Model: Token-classification results



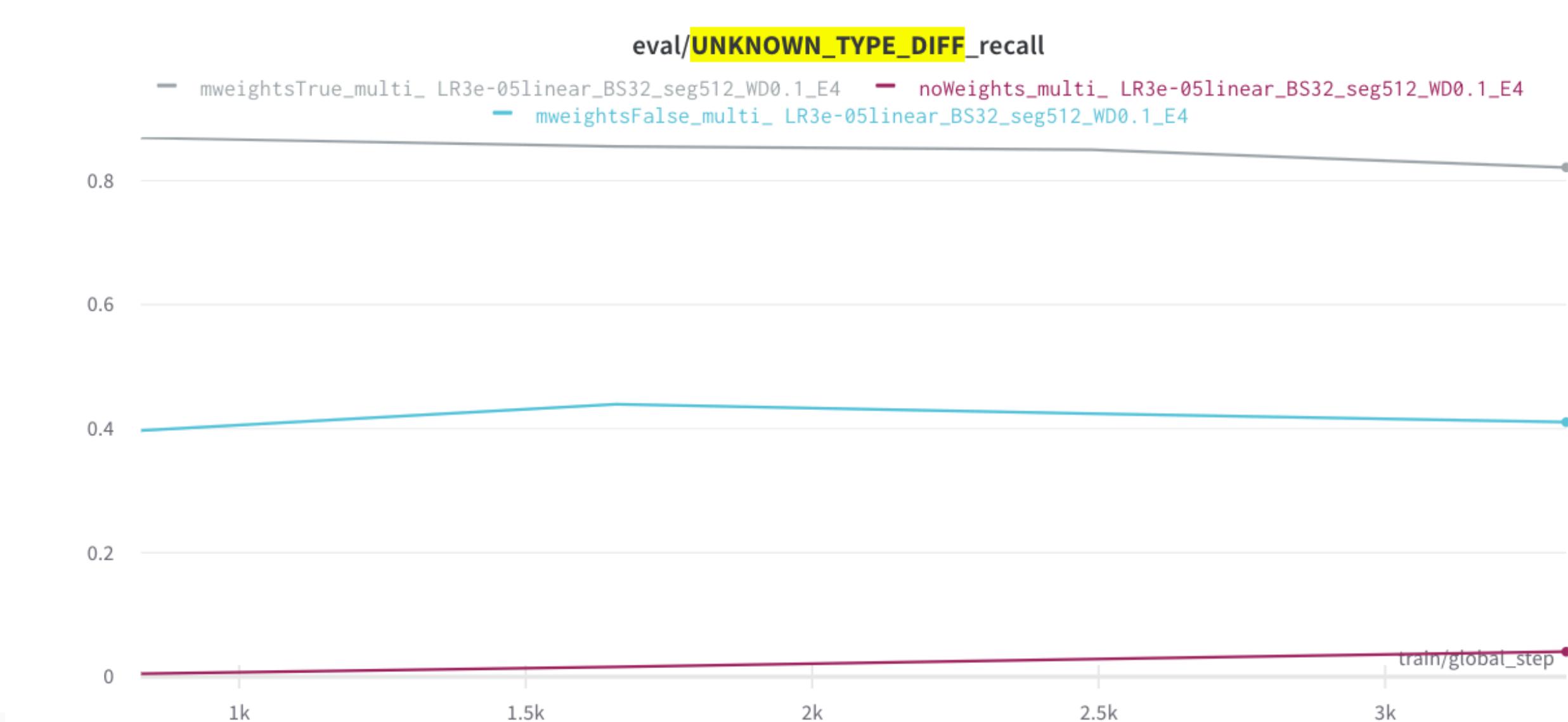
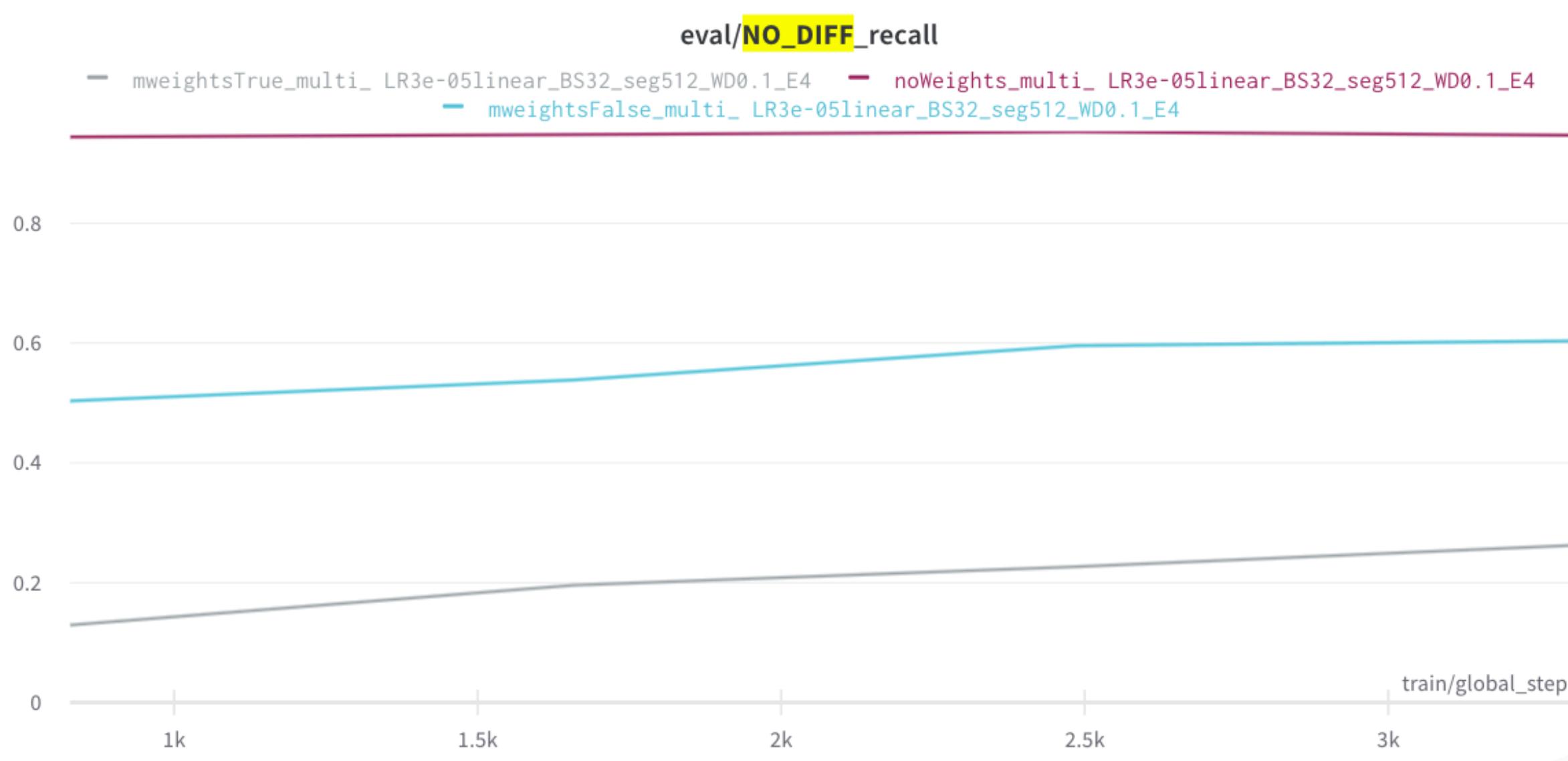
Model: Token-classification results



Model: Token-classification results



Model: Token-classification results







Next steps

Dataset

- Increase dataset size
- Include insertion/deletion errors in HuggingFace dataset

Token classification model:

- Train different models for different types of errors?
- Improve loss function to weigh important error types more heavily

Masked Language Model:

- Train model with MLM objective:
 - Mask out words labeled incorrect
- Attempt an end to end model:
 - Mask out words predicted as incorrect