Semantic Search on PyTorch discussions

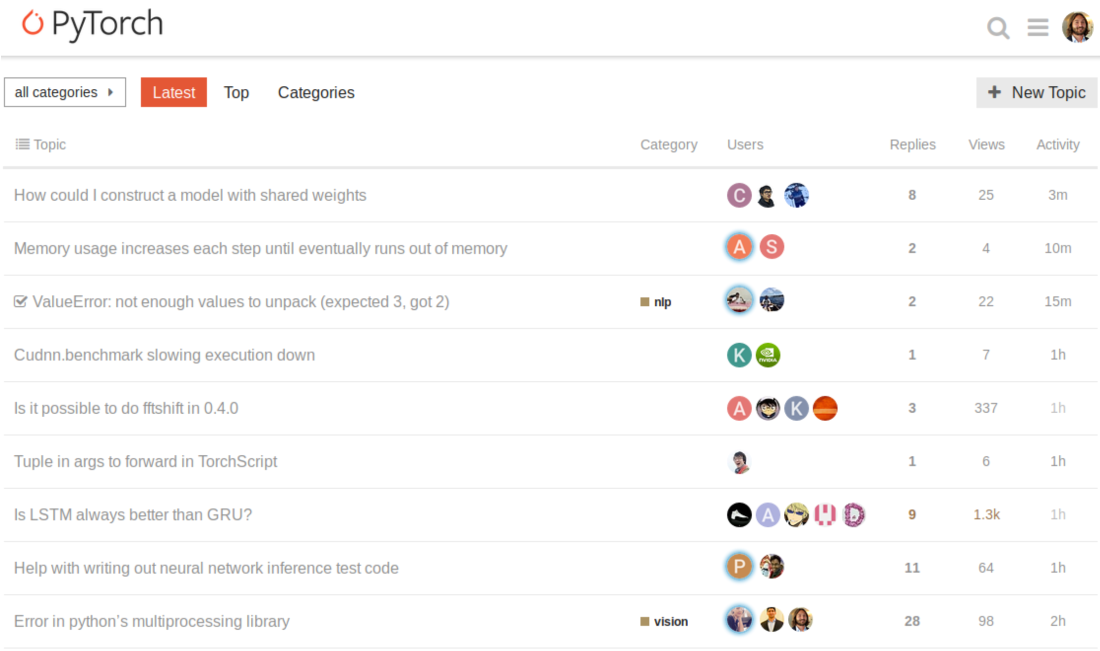
PyTorch in Munich at Microsoft

Munich Applied Deep Learning Meetup
Dec 11th, 2018

Piotr Bialecki, @ptrblck_de, github.com/ptrblck

PyTorch Discussion Board

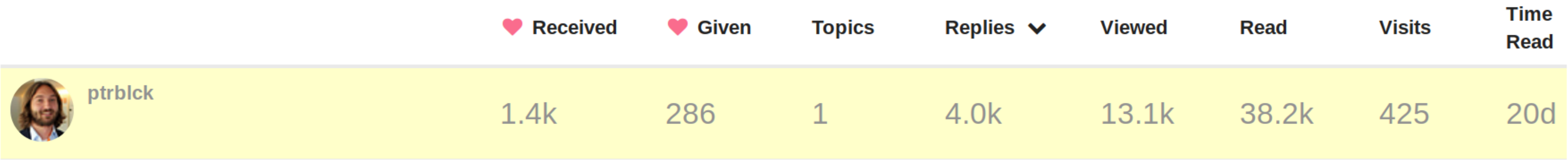
- https://discuss.pytorch.org/
- Discussions on deep learning with PyTorch
  - Help with debugging
  - Performance issues
  - Model/training support
  - Feedback
  - ...
- Welcoming community!



Topic	Category	Users	Replies	Views	Activity
How could I construct a model with shared weights		C	8	25	3m
Memory usage increases each step until eventually runs out of memory		A S	2	4	10m
☑ ValueError: not enough values to unpack (expected 3, got 2)	nlp		2	22	15m
Cudnn.benchmark slowing execution down		K	1	7	1h
Is it possible to do fftshift in 0.4.0		A K	3	337	1h
Tuple in args to forward in TorchScript			1	6	1h
Is LSTM always better than GRU?		A D	9	1.3k	1h
Help with writing out neural network inference test code		P	11	64	1h
Error in python's multiprocessing library	vision		28	98	2h

PyTorch Discussion Board

… and I like to hangout in the board!

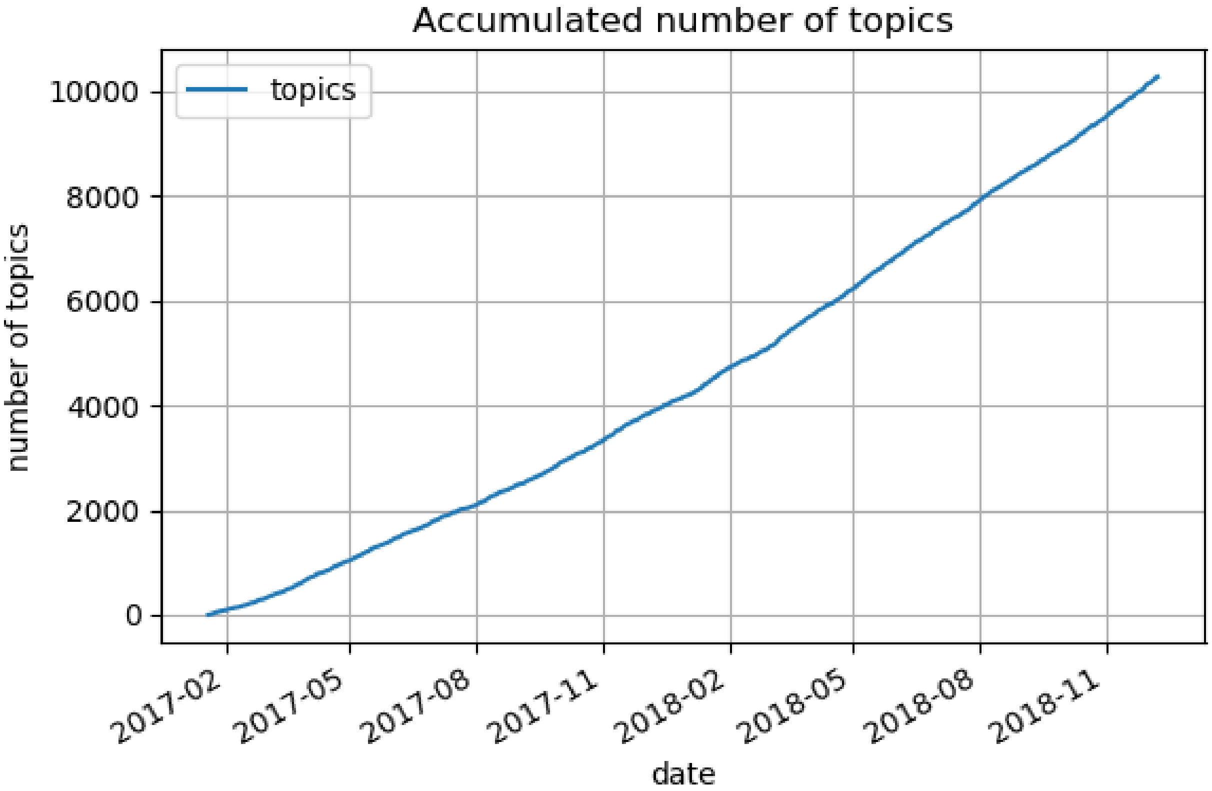


	❤ Received	❤ Given	Topics	Replies	Viewed	Read	Visits	Time Read
ptrblck	1.4k	286	1	4.0k	13.1k	38.2k	425	20d

So you might have seen me there.

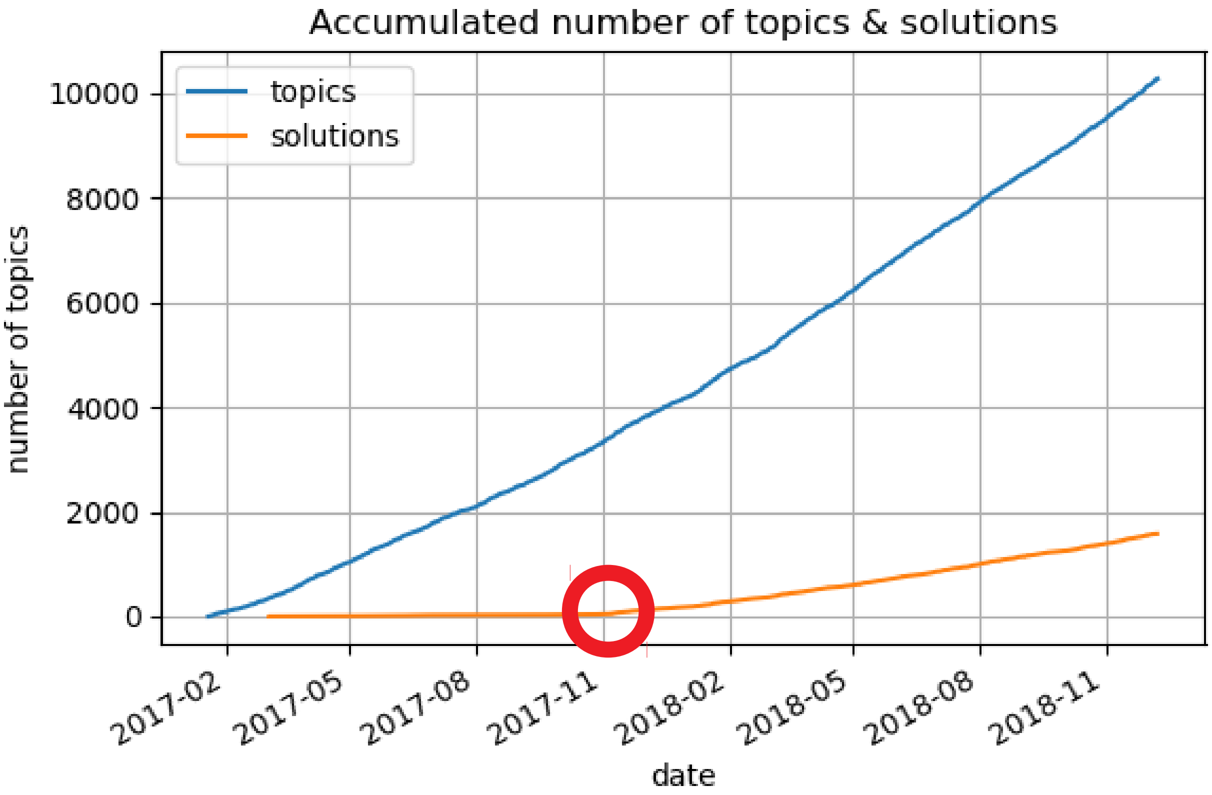
PyTorch Discussion Board

- Some stats
  - ~13,000 topics
  - ~52,000 posts
  - ~1,700 marked solutions



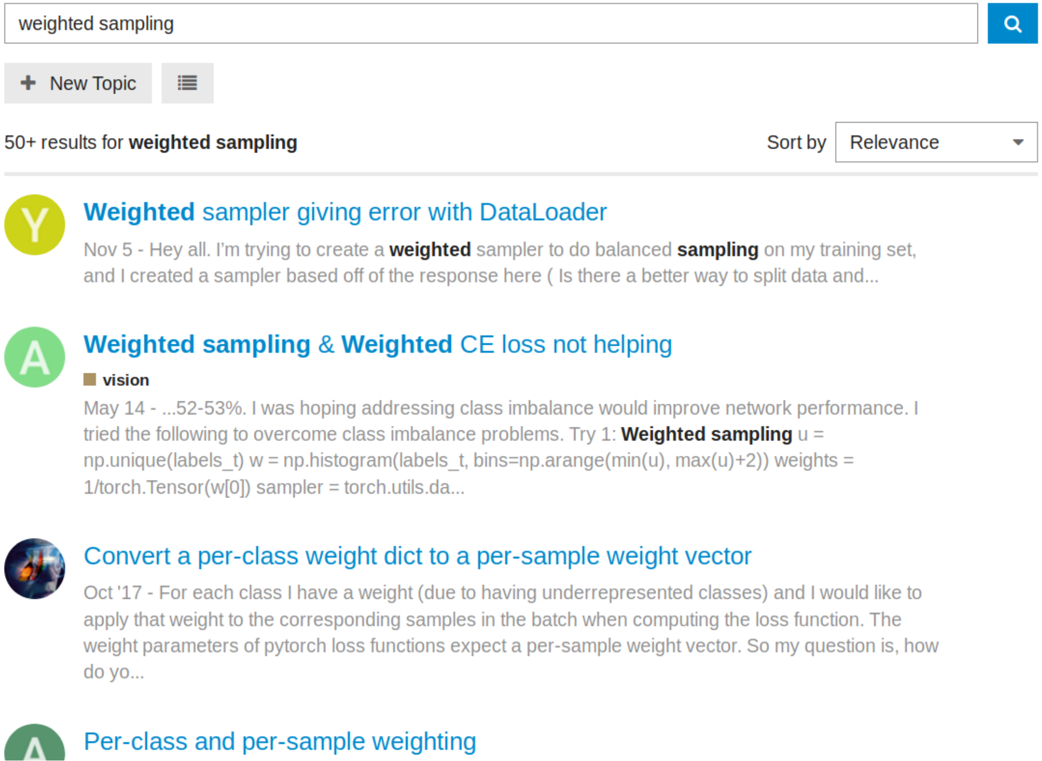
PyTorch Discussion Board

- Some stats
  - ~13,000 topics
  - ~52,000 posts
  - ~1,700 marked solutions



PyTorch Discussion Board

- Search works with **keywords** (lexical search)
- Fine in a lot of cases
- **Semantic search:** search with **meaning**



weighted sampling

50+ results for **weighted sampling**

Weighted sampler giving error with DataLoader

Nov 5 - Hey all. I'm trying to create a **weighted** sampler to do balanced **sampling** on my training set, and I created a sampler based off of the response here (Is there a better way to split data and...

Weighted sampling & **Weighted** CE loss not helping

vision

May 14 - ...52-53%. I was hoping addressing class imbalance would improve network performance. I tried the following to overcome class imbalance problems. Try 1: **Weighted sampling** u = np.unique(labels_t) w = np.histogram(labels_t, bins=np.arange(min(u), max(u)+2)) weights = 1/torch.Tensor(w[0]) sampler = torch.utils.da...

Convert a per-class weight dict to a per-sample weight vector

Oct '17 - For each class I have a weight (due to having underrepresented classes) and I would like to apply that weight to the corresponding samples in the batch when computing the loss function. The weight parameters of pytorch loss functions expect a per-sample weight vector. So my question is, how do yo...

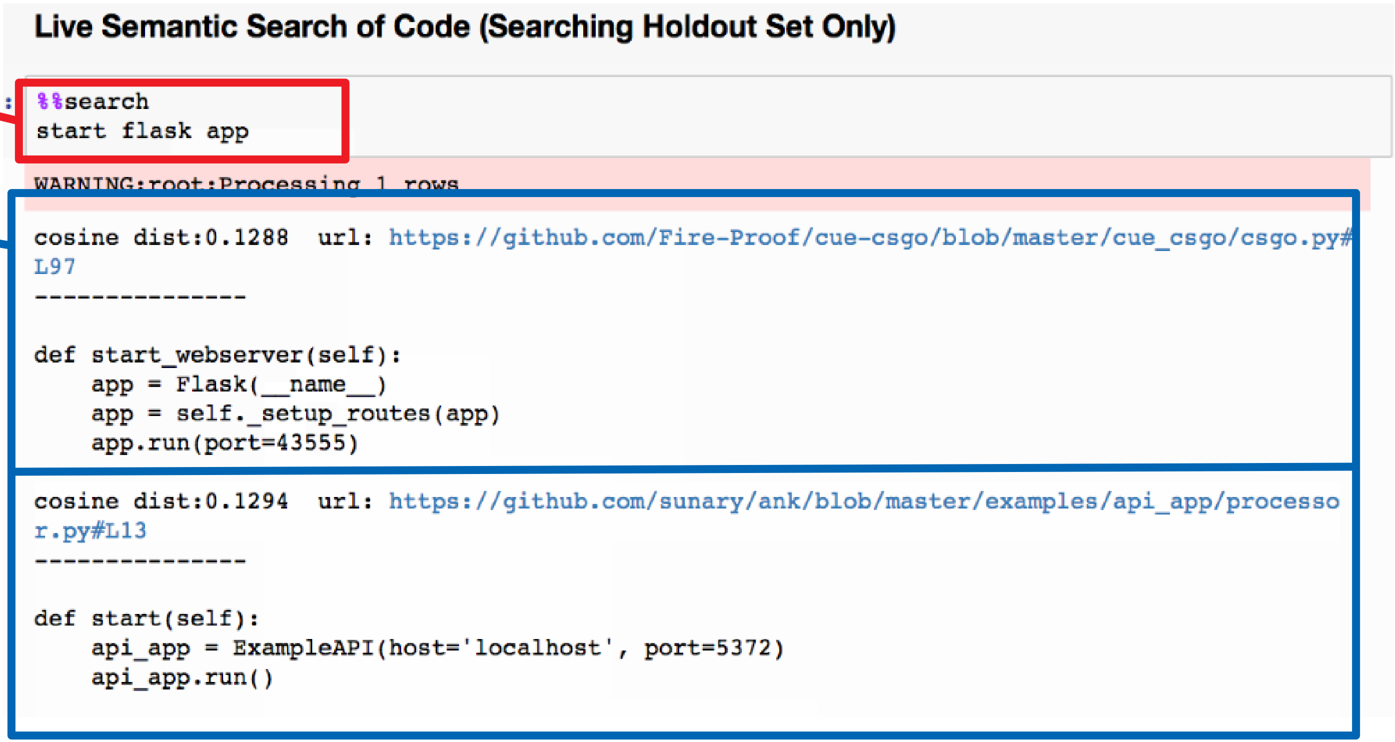
Per-class and per-sample weighting

Semantic search @github

- **Hamel Husain & Ho-Hsiang Wu** created semantic search demo for Github code search using deep learning
- Nice blog post: https://towardsdatascience.com/semantic-code-search-3cd6d244a39c
- Used **function – docstring** pairs
- This work is highly inspired by Husain and Wu (thanks a lot for the great blog post and explanations)

Semantic search @github

- Search
- Result



Live Semantic Search of Code (Searching Holdout Set Only)

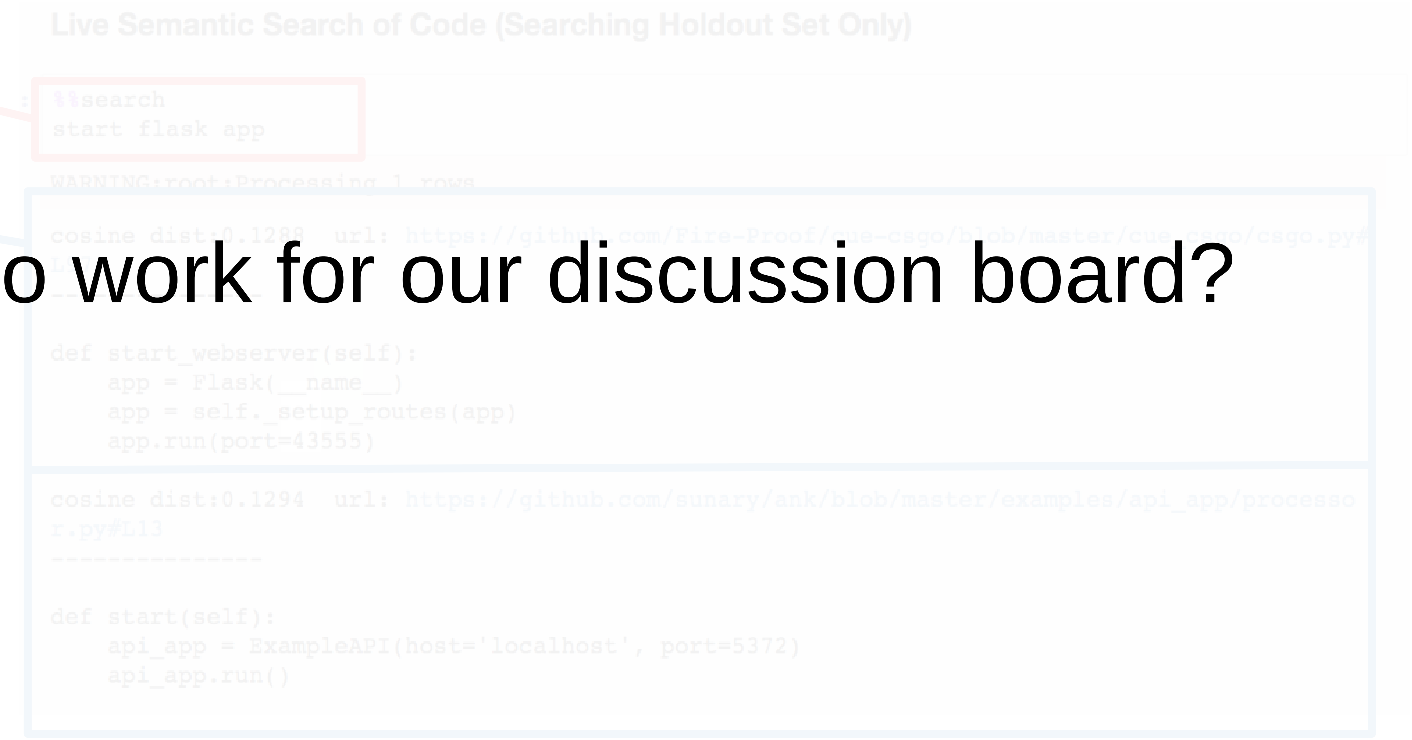
```
%%search
start flask app
```

```
WARNING:root:Processing 1 rows

cosine dist:0.1288  url: https://github.com/Fire-Proof/cue-csgo/blob/master/cue_csgo/csgo.py#L97
---------------

def start_webserver(self):
    app = Flask(__name__)
    app = self._setup_routes(app)
    app.run(port=43555)

cosine dist:0.1294  url: https://github.com/sunary/ank/blob/master/examples/api_app/processor.py#L13
---------------

def start(self):
    api_app = ExampleAPI(host='localhost', port=5372)
    api_app.run()
```

Semantic search @github

- Search
- Result

Would this also work for our discussion board?



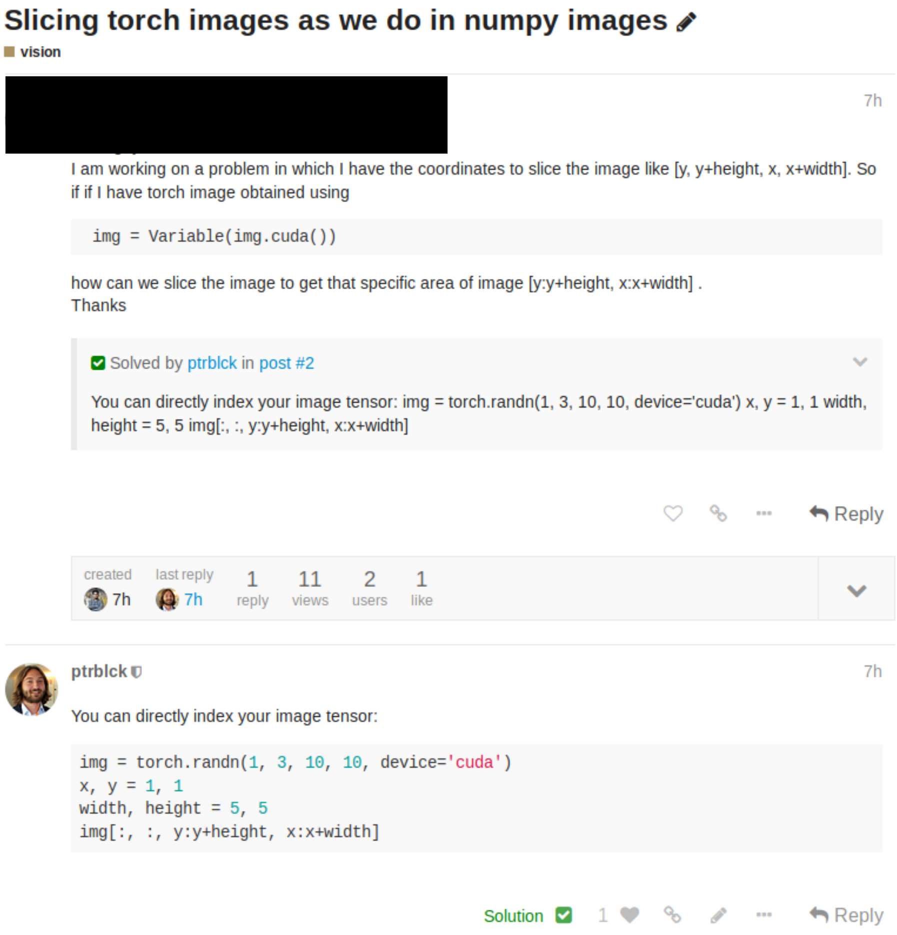
Live Semantic Search of Code (Searching Holdout Set Only)

```
%%search
start flask app
```

```
WARNING:root:Processing 1 rows

cosine dist:0.1288  url: https://github.com/Fire-Proof/cue-csgo/blob/master/cue_csgo/csgo.py#
L[illegible]
---------------

def start_webserver(self):
    app = Flask(__name__)
    app = self._setup_routes(app)
    app.run(port=43555)

cosine dist:0.1294  url: https://github.com/sunary/ank/blob/master/examples/api_app/processo
r.py#L13
---------------

def start(self):
    api_app = ExampleAPI(host='localhost', port=5372)
    api_app.run()
```

PyTorch Discussion Board

- Threads/topics:
  - Title (name)
  - Question (start post)
  - Stats
  - Solution?
  - Posts



Slicing torch images as we do in numpy images

[redacted]

I am working on a problem in which I have the coordinates to slice the image like [y, y+height, x, x+width]. So if if I have torch image obtained using

```
img = Variable(img.cuda())
```

how can we slice the image to get that specific area of image [y:y+height, x:x+width] .
Thanks

✅ Solved by ptrblck in post #2

You can directly index your image tensor: img = torch.randn(1, 3, 10, 10, device='cuda') x, y = 1, 1 width, height = 5, 5 img[:, :, y:y+height, x:x+width]

created	last reply	1	11	2	1
7h	7h	reply	views	users	like

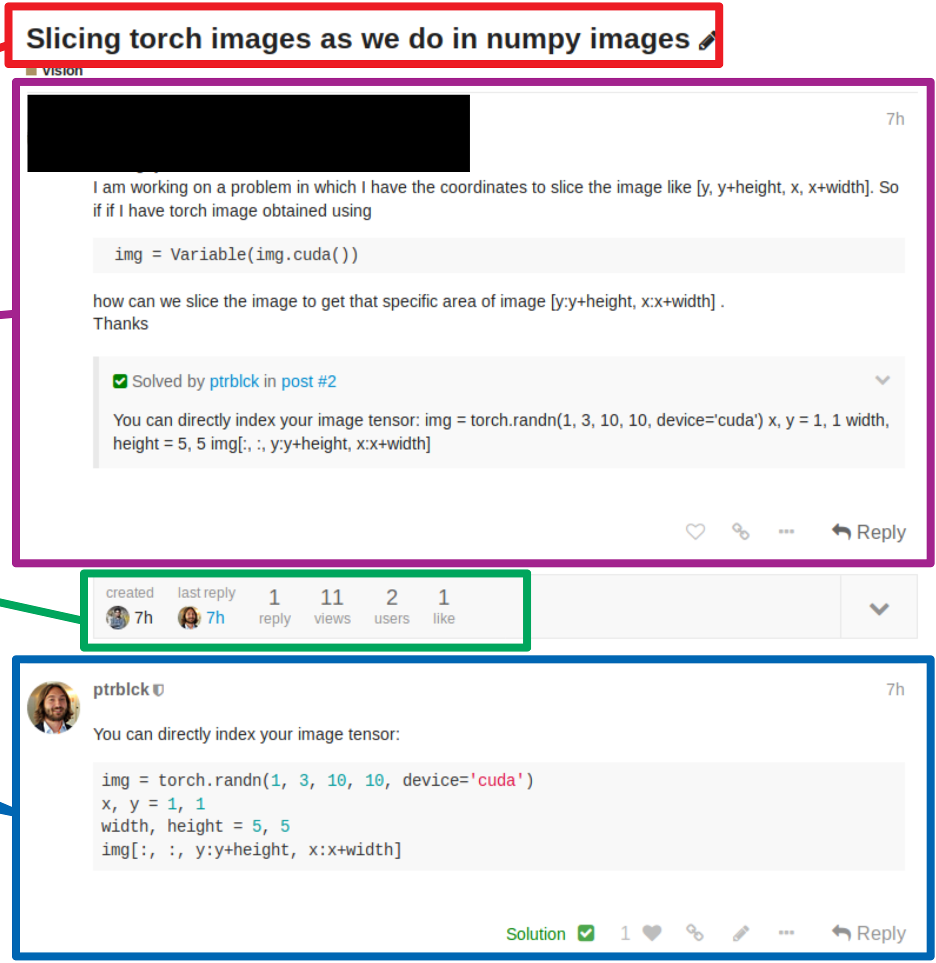
ptrblck

You can directly index your image tensor:

```
img = torch.randn(1, 3, 10, 10, device='cuda')
x, y = 1, 1
width, height = 5, 5
img[:, :, y:y+height, x:x+width]
```

PyTorch Discussion Board

- Threads/topics:
  - Title (name)
  - Question (start post)
  - Stats
  - Solution?
  - Posts?



Slicing torch images as we do in numpy images

Vision

[redacted]

7h

I am working on a problem in which I have the coordinates to slice the image like [y, y+height, x, x+width]. So if if I have torch image obtained using

```
img = Variable(img.cuda())
```

how can we slice the image to get that specific area of image [y:y+height, x:x+width] .
Thanks

☑ Solved by ptrblck in post #2

You can directly index your image tensor: img = torch.randn(1, 3, 10, 10, device='cuda') x, y = 1, 1 width, height = 5, 5 img[:, :, y:y+height, x:x+width]

Reply

created 7h | last reply 7h | 1 reply | 11 views | 2 users | 1 like

ptrblck

7h

You can directly index your image tensor:

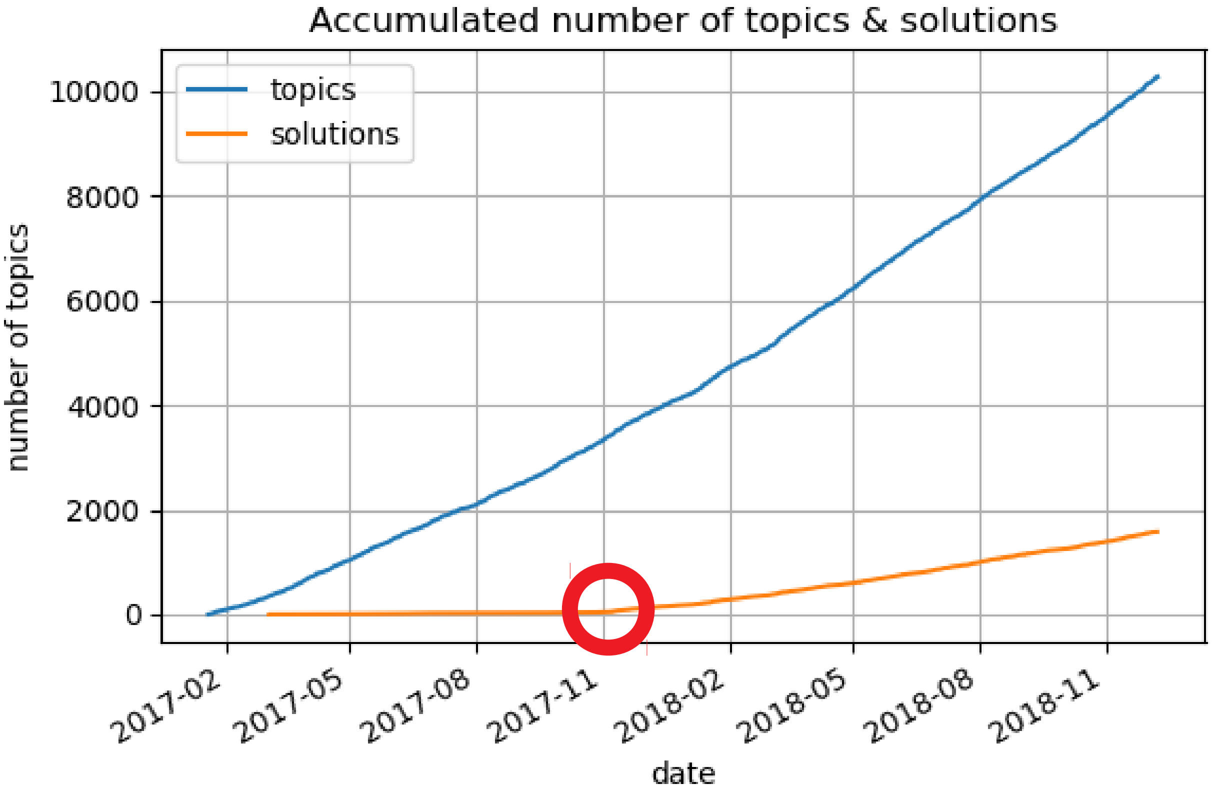
```
img = torch.randn(1, 3, 10, 10, device='cuda')
x, y = 1, 1
width, height = 5, 5
img[:, :, y:y+height, x:x+width]
```

Solution ☑ 1

Reply

PyTorch Discussion Board

- Some stats
  - ~**13,000** topics
  - ~52,000 posts
  - ~**1,700** marked solutions



PyTorch Discussion Board

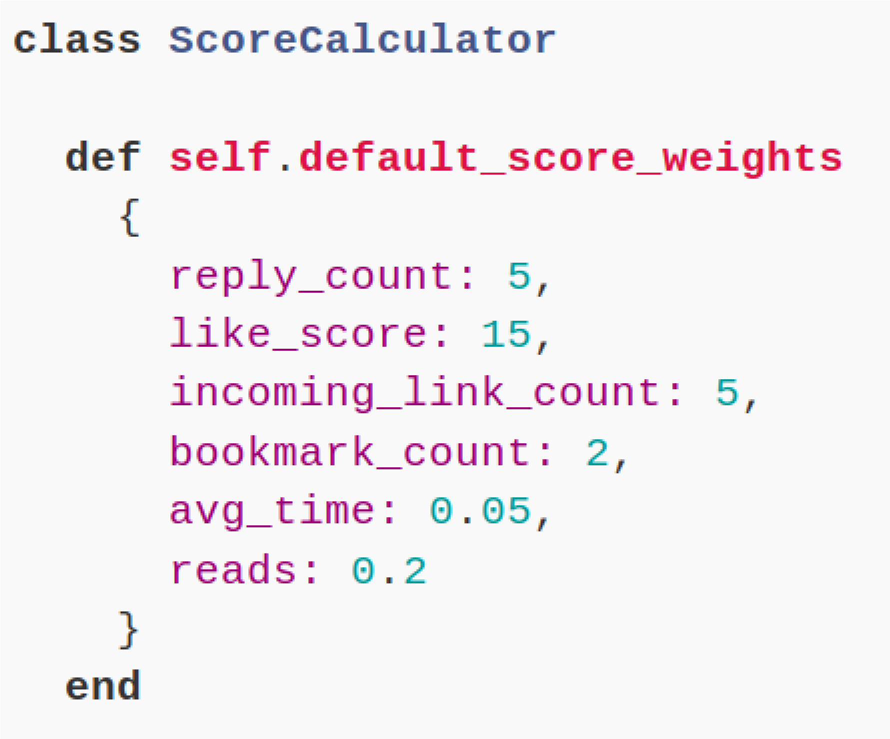
- **1,700** marked solutions might not be enough data to train deep learning model

PyTorch Discussion Board

- **1,700** marked solutions might not be enough data to train deep learning model
- But we have **13,000** topics!
- Workflow:
  - Use solution if available
  - Else: take post with highest **score** (not start post)

Discourse post score

- Uses
  - Reply count
  - Likes
  - Links
  - Bookmark count
  - Reading time?
  - Number of reads



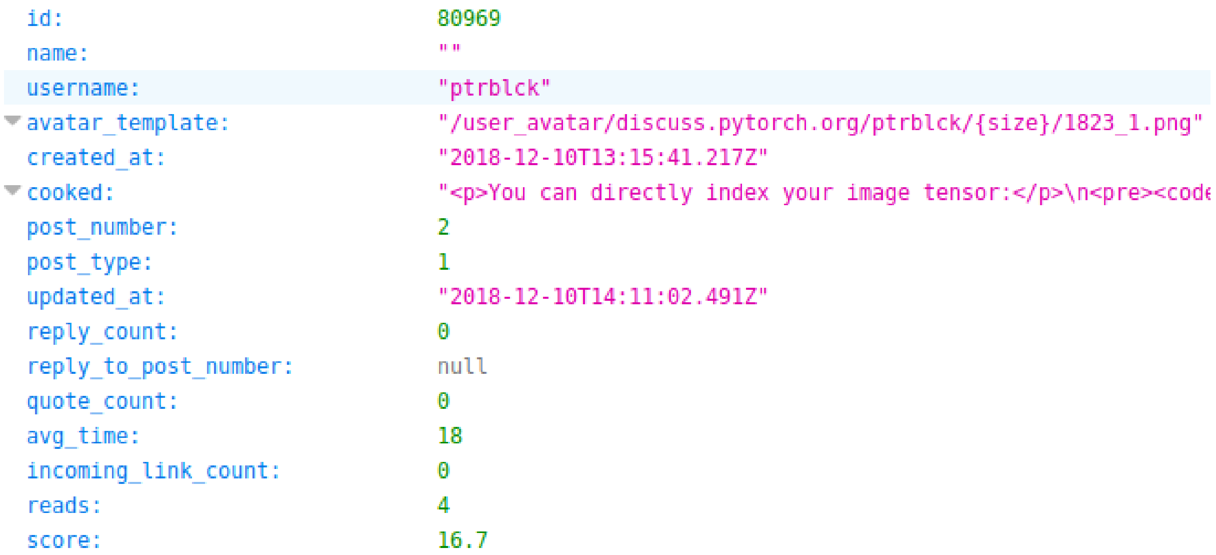
```ruby
class ScoreCalculator

  def self.default_score_weights
    {
      reply_count: 5,
      like_score: 15,
      incoming_link_count: 5,
      bookmark_count: 2,
      avg_time: 0.05,
      reads: 0.2
    }
  end
```

Overview

- Data
- Model
- Loss function
- Training
- Testing

Get the data

- All pulled information is public (indexed by Google)
- Use discourse REST API to get all posts
- Save title, question, solution (or highest scored post)
- Create two datasets
  - small dataset (only solutions)
  - Bigger dataset (solutions or best post)



```
id:                       80969
name:                     ""
username:                 "ptrblck"
avatar_template:          "/user_avatar/discuss.pytorch.org/ptrblck/{size}/1823_1.png"
created_at:               "2018-12-10T13:15:41.217Z"
cooked:                   "<p>You can directly index your image tensor:</p>\n<pre><code
post_number:              2
post_type:                1
updated_at:               "2018-12-10T14:11:02.491Z"
reply_count:              0
reply_to_post_number:     null
quote_count:              0
avg_time:                 18
incoming_link_count:      0
reads:                    4
score:                    16.7
```

Get the data

- REST API was quite easy to use (although the docs could get some more examples)
- Saved datasets:
  - Small dataset: 1582 threads
  - Bigger dataset: 10,280 threads

Get data

- Example of “raw” markdown data:
-
  ~~~
  "You can directly index your image tensor:\n```python\nimg = torch.randn(1, 3, 10, 10, device='cuda')\nx, y = 1, 1\nwidth, height = 5, 5\nimg[:, :, y:y+height, x:x+width]\n```"
  ~~~
- How to clean and preprocess this kind of data?

Preprocessing data

- Basic approach:
  - Tokenize the raw text
  - Lower all words
  - …
  - Create language (dictionary)
  - Done!

Preprocessing data

- Basic approach:
  - Tokenize the raw text
  - Lower all words
  - …
  - Create language (dictionary)
  - Done!
  - Maybe not :(
- Tokenization of code seems to fail

```
['you', 'can', 'directly',
'index', 'your', 'image',
'tensor', ':', '``', '`',
'python', 'img', '=',
'torch.randn(1', ',', '3',
',', '10', ',', '10', ',',
"device='cuda", "'", ')', 'x',
',', 'y', '=', '1', ',', '1',
'width', ',', 'height', '=',
'5', ',', '5', 'img', '[',
':', ',', ':', ',', 'y', ':',
'y+height', ',', 'x', ':',
'x+width', ']', '``', '`']
```

Preprocessing data

- New approach:
  - Tokenize text and code separately
    - Use regex to get markdown code
    - `` re_code = r'(?:(?<!\\)((?:\\{2})+)(?=`+)|(?<!\\)(`+)(.+?)(?<!`)\2(?!`))' ``
    - (Taken from https://github.com/Python-Markdown/markdown)
    - Also, remove all links (+ image links)
  - Use tokens to create language

Preprocessing data

- New approach:
  - Tokenize text and code separately
  - Use tokens to create language (dictionary)
    - Represent each word as one-hot encoded vector
    - Create lookup table for word – index
    - See PyTorch Seq2Seq Tutorial `(class Lang)`
      https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html

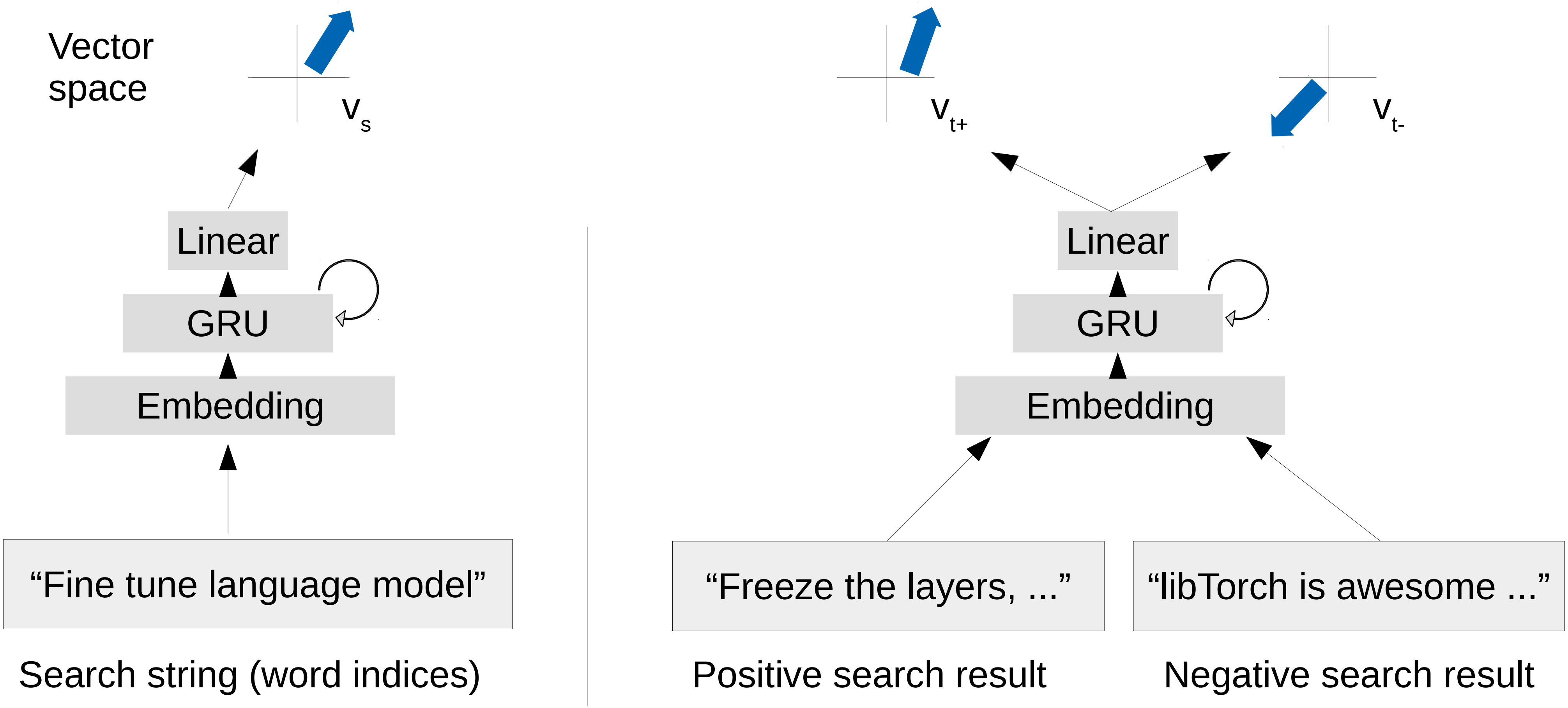
Overview

- ~~Data~~
- Model
- Loss function
- Training
- Testing

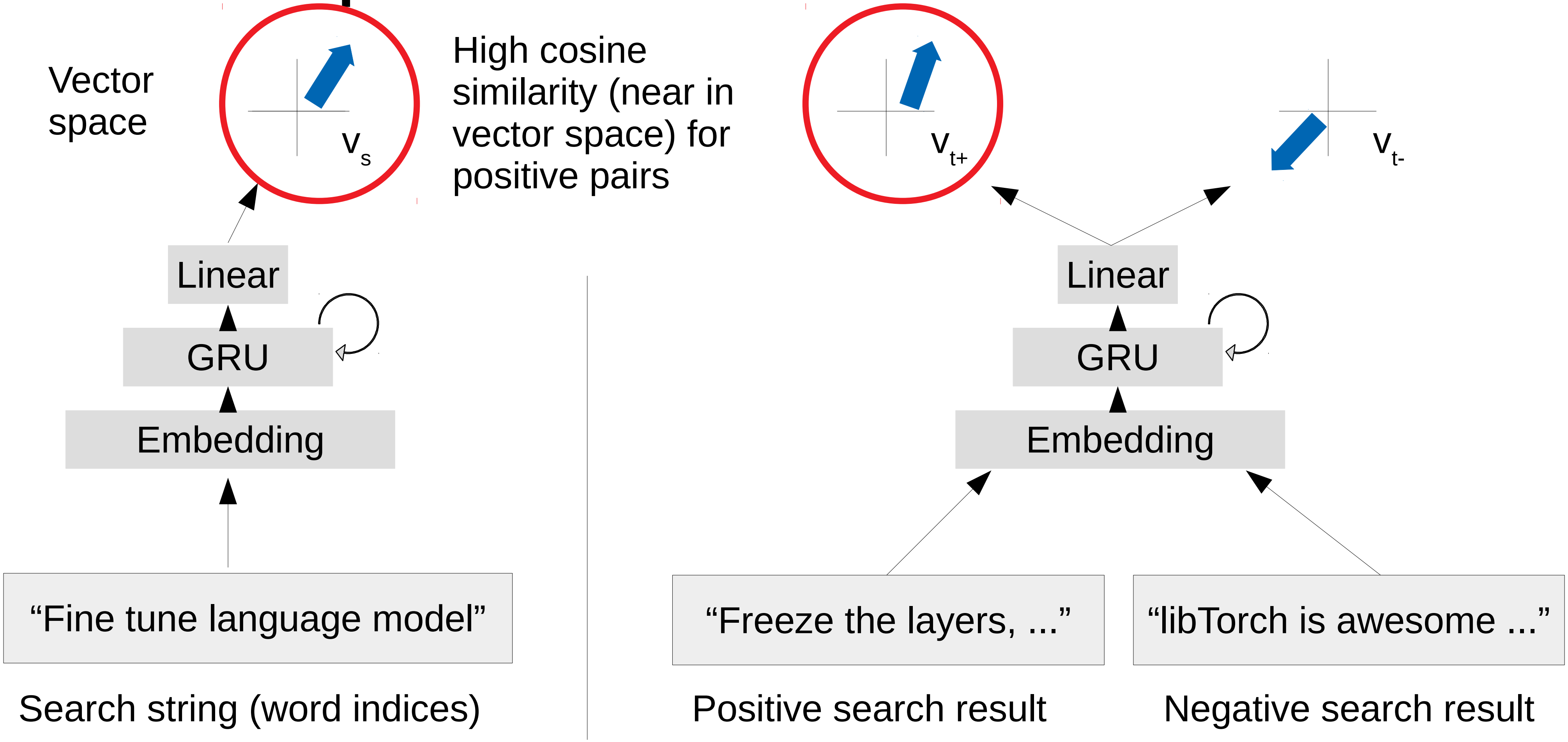
Simple baseline architecture

- Basic idea
  - Represent **search string** and **target** (answer/thread) in a shared vector space
  - Use two neural networks for mapping
  - Search strings and targets with same “meaning” should be close in vector space
  - Different meaning → far apart in vector space
  - Use cosine similarity to measure distance

Simple baseline architecture



Simple baseline architecture



Vector space

v_s

High cosine similarity (near in vector space) for positive pairs

v_{t+}

v_{t-}

Linear

GRU

Embedding

“Fine tune language model”

Search string (word indices)

Linear

GRU

Embedding

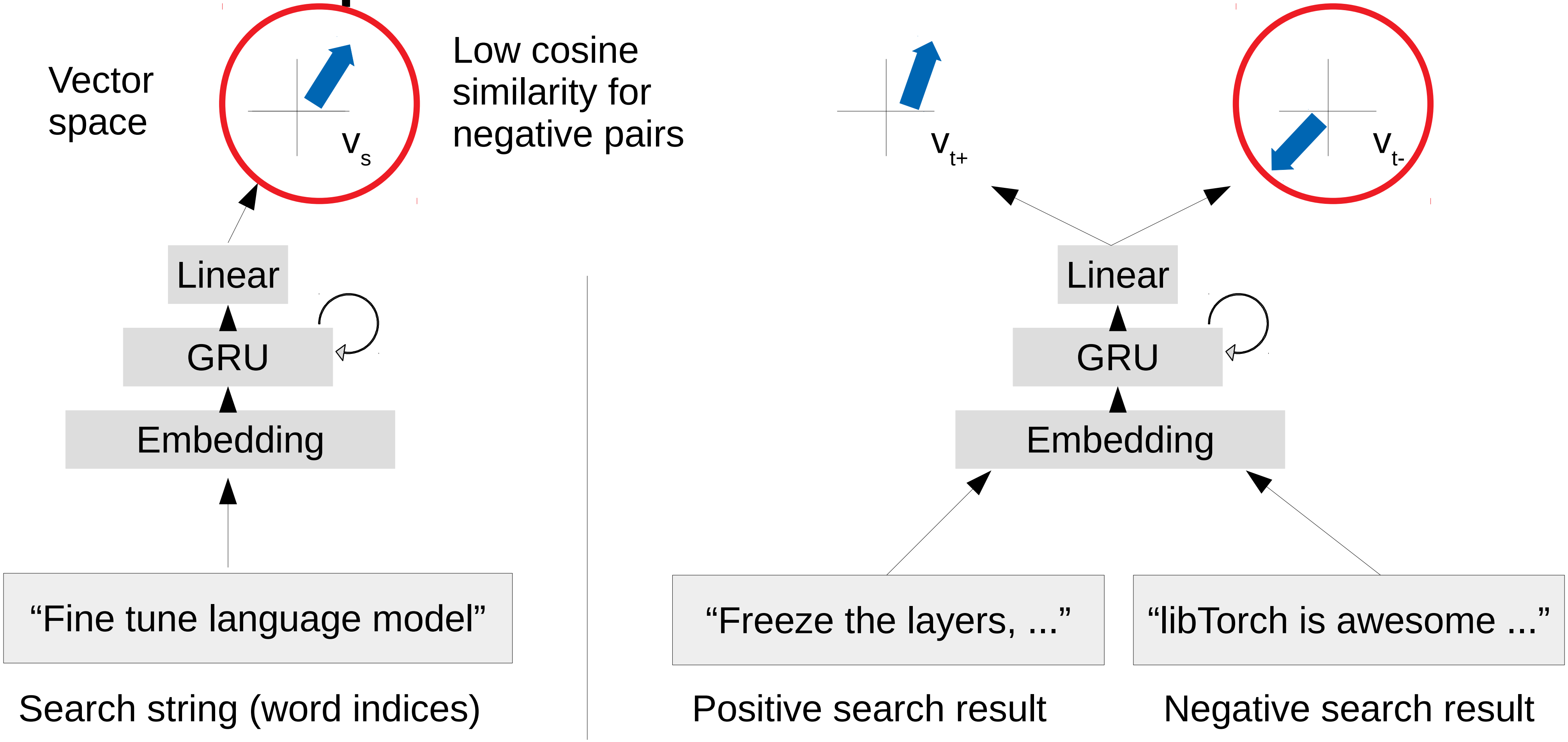
“Freeze the layers, ...”

“libTorch is awesome ...”

Positive search result

Negative search result

Simple baseline architecture



Vector space

v_s

Low cosine similarity for negative pairs

v_{t+}

v_{t-}

Linear

GRU

Embedding

Linear

GRU

Embedding

“Fine tune language model”

Search string (word indices)

“Freeze the layers, ...”

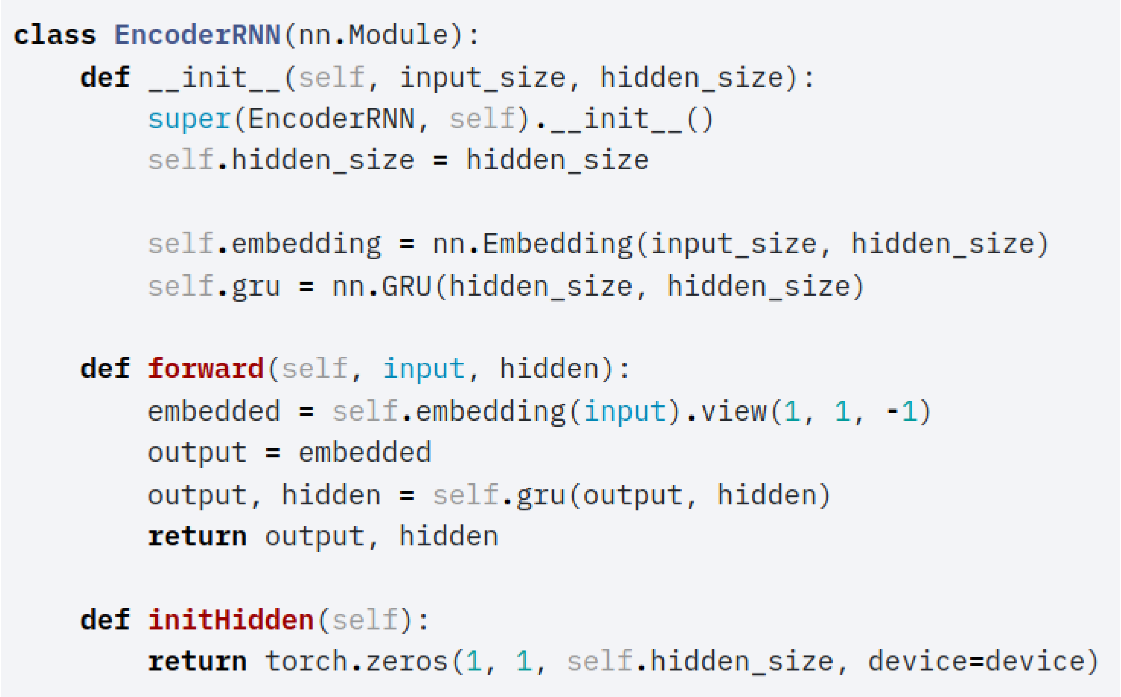
“libTorch is awesome ...”

Positive search result

Negative search result

Simple baseline architecture

- Use model (encoder) from PyTorch Seq2Seq tutorial
- Add linear layer(s) to learn vector space



```
class EncoderRNN(nn.Module):
    def __init__(self, input_size, hidden_size):
        super(EncoderRNN, self).__init__()
        self.hidden_size = hidden_size

        self.embedding = nn.Embedding(input_size, hidden_size)
        self.gru = nn.GRU(hidden_size, hidden_size)

    def forward(self, input, hidden):
        embedded = self.embedding(input).view(1, 1, -1)
        output = embedded
        output, hidden = self.gru(output, hidden)
        return output, hidden

    def initHidden(self):
        return torch.zeros(1, 1, self.hidden_size, device=device)
```

Overview

- ~~Data~~
- ~~Model~~
- Loss function
- Training
- Testing

Loss function

- Start with simple cosine similarity [-1, 1]
  - Should be high for positive pairs (sim+)
  - Low for negative samples (sim-)
  - Shift by 1 to get zero loss instead of negative values

    `loss = (1 – sim+) + (1 + sim–)`
  - Sum both similarities together
  - `nn.CosineSimilarity()`

Overview

- ~~Data~~
- ~~Model~~
- ~~Loss function~~
- Training
- Testing

Train the baseline models

- Use standard setup
  (SGD, lr=1e-3, batch_size=64, …)
- Start with small dataset
- …

Train the baseline models

- Use standard setup
  (SGD, lr=1e-3, batch_size=64, …)
- Start with small dataset
- …
- Fail: Training+Validation loss hardly moving

Train the baseline models

- Tune model hyperparameters (layer size)

Train the baseline models

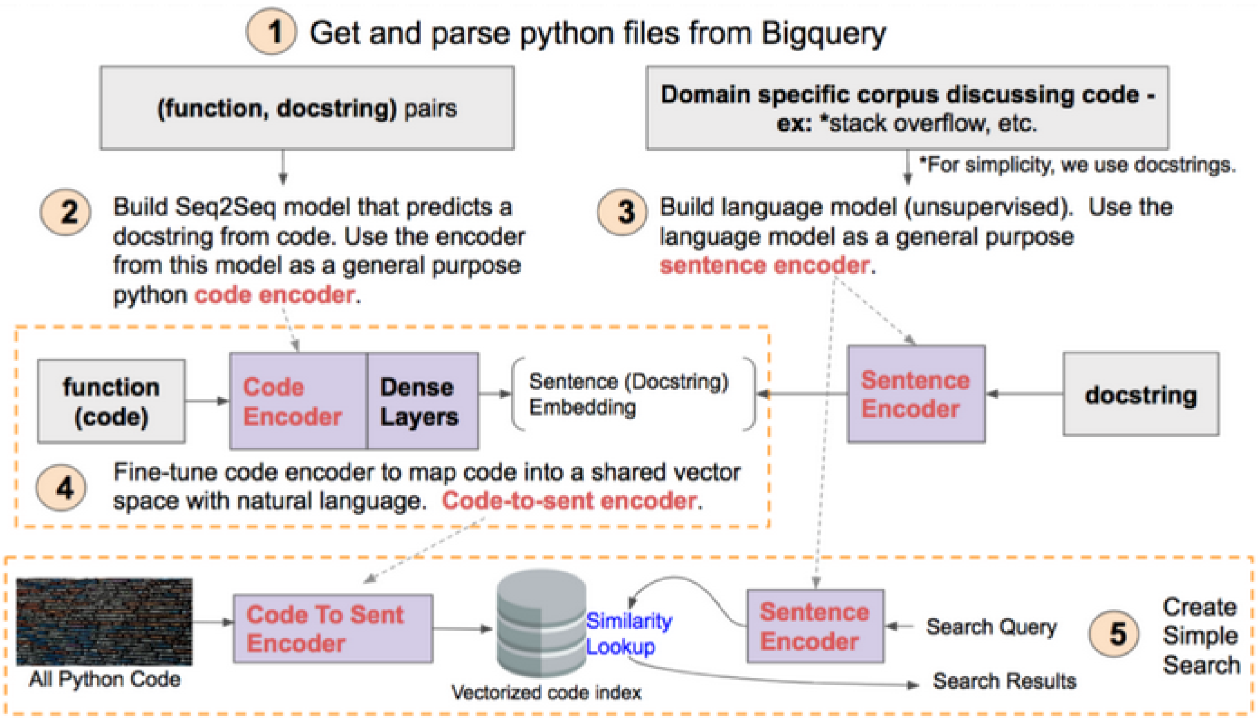
- Tune model hyperparameters (layer size)
- Fail

Train the baseline models

- Things that have failed:
  - Model hyperparameter tuning
  - Tuning of optimization hyperparams
    (lr, weight decay, different optimizer)
  - Adding some regularization (BatchNorm, Dropout)
  - Change GRU (bidirectional, more layers)
  - Use the bigger dataset
  - Use shorter sequences (cut or remove longer sequences)
- Nothing seems to be working!

Train the baseline models

- What have Husain & Wu done?
- Steps 2 and 3 create a “language model” for both networks
- Step 4 learn the shared vector space



1 Get and parse python files from Bigquery

2 Build Seq2Seq model that predicts a docstring from code. Use the encoder from this model as a general purpose python code encoder.

3 Build language model (unsupervised). Use the language model as a general purpose sentence encoder.

*For simplicity, we use docstrings.

4 Fine-tune code encoder to map code into a shared vector space with natural language. Code-to-sent encoder.

5 Create Simple Search

Train the baseline models

- What have Husain & Wu done?
- Two different approaches for pretraining
  - Seq2Seq model (use only encoder)
  - Try to learn to predict next word

Train the baseline models

- What have Husain & Wu done?
- Two different approaches for pretraining
  - Seq2Seq model (use only encoder)
    - Failed: probably too little data?
  - Try to learn to predict next word
    - Failed: No natural language (mixture of text + code)?

Overview

- Data
- ~~Model~~
- ~~Loss function~~
- Training
- Testing

Back to Step1!

Review the data

- Data consists of
  - A lot of numbers
    - Tensor/model shapes, random values etc.
  - A lot of single letter words
    - Variable names, etc.

Review the data

- Data consists of
  - A lot of numbers
    - Tensor/model shapes, random values etc.
  - A lot of single letter words
    - Variable names, etc.
- Remove these and try training again with small dataset

Review the data

- Data consists of
  - A lot of numbers
    - Tensor/model shapes, random values etc.
  - A lot of single letter words
    - Variable names, etc.
- Remove these and try training again with small dataset
- (Half) Fail: Model trades sim+ for sim- (at least moving at all!)

Train the baseline model

- Pretrain models using just search strings
  - Maybe this way the “language” will be learned?
- Then add targets to datasets

Train the baseline model

- Pretrain models using just search strings
  - Maybe this way the “language” will be learned?
- Then add targets to datasets
- Works OK! First success!
- Validation loss is still high
  - … but it’s a first step ;)

Train the baseline model

- Pretrain models using just search strings
- Then add targets to datasets
- Change hyperparameters around
  - Add or remove capacity from models
  - Observe the losses
- Switch back to bigger dataset
- Change loss function to `log(1 + exp(-1.0*((sim+)-(sim-))))`
  - Taken from Geo et. al, “An Introduction to Deep Learning for Natural Language Processing”, Microsoft Research

Train the baseline model

- Works alright!
- Training and validation losses going down
  - Not as I would have wished, but anyway

Overview

- ~~Data~~
- ~~Model~~
- ~~Loss function~~
- ~~Training~~
- Testing

Testing on hold-out set

- Question: 'tensor is not contiguous'
- Top10 Answers:

- how to compile pytorch from source without cuda default location
- how to merge by avg multiple inputs to layer
- how to specify gpu usage
- why does this assignment operation of variable not work
- apply part of tensor on function to avoid out of memory
- **shuffle elements of tensor**
- method object does not support item assignment
- **tensor slicing on 3 dim tensors**
- how to extend tensors inside variable
- tensor and variable are the same now

Testing on hold-out set

- Question: 'how to broadcast tensor'
- Top10 Answers:

- **how to flip a tensor along a specified axis**
- how to get the current value of a variable
- **how to broadcast a 1d tensor with a 4d tensor**
- how to transfer an existing tensor to another device based on other tensor
- how to get all registerd buffer by self register buffer
- how can l use the pre trained resnet to extract feautres from my own dataset
- how to convert a normal variable into a regular variable that can be inputted to a loss function
- **how to merge tensor with weights**
- how to choose a suitable weight decay
- how to keey the weight of conv layer unchanged

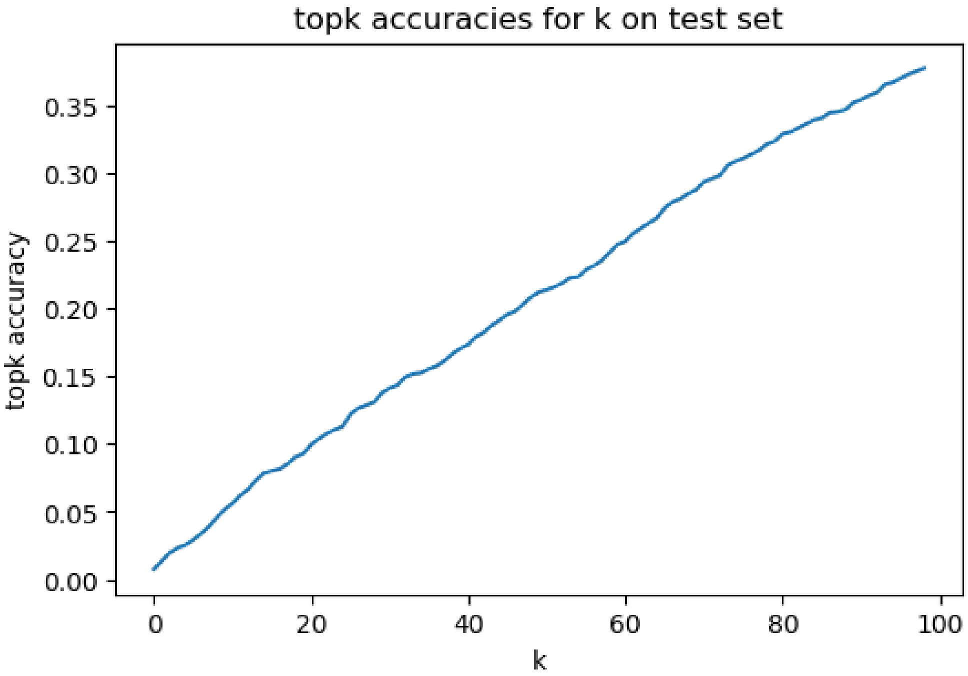
Testing on hold-out set

- Question: 'model is performing bad'
- Top10 Answers:

- unusual large memory for con2d with batch size 1
- gpu high memory usage low gpu volatile util
- pytorch example with cnn based object detection
- **error loading bidirectional lstm model**
- question about thstorage
- too many resources requested for launch
- what is pytorch
- **what is nn embedding exactly doing**
- will conda install pytorch torchvision c pytorch also install cuda and cudnn
- loading pytorch checkpoint in tf keras

Testing on hold-out set

- Based on these results let’s rather call this talk ”*First steps towards* semantic search on PyTorch discussions”
- Top10 accuracy: ~7%
- Top10 random: 10/1337=~0.7%



PyTorch

Thanks a lot to **all of you** for
being such a great community!

Make sure to create an account at
https://discuss.pytorch.org ;)

Now let’s have some beers and pizza,
and hang out together!

Semantic Search on
PyTorch discussions

PyTorch in Munich at
Microsoft

Munich Applied
Deep Learning Meetup
Dec 11th, 2018
Piotr Bialecki