AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

CSCE 771: Computer Processing of Natural Language
Lecture: Pytorch and BERT

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

24th SEP 2024

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 11

- Opening Segment
  - Review of Lecture 10
- Main Lecture
- Concluding Segment
  - About Next Lecture – Lecture 12

Main Section
- PyTorch
- Evaluating LMs
- BERT

Recap of Lecture 10

- We reviewed Machine Learning methods
  - Data preparation is the key
  - Watch out for evaluation
  - ML is just a step, what happens to the model is also important
- Language models
  - We reviewed connections between parsing and language model
  - We discussed language models and are focusing on pre-requisites needed to understand them
  - We discussed contextual word representations as a stepping stone

Main Lecture

Kaushik Roy

South Carolina

University of South Carolina

[Webpage]

I am Kaushik Roy, a Ph.D. candidate at the Artificial Intelligence Institute, University of South Carolina. My research focuses on developing neurosymbolic methods for declarative and process knowledge-infused learning, reasoning, and sequential decision-making, with a particular emphasis on social good applications. My academic journey has taken me from R.V. College of Engineering in Bangalore for my Bachelor's to Indiana University Bloomington for my Master's, and briefly to the University of Texas at Dallas before settling at the University of South Carolina for my doctoral studies. My research interests span machine learning, artificial intelligence, and their application in social good settings. I'm passionate about pushing the boundaries of AI, particularly in areas where it intersects with human understanding and decision-making.

Topics of Interest to Me: Neurosymbolic AI Knowledge-infused Learning AI for Social Good Healthcare Informatics

Organization of This Lecture
Part 1: Pytorch

- Introduction Section
  - Pytorch
- Main Section

Main Section

- What is a Pytorch?
  - Old-school differentiation
  - Automatic differentiation
- Processing data
  - Scalars
  - Vectors, Matrices
  - Tensors

Organization of This Lecture
PART 2: BERT

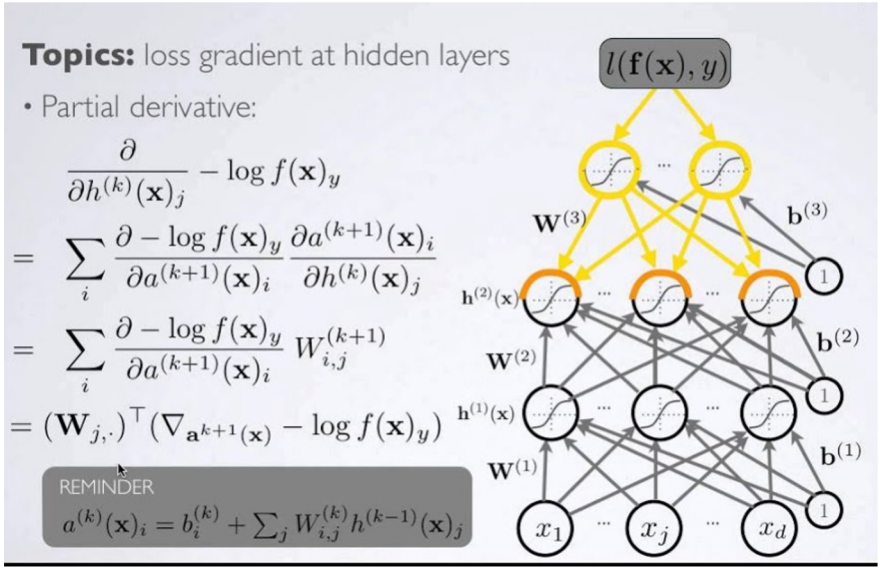
- Introduction Section
  - Language Modeling
- Main Section 
- Concluding Section
    - Next lecture Mamba

Main Section

- What is Language Modeling?
  - BERT-and family
  - Live Coding - Two CV tasks
- Concluding Comments
  - BERT lib on CV
  - BERT scratch on CV
  - Abstract Math Objects

Image Source

What is Pytorch?



Old-school differentiation

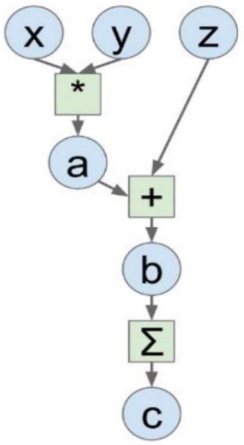
- Processing data
  - Scalars?
  - Vectors, Matrices?
  - Tensors?
- Abstract Mathematical Objects?

- Prior to libraries such as pytorch, we needed to hand derive the derivatives of neural network architectures, which could end up being prone to manual errors!

Image Source

Why is Pytorch?

Computation Graph




```
import torch

N, D = 3, 4

x = torch.rand((N, D),requires_grad=True)
y = torch.rand((N, D),requires_grad=True)
z = torch.rand((N, D),requires_grad=True)

a =x * y
b =a + z
c=torch.sum(b)

c.backward()
```

Automating differentiation

- Processing data
  - Scalars ✓
  - Vectors, Matrices ✓
  - Tensors ✓
- Abstract Mathematical Objects?

- Pytorch allows arbitrary compositions of mathematical object as functions, while also simultaneously and **automatically** maintaining the derivative of the function with respect to it's parameters!

Brief History

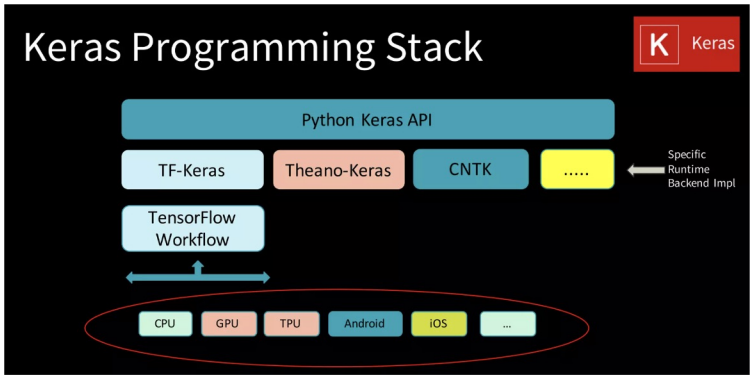
- **Big data**
  - Very large volumes of raw data
- **Deep Learning boom**

Image Source

Automatic Differentiation

- Tensorflow
  - Released by Google
  - Used software artifacts such as constants, variables, placeholders, operations, sessions, tensors, graphs
  - Hugely popular, but a bit difficult to manage (complex) – APIs built to make it easier (Keras)

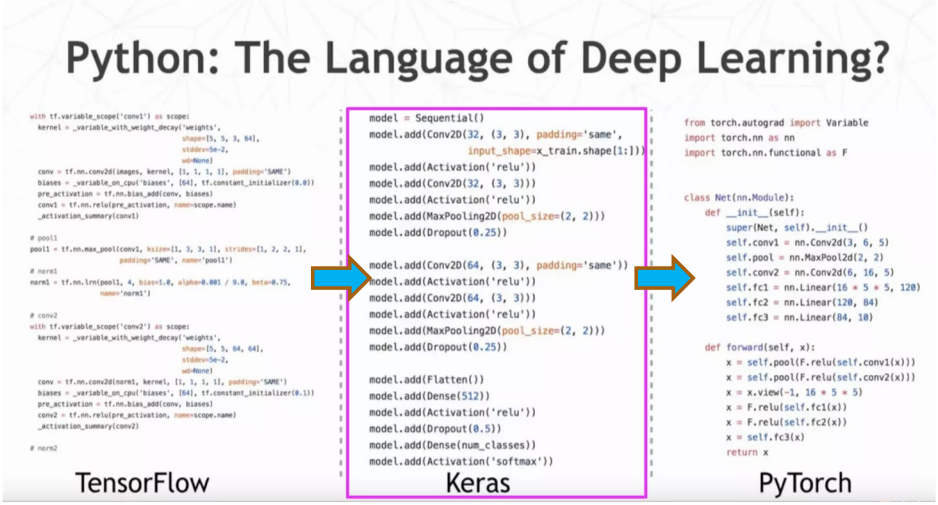
Image Source



Brief History

- **Python as language of deep learning**
  - Many libraries in python
- **Python-inspired syntax**

Image Source



Automatic Differentiation

- Pytorch
  - Released by Meta (Formerly Facebook)
  - Used software artifacts such as Tensors of various dimensions, optimizers, and extremely intuitive due to it's **pythonic** nature
  - Hugely popular, and today most popular for **rapid research prototyping**

Image Source

Tiny Example

Autograd

- Automatic Differentiation Package
- Don't need to worry about partial differentiation, chain rule etc.
  - `backward()` does that
- Gradients are accumulated for each step by default:
  - Need to zero out gradients after each update
  - `tensor.grad_zero()`

```
# Create tensors.
x = torch.tensor(1., requires_grad=True)
w = torch.tensor(2., requires_grad=True)
b = torch.tensor(3., requires_grad=True)

# Build a computational graph.
y = w * x + b    # y = 2 * x + 3

# Compute gradients.
y.backward()

# Print out the gradients.
print(x.grad)    # x.grad = 2
print(w.grad)    # w.grad = 1
print(b.grad)    # b.grad = 1
```

- This code shows a simple composition of scalars into a linear function
- Note how we never specify the gradients manually, and **pytorch automatically calculates it during a call to .backward()**

Image Source

Tiny Example

Optimizer and Loss

Optimizer

- Adam, SGD etc.
- An optimizer takes the parameters we want to update, the learning rate we want to use along with other hyper-parameters and performs the updates

Loss

- Various predefined loss functions to choose from
- L1, MSE, Cross Entropy

```
a = torch.randn(1, requires_grad=True, dtype=torch.float, device=device)
b = torch.randn(1, requires_grad=True, dtype=torch.float, device=device)


# Defines a SGD optimizer to update the parameters
optimizer = optim.SGD([a, b], lr=lr)

for epoch in range(n_epochs):
    yhat = a + b * x_train_tensor
    error = y_train_tensor - yhat
    loss = (error ** 2).mean()

    loss.backward()

    optimizer.step()

    optimizer.zero_grad()

print(a, b)
```

- Pytorch allows a wide range of optimization techniques, such as **stochastic gradient descent (SGD)**, already built in it's optimizer toolkit

Image Source

Tiny Example

Model

In PyTorch, a model is represented by a regular Python class that inherits from the Module class.

- Two components
  - `__init__(self)` : it defines the parts that make up the model- in our case, two parameters, a and b
  - `forward(self, x)` : it performs the actual computation, that is, it outputs a prediction, given the inputx

```
class ManualLinearRegression(nn.Module):
    def __init__(self):
        super().__init__()
        # To make "a" and "b" real parameters of the model, we need to wrap them with nn.Parameter
        self.a = nn.Parameter(torch.randn(1, requires_grad=True, dtype=torch.float))
        self.b = nn.Parameter(torch.randn(1, requires_grad=True, dtype=torch.float))

    def forward(self, x):
        # Computes the outputs / predictions
        return self.a + self.b * x
```

- A pytorch model is easily specified as a class
- Next we will demonstrate how effective pytorch is using the CV dataset

Data - CV data as a PDF

Follow Links

Links to my Google Scholar and Socials

South Carolina
University of South Carolina
Email
Google Scholar
Github
LinkedIn

Curriculum Vitae, Updated September 14th 2024.

🎓 Education

- **2020 – Present**: Ph.D. Candidate, University of South Carolina
  - Topic: Process Knowledge-infused Learning and Reasoning
- **2017 – 2020**: Ph.D. Student, University of Texas at Dallas (Transferred in 2020)
  - Topic: Relational Sequential Decision Making [Qualifier Document]
- **2015 – 2017**: M.Sc. Computer Science, Indiana University Bloomington
  - Specialization: Artificial Intelligence and Machine Learning
- **2011 – 2015**: B.E. Computer Science, RV College of Engineering
  - Thesis title: Computer Vision Algorithms for Background Understanding [Thesis Document]

📚 Publications

📘 Book Chapters

Tasks

- CV-related tasks
  - Extractive Summarization – Extract Institution, Degree, and Year
    - Institution: UofSC
    - Degree: Ph.D.
    - Year: 2024
  - Abstractive Summarization - Masked Language Filling with Partial Queries
    - Kaushik Roy is a ?__

Data - Preprocessing Steps

Key Steps in Preprocessing the CV Text:

1. **Convert the PDF into Plain Text**:
   - Use a library like **pdfplumber** to extract the text from the PDF file.
2. **Text Cleaning**:
   - Remove or replace unnecessary characters such as newline characters (\n), multiple spaces, tabs (\t), etc.
   - Normalize any irregular characters.
3. **Segment the Text**:
   - Break the text into sentences or paragraphs, which can be useful when preparing data for sequence-based tasks.
4. **Tokenization**:
   - Use a tokenizer like **BERT's tokenizer** to convert the cleaned text into a format that can be used by your model.

Step 1 - Convert PDF into Plain Text

You can use **pdfplumber** to extract text from a PDF file. The extracted text may contain newline characters, extra spaces, and other artifacts that need to be cleaned.

Code Snapshot:

```python
import pdfplumber

def extract_text_from_pdf(pdf_path):
    """
    Extract text from a PDF file using pdfplumber.

    Args:
        pdf_path (str): Path to the PDF file.

    Returns:
        str: Extracted text from the PDF.
    """
    with pdfplumber.open(pdf_path) as pdf:
        text = ''
        for page in pdf.pages:
            text += page.extract_text()  # Extract text from each p
    return text
```

Step 2 - Text Cleaning

Once you've extracted the text from the PDF, it might contain newlines, multiple spaces, and other irregular characters that need to be cleaned. Here's what we can do:

- **Remove newline characters** (\n) and replace them with spaces (or punctuation marks like periods, depending on how the text is structured).
- **Normalize white spaces**: Collapse multiple spaces into a single space.
- **Remove special characters**: Remove characters like tabs, and optionally remove non-alphabetic characters.

Code Snapshot:

```python
import re

def clean_text(text):
    """
    Clean the extracted text by removing newlines, multiple spaces,

    Args:
        text (str): The raw extracted text.

    Returns:
        str: Cleaned text.
    """
    # Replace newline characters with spaces
    text = text.replace('\n', ' ').replace('\r', ' ')

    # Remove tabs and multiple spaces
    text = re.sub(r'\s+', ' ', text)  # Replace multiple spaces wi

    # Optionally remove non-alphanumeric characters (if needed)
    # text = re.sub(r'[^a-zA-Z0-9.,!?\'\s]', '', text)

    # Strip leading/trailing spaces
    text = text.strip()

    return text
```

Step 3 - Segment the Text

- Depending on how you want to use the text in your model, you might want to segment it into smaller units like **sentences** or **paragraphs**.
- If the extracted text is structured, you can split it into sentences using **nltk** or **spaCy** to make the text more manageable.
- **Code Example for Segmenting Text into Sentences Using nltk:**

Code Snapshot:

```
import nltk
nltk.download('punkt')  # Download the sentence tokenizer

def segment_text_into_sentences(text):
    """
    Segment the cleaned text into sentences.

    Args:
        text (str): The cleaned text.

    Returns:
        list: A list of sentences.
    """
    from nltk.tokenize import sent_tokenize
    sentences = sent_tokenize(text)  # Tokenize the text into sent
    return sentences
```

Step 4 - Tokenization

- Once the text is cleaned and segmented, you can tokenize it using **BERT's tokenizer** or any other suitable tokenizer.

Code Snapshot:

```
from transformers import BertTokenizer

# Load the BERT tokenizer
tokenizer = BertTokenizer.from_pretrained("bert-base-uncased")

def tokenize_text(sentences):
    """
    Tokenize a list of sentences using the BERT tokenizer.

    Args:
        sentences (list): A list of sentences.

    Returns:
        list: A list of tokenized sentences (token IDs).
    """
    tokenized_sentences = [tokenizer.encode(sentence, return_tensor
    return tokenized_sentences
```

Live Coding

Github:
https://github.com/kauroy1994/Teaching
(Neural Network-based CV information filling)

- Simply using a feedforward neural network with a position embedding layer to solve the CV tasks

Code Snapshot:

```
class generator(nn.Module):

  def __init__(self,
               n_tokens = None,
               emb_size = None,
               context_size = None,
               n_layers = 2,
               h_size = 100):

    super().__init__()

    self.n_tokens = n_tokens
    self.emb_size = emb_size
    self.context_size = context_size
    self.n_layers = n_layers
    self.h_size = h_size

    self.embeddings = nn.Embedding(self.n_tokens, self.emb_size)
    self.pos_embeddings = nn.Embedding(self.context_size, self.emb_size)

    self.fc1 = nn.Linear(self.emb_size,self.h_size,bias=False)
    self.fc2 = nn.Linear(self.h_size,self.h_size,bias=False)
    self.head = nn.Linear(self.h_size,self.n_tokens)

  def forward(self,
              token_encodings):

    n_tokens = len(token_encodings)
    token_encodings = torch.tensor(token_encodings)
    token_encodings.to(device)
    token_embeddings = self.embeddings(token_encodings)
    pos_embeddings = self.pos_embeddings(torch.arange(n_tokens))
    token_embeddings += pos_embeddings

    reps = token_embeddings
    reps = F.leaky_relu(self.fc1(reps))
    reps = F.leaky_relu(self.fc2(reps))
    reps = self.head(reps)

    logits = reps[-1]
    return logits
```

Evaluating Language Models

Evaluation – Language Model

- **Intrinsic evaluation**: measure the quality of a model independent of any application
- E**xtrinsic evaluation**: situate model in an application and evaluate the whole application for improvement. Also called in-vivo evaluation

Perplexity

$$\text{Average NLL} = -\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_1, w_2, ..., w_{i-1})$$

where N is the total number of words in the test dataset, and $P(w_i|w_1, w_2, ..., w_{i-}$ is the probability of word w_i given the previous words $w_1, w_2, ..., w_{i-1}$.

$$\text{Perplexity} = e^{\text{Average NLL}}$$

Value range: best: 1, worst: positive infinite; practical upper bound: number of words in vocabulary

1. P(John) = 0.1
2. P(bought | John) = 0.4
3. P(apples | bought) = 0.3
4. P(from | apples) = 0.5
5. P(the | from) = 0.6
6. P(market | the) = 0.7

Now, let's compute the probability of the generated sequence:

$P(\text{"John bought apples from the market"}) = P(John) \times P(bought|John) \times P(apples|bought) \times P(from|apples) \times P(the|from) \times P(market|the)$

$P(\text{"John bought apples from the market"}) = 0.1 \times 0.4 \times 0.3 \times 0.5 \times 0.6 \times 0.7$

Hence, P("John bought apples from the market") = 0.00252

Average NLL = – log(0.00252) / 6. [N = 6 as the model generated six words]

Hence, Average NLL = 0.99725

Perplexity = Exp(Average NLL) = 2.71

Credit:

- https://blog.uptrain.ai/decoding-perplexity-and-its-significance-in-llms/
- https://d2l.ai/chapter_recurrent-neural-networks/language-model.html

Perplexity Comments

1. A model with a vocabulary of 10,000 words and a perplexity of 2.71 is much better than a model with a vocabulary of 100 words and the same perplexity score of 2.71.
2. Lower perplexity results in higher consistency. As we know, LLMs are non-deterministic, i.e., the same inputs can result in two different outputs; a lower perplexity means that the model is more likely to produce the same output over multiple runs.
3. Perplexity is the inverse of the geometric mean of the probability of each word. Hence, the inverse of average probability (2.3077 in the previous case) can be considered a good proxy for quick calculations.
4. This calculation happens in the tokens space (compared to the words space), but the core principle remains the same.

Credit:

- https://blog.uptrain.ai/decoding-perplexity-and-its-significance-in-llms/

Perplexity

- Suppose
  - P('X') = 0.25
  - P('Y') = 0.5
  - P('Z') = 0.25
- Perplexity
  - ('XXX') = - exp (log (0.25 x 0.25 x 0.25) * (1/3)) = 3.94
  - Perplexity ('XYX') = - exp(log(0.25 x 0.5 x 0.25) * (1/3)) =
- Lower the number, the better is the model

Perplexity (Approx Calculation)

- Intrinsic evaluation
- **Definition**: perplexity of a language model on a test set is the inverse probability of the test set, normalized by the number of words

$$\begin{aligned} \text{PP}(W) &= P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}} \\ &= \sqrt[N]{\frac{1}{P(w_1 w_2 \ldots w_N)}} \end{aligned}$$

Of bi-grams

$$\text{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_{i-1})}}$$

Example: digits – 0 ..9, assuming equal probab. of 0.1

$$\begin{aligned} \text{PP}(W) &= P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}} \\ &= \left(\frac{1}{10}^{N}\right)^{-\frac{1}{N}} \\ &= \frac{1}{10}^{-1} \\ &= 10 \end{aligned}$$

From Jurafsky & Martin

Language Model: BERT/ Transformer

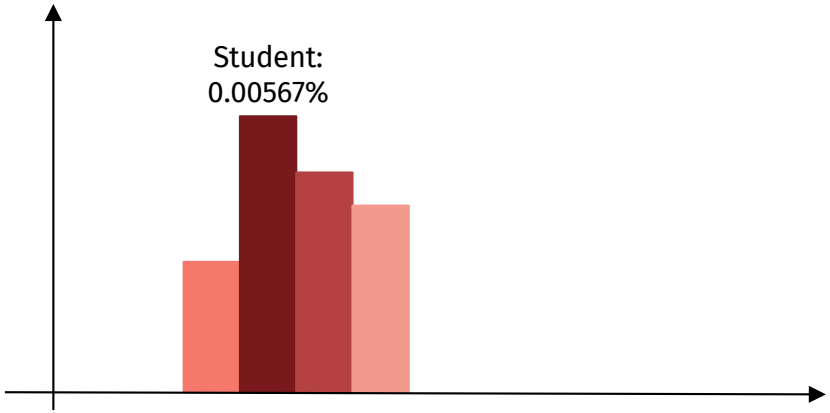
Language Modeling

Input Kaushik Roy is a PhD Student

Sequence	Kaushik	Roy	is	a	PhD	Student
	1	2	3	4	5	6

Student

Did you mean: Student
Did you mean: Student Square
Did you mean: Student Success



BERT - **B**idirectional **E**ncoder **R**epresentations from **T**ransformers

Learns with two tasks

- Predicting missing words in sentences
  - mask out 15% of the words in the input, predict the masked words.
- Given two sentences A and B, is B the actual next sentence that comes after A, or just a random sentence from the corpus?

(12-layer to 24-layer Transformer)
on (Wikipedia + BookCorpus)

Input: the man went to the [MASK1] . he bought a [MASK2] of milk.
Labels: [MASK1] = store; [MASK2] = gallon

Sentence A: the man went to the store .
Sentence B: he bought a gallon of milk .
Label: IsNextSentence

Sentence A: the man went to the store .
Sentence B: penguins are flightless .
Label: NotNextSentence

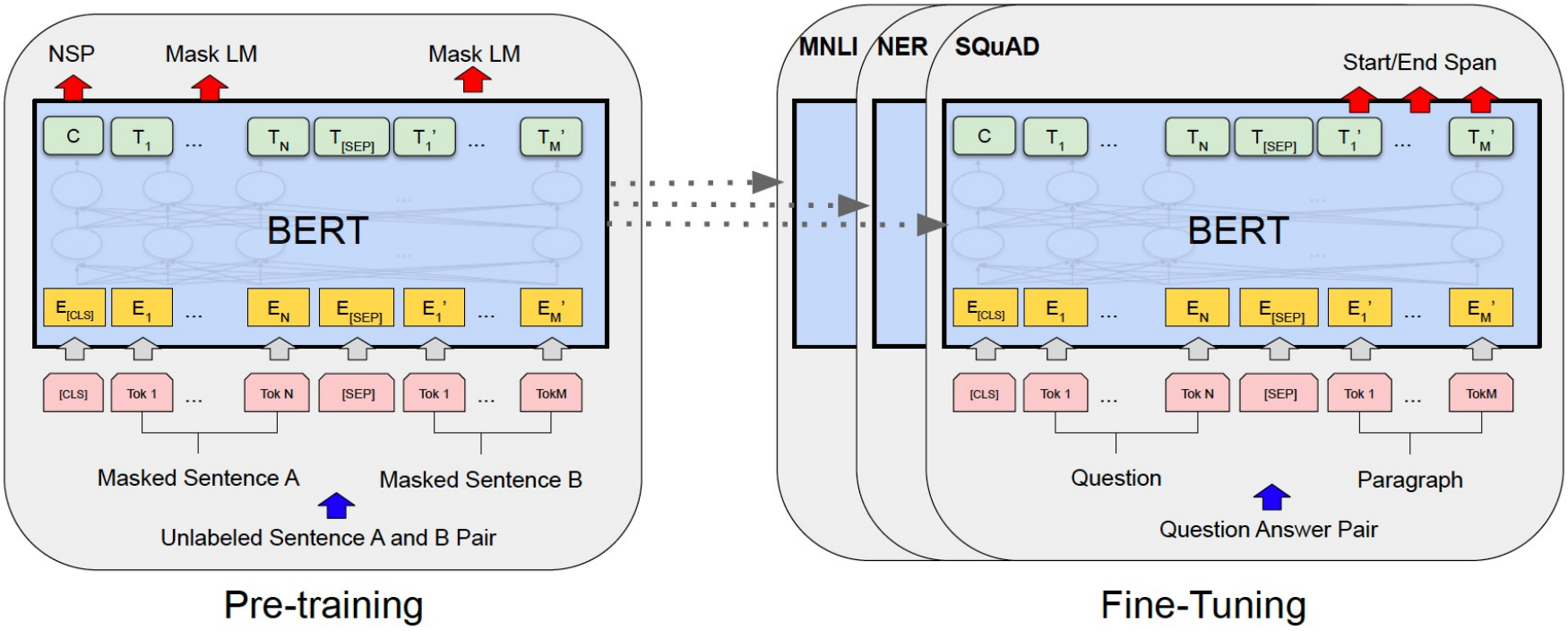
Credit and details: https://github.com/google-research/bert

Transformer

- RNN/ LSTM with
  - Attention
    - attention layer can access all previous states and weighs them according to some learned measure of relevancy to the current token, providing sharper information about far-away relevant tokens
    - **Query** vector, **Key** vector, and **Value** vectors introduced during encoding and decoding phase
  - Parallelization of learning
  - See Dr. Amitava Das's slide for Attention/ BERT video
    - Material: https://prezi.com/view/amx5hBo8UhMOn1rPyJ02/

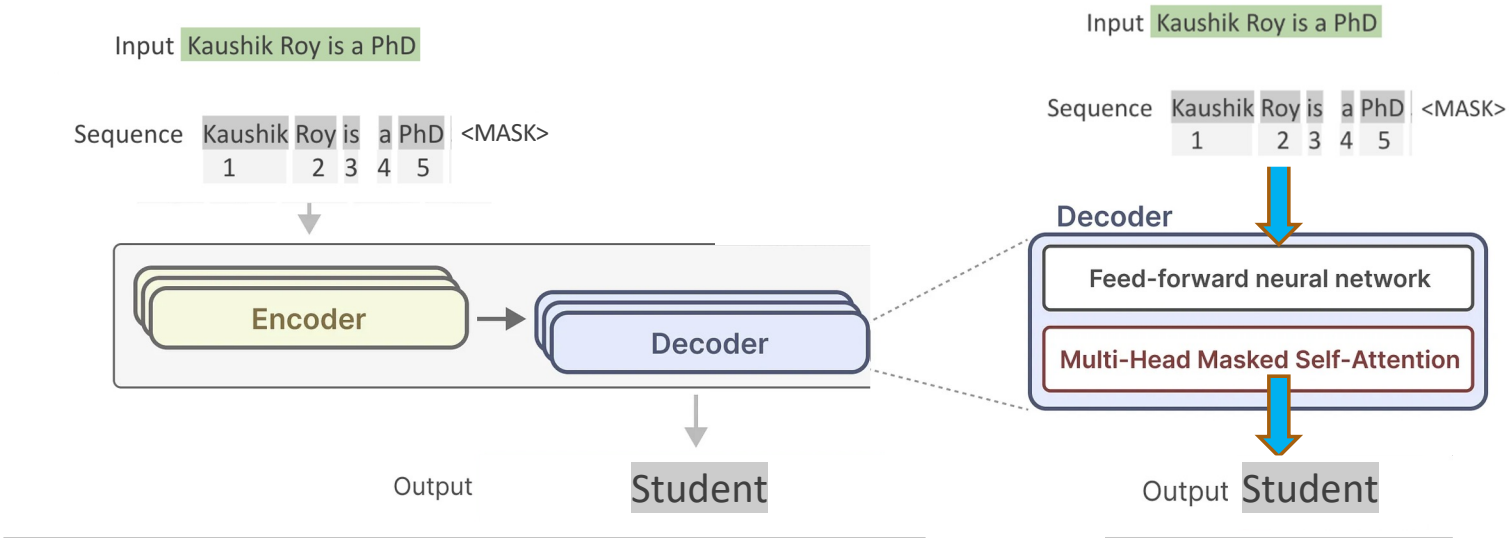
Source and details: https://en.wikipedia.org/wiki/Transformer_(machine_learning_model), http://jalammar.github.io/illustrated-transformer/

BERT: Before and During Usage

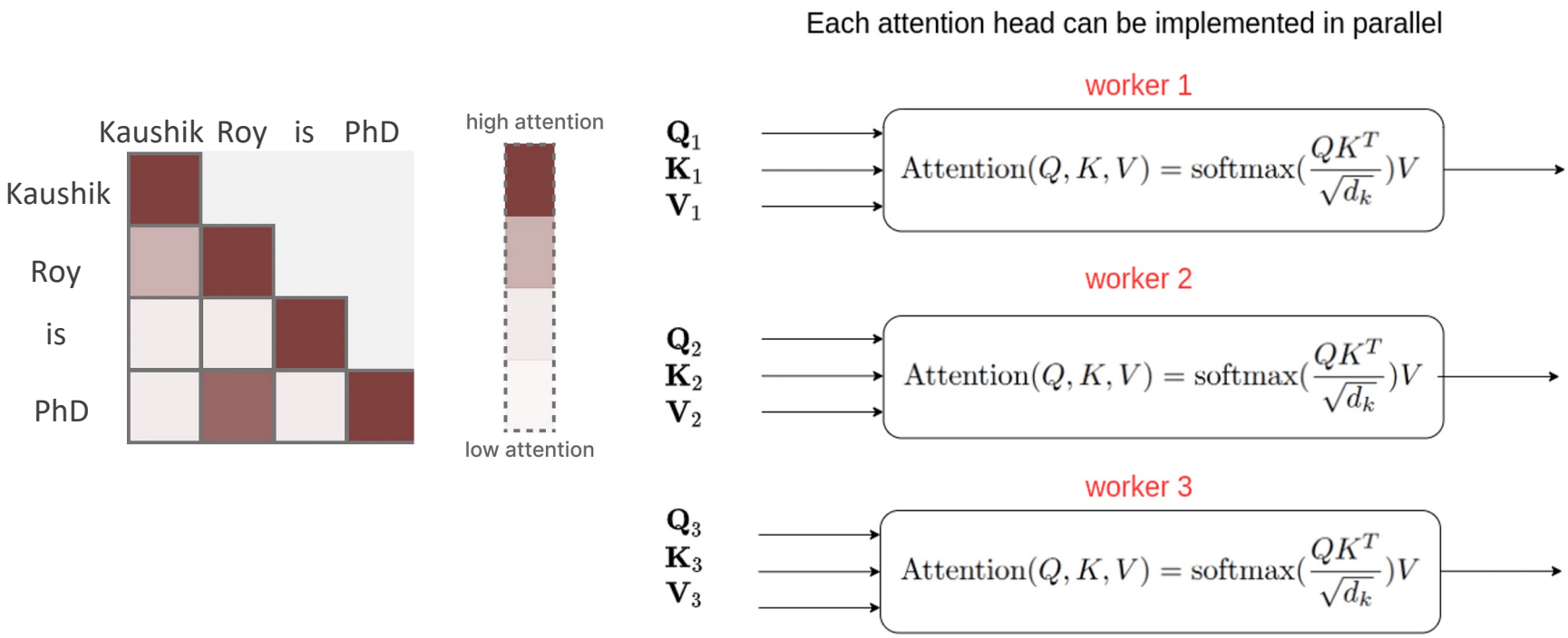


Credit and details: **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**
Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, 2018

BERT and family



BERT and family - Multi-headed Self-Attention



Self Attention Snippet and Live Coding - BERT from Scratch

```python
class SelfAttention(nn.Module):
    def __init__(self, embed_size, heads):
        super(SelfAttention, self).__init__()
        self.embed_size = embed_size
        self.heads = heads
        self.head_dim = embed_size // heads

        assert (
            self.head_dim * heads == embed_size
        ), "Embedding size needs to be divisible by heads"

        self.values = nn.Linear(self.head_dim, embed_size, bias=False)
        self.keys = nn.Linear(self.head_dim, embed_size, bias=False)
        self.queries = nn.Linear(self.head_dim, embed_size, bias=False)
        self.fc_out = nn.Linear(embed_size, embed_size)
```

BERT-based CV Processing

Live Coding - BERT using Libraries

Github: https://github.com/kauroy1994/Teaching (BERT-based CV Processing)

- Use the transformers library to extract information from the CV

```python
import pdfplumber
from transformers import AutoTokenizer, AutoModelForTokenClassification
from transformers import pipeline

# Step 1: Load the CV PDF and extract text
def extract_text_from_pdf(pdf_path):
    with pdfplumber.open(pdf_path) as pdf:
        pages = [page.extract_text() for page in pdf.pages]
    return ''.join(pages)

# Extract text from the CV
pdf_path = "CV.pdf"
cv_text = extract_text_from_pdf(pdf_path)

# Step 2: Load pre-trained BERT model and tokenizer
tokenizer = AutoTokenizer.from_pretrained("dslim/bert-base-NER")
model = AutoModelForTokenClassification.from_pretrained("dslim/bert-base-NER")

# Step 3: Use pipeline for Named Entity Recognition
nlp = pipeline("ner", model=model, tokenizer=tokenizer, grouped_entities=True)

# Step 4: Extract entities from the CV text
ner_results = nlp(cv_text)

# Display the recognized entities
for entity in ner_results:
    print(f"Entity: {entity['word']}, Label: {entity['entity_group']}")

# Step 5: Post-process the entities for CV data extraction (optional)
# For example, grouping entities like degree, institution, and dates
def extract_education_details(ner_results):
    education = []
    current_education = {}
    for entity in ner_results:
        if entity['entity_group'] == 'ORG':
            current_education['institution'] = entity['word']
        elif entity['entity_group'] == 'MISC':  # Assuming degrees are labeled as MISC
            current_education['degree'] = entity['word']
        elif entity['entity_group'] == 'DATE':
            current_education['year'] = entity['word']

        # Save the current education entry
        if 'institution' in current_education and 'degree' in current_education and 'year' in current_education:
            education.append(current_education)
            current_education = {}
    return education

education_details = extract_education_details(ner_results)

# Display the structured education data
print("Extracted Education Details:", education_details)
```

Using BERT in Practice – Huggingface Libraries

- Transformers – https://github.com/huggingface/transformers
- APIs to download and use pre-trained models, fine-tune them on own datasets and tasks
  - Code Sample

```
# Loading BERT
model_class, tokenizer_class, pretrained_weights =  (ppb.DistilBertModel,
ppb.DistilBertTokenizer, 'distilbert-base-uncased')

# Load pretrained model/tokenizer
tokenizer = tokenizer_class.from_pretrained(pretrained_weights)
model = model_class.from_pretrained(pretrained_weights)
```

- Provides pretrained models in 100+ languages.
- Use with popular deep learning libraries, PyTorch and TensorFlow,
  - Possible to train / fine-tune models with one, and load it for inference with another

Using BERT in Practice – Huggingface Libraries

- DistilBERT
  - Details: https://medium.com/huggingface/distilbert-8cf3380435b5
  - Teacher-student learning, also called model distillation
    - Teacher: bert-base-uncased
    - Student: dstilBERT - BERT without *the token-type embeddings and the pooler* , and half the layers
  - **"DistilBERT,** has **about half** the total number of parameters of BERT base and retains 95% of BERT's performances on the language understanding benchmark GLUE"

- Sample code of usage for sentiment classification: https://github.com/biplav-s/course-nl/blob/master/l12-langmodel/UsingLanguageModel.ipynb

Example Pre-Trained Models

1. ALBERT (from Google Research and the Toyota Technological Institute at Chicago) released with the paper ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, by Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut.
2. BART (from Facebook) released with the paper BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension by Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov and Luke Zettlemoyer.
3. BERT (from Google) released with the paper BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding by Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova.
4. BERT For Sequence Generation (from Google) released with the paper Leveraging Pre-trained Checkpoints for Sequence Generation Tasks by Sascha Rothe, Shashi Narayan, Aliaksei Severyn.
5. CamemBERT (from Inria/Facebook/Sorbonne) released with the paper CamemBERT: a Tasty French Language Model by Louis Martin*, Benjamin Muller*, Pedro Javier Ortiz Suárez*, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah and Benoît Sagot.
6. CTRL (from Salesforce) released with the paper CTRL: A Conditional Transformer Language Model for Controllable Generation by Nitish Shirish Keskar*, Bryan McCann*, Lav R. Varshney, Caiming Xiong and Richard Socher.
7. DeBERTa (from Microsoft Research) released with the paper DeBERTa: Decoding-enhanced BERT with Disentangled Attention by Pengcheng He, Xiaodong Liu, Jianfeng Gao, Weizhu Chen.
8. DialoGPT (from Microsoft Research) released with the paper DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation by Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, Bill Dolan.
9. DistilBERT (from HuggingFace), released together with the paper DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter by Victor Sanh, Lysandre Debut and Thomas Wolf. The same method has been applied to compress GPT2 into DistilGPT2, RoBERTa into DistilRoBERTa, Multilingual BERT into DistilmBERT and a German version of DistilBERT.
10. DPR (from Facebook) released with the paper Dense Passage Retrieval for Open-Domain Question Answering by Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih.
11. ELECTRA (from Google Research/Stanford University) released with the paper ELECTRA: Pre-training text encoders as discriminators rather than generators by Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning.
12. FlauBERT (from CNRS) released with the paper FlauBERT: Unsupervised Language Model Pre-training for French by Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, Didier Schwab.
13. Funnel Transformer (from CMU/Google Brain) released with the paper Funnel-Transformer: Filtering out Sequential Redundancy for Efficient Language Processing by Zihang Dai, Guokun Lai, Yiming Yang, Quoc V. Le.
14. GPT (from OpenAI) released with the paper Improving Language Understanding by Generative Pre-Training by Alec Radford, Karthik Narasimhan, Tim Salimans and Ilya Sutskever.
15. GPT-2 (from OpenAI) released with the paper Language Models are Unsupervised Multitask Learners by Alec Radford*, Jeffrey Wu*, Rewon Child, David Luan, Dario Amodei** and Ilya Sutskever**.
16. LayoutLM (from Microsoft Research Asia) released with the paper LayoutLM: Pre-training of Text and Layout for Document Image Understanding by Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou.
17. Longformer (from AllenAI) released with the paper Longformer: The Long-Document Transformer by Iz Beltagy, Matthew E. Peters, Arman Cohan.
18. LXMERT (from UNC Chapel Hill) released with the paper LXMERT: Learning Cross-Modality Encoder Representations from Transformers for Open-Domain Question Answering by Hao Tan and Mohit Bansal.
19. MarianMT Machine translation models trained using OPUS data by Jörg Tiedemann. The Marian Framework is being developed by the Microsoft Translator Team.
20. MBart (from Facebook) released with the paper Multilingual Denoising Pre-training for Neural Machine Translation by Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, Luke Zettlemoyer.
21. MMBT (from Facebook), released together with the paper a Supervised Multimodal Bitransformers for Classifying Images and Text by Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Davide Testuggine.
22. Pegasus (from Google) released with the paper PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization> by Jingqing Zhang, Yao Zhao, Mohammad Saleh and Peter J. Liu.
23. Reformer (from Google Research) released with the paper Reformer: The Efficient Transformer by Nikita Kitaev, Łukasz Kaiser, Anselm Levskaya.
24. RoBERTa (from Facebook), released together with the paper a Robustly Optimized BERT Pretraining Approach by Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. ultilingual BERT into DistilmBERT and a German version of DistilBERT.
25. SqueezeBert released with the paper SqueezeBERT: What can computer vision teach NLP about efficient neural networks? by Forrest N. Iandola, Albert E. Shaw, Ravi Krishna, and Kurt W. Keutzer.
26. T5 (from Google AI) released with the paper Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer by Colin Raffel and Noam Shazeer and Adam Roberts and Katherine Lee and Sharan Narang and Michael Matena and Yanqi Zhou and Wei Li and Peter J. Liu.
27. Transformer-XL (from Google/CMU) released with the paper Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context by Zihang Dai*, Zhilin Yang*, Yiming Yang, Jaime Carbonell, Quoc V. Le, Ruslan Salakhutdinov.
28. XLM (from Facebook) released together with the paper Cross-lingual Language Model Pretraining by Guillaume Lample and Alexis Conneau.
29. XLM-RoBERTa (from Facebook AI), released together with the paper Unsupervised Cross-lingual Representation Learning at Scale by Alexis Conneau*, Kartikay Khandelwal*, Naman Goyal, Vishrav Chaudhary*, Guillaume Wenzek*, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer and Veselin Stoyanov.
30. XLNet (from Google/CMU) released with the paper XLNet: Generalized Autoregressive Pretraining for Language Understanding by Zhilin Yang*, Zihang Dai*, Yiming Yang, Jaime Carbonell, Ruslan

https://github.com/huggingface/transformers

Concluding Segment

Concluding Comments

- We saw how model's are constructed and trained using a convenience library called pytorch
- We saw how BERT can be used in two ways for two CV-related tasks
- What about abstract mathematical objects?

TextGrad: Automatic ''Differentiation'' via Text

An autograd engine -- for textual gradients!

TextGrad is a powerful framework building automatic ``differentiation'' via text. TextGrad implements backpropagation through text feedback provided by LLMs, strongly building on the gradient metaphor

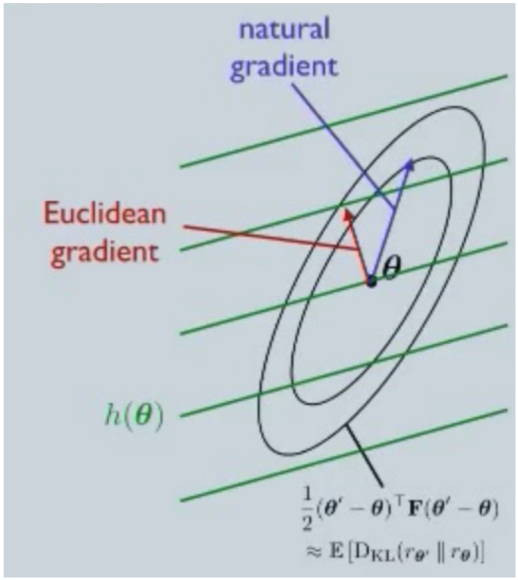
1 Analogy in abstractions

	Math	PyTorch	TextGrad
Input	x	`Tensor(image)`	`tg.Variable(article)`
Model	$\hat{y} = f_\theta(x)$	`ResNet50()`	`tg.BlackboxLLM("You are a summarizer.")`
Loss	$L(y, \hat{y}) = \sum_i y_i \log(\hat{y}_i)$	`CrossEntropyLoss()`	`tg.TextLoss("Rate the summary.")`
Optimizer	$\mathrm{GD}(\theta, \frac{\partial L}{\partial \theta}) = \theta - \frac{\partial L}{\partial \theta}$	`SGD(list(model.parameters()))`	`tg.TGD(list(model.parameters()))`

2 Automatic differentiation

PyTorch and TextGrad share the same syntax for backpropagation and optimization.

Forward pass	Backward pass	Updating variable
`loss = loss_fn(model(input))`	`loss.backward()`	`optimizer.step()`



Course Project

Discussion: Course Project

Theme: Analyze quality of official information available for elections in 2024 [in a state]

- Take information available from
  - Official site: State Election Commissions
  - Respected non-profits: League of Women Voters
- Analyze information
  - State-level: Analyze quality of questions, answers, answers-to-questions
  - Comparatively: above along all states (being done by students)
- Benchmark and report
  - Compare analysis with LLM
  - Prepare report

- Process and analyze using NLP
  - Extract entities
  - Assess quality – metrics
    - Content – *Englishness*
    - Content – *Domain* -- election
  - … other NLP tasks
  - Analyze and communicate overall

Major dates for project check

- Sep 10: written – project outline
- Oct 8: in class
- Oct 31: in class // LLM
- Dec 5: in class // Comparative

About Next Lecture – Lecture 12

Lecture 12 Outline

- Mamba, Finetuning

7	Sep 10 (Tu)	Statistical parsing, QUIZ
8	Sep 12 (Th)	Evaluation, Semantics
9	Sep 17 (Tu)	Semantics, Machine Learning for NLP, Evaluation - Metrics
10	Sep 19 (Th)	Towards Language Model: Vector embeddings, Embeddings, CNN/ RNN
11	Sep 24 (Tu)	Language Model – PyTorch, BERT, {Resume data, two tasks} **– Guest Lecture**
12	Sep 26 (Th)	Language Model – Finetuning, Mamba - **Guest Lecture**
13	Oct 1 (Tu)	Language model – comparing arch, finetuning - **Guest Lecture**
14	Oct 3 (Th)	Language model – comparison of results, discussion, ongoing trends– **Guest Lecture**
15	Oct 8 (Tu)	PROJ REVIEW