Natural Language Understanding, Generation, and Machine Translation

Lecture 11: Large Pretrained Models and Prompting

Shay Cohen
based on slides by Alexandra Birch

7 February 2024 (week 4)

School of Informatics
University of Edinburgh
scohen@inf.ed.ac.uk

Paradigm Shift: Pre-train fine-tune

Classification task: $p(y|x)$

- Traditional: hand-crafted features to represent x, and then apply machine learning
- Deep learning: learn latent features of x
- Idea: learn a **generic** latent feature once, and then share it across all NLP tasks.
- Language modelling is such a generic task:

$$p(x_i|x_0, x_1, \ldots, x_{i-1})$$

- Abundant amount of naturally occuring text

Refresher

ELMo

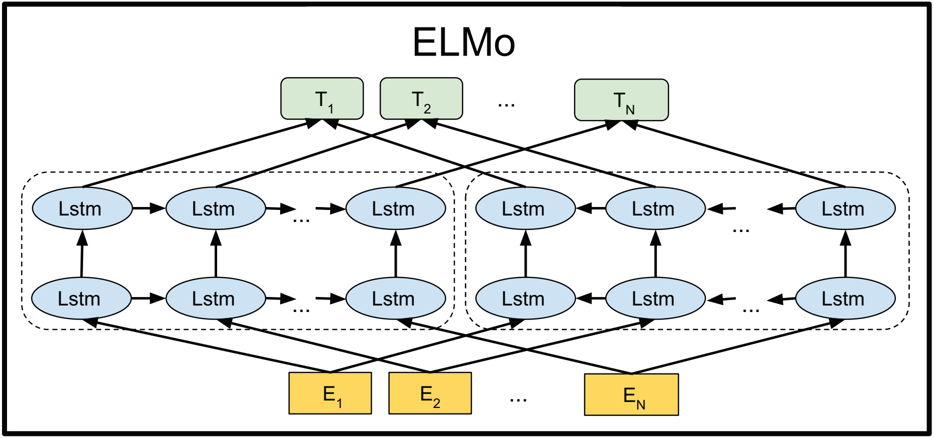


Figure from [Devlin et al., 2019]

Deep contextualized word representations [Peters et al., 2018]

Transformer

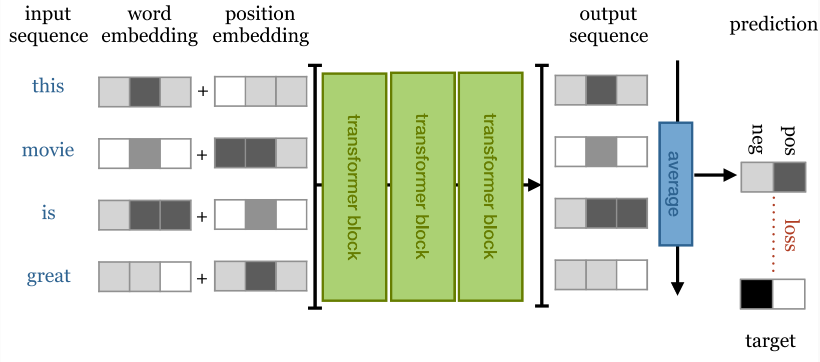


Figure from [Bloem, 2019]

BERT vs. GPT

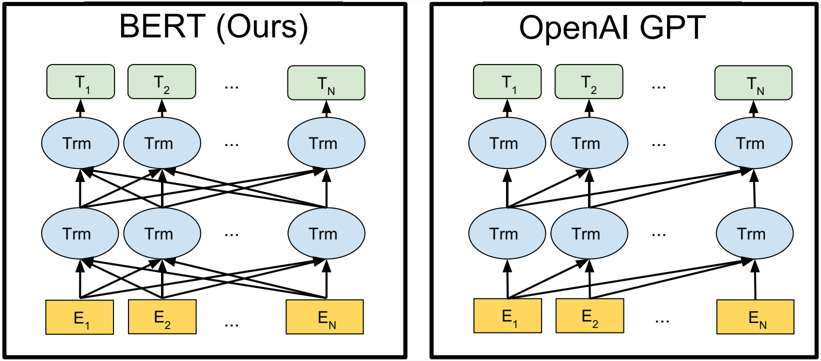


Figure from [Devlin et al., 2019].

Pre-training and Finetuning Bert

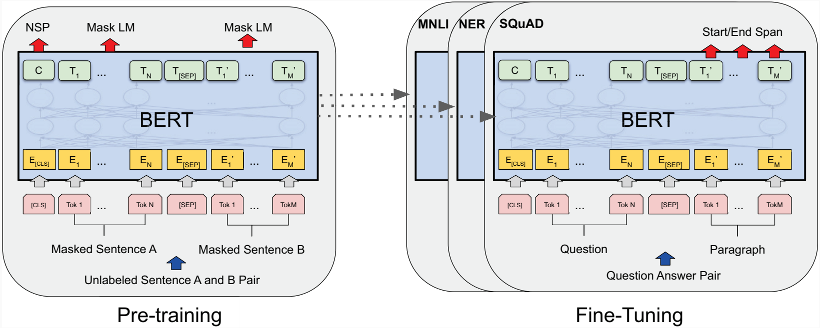


Figure from [Devlin et al., 2019].

Pre-training

Pre-trained Language Models

- Three main dimensions across they vary:
    - Objective Functions (main and auxiliary)
    - Noising Functions
    - Directionality
- Examples of Model Architectures

Main Training Objectives

- Standard Language Model Objective:

  Task is to predict: $P(x_i | x_0 \dots x_{i-1})$

- Denoising Objective:

  Noising function: $\tilde{x} = f_{noise}(x)$

  Task is to predict: $P(x | \tilde{x})$

  - Corrupted Text Reconstruction: loss over noised part
  - Full Text Reconstruction: loss over entire input

Auxiliary Objectives

- Can apply multiple learning objectives
- Learn specific things about language which will be useful in the downstream tasks eg.
  - Next Sentence Prediction: do two segments appear consecutively - better sentence representations BERT [Devlin et al., 2019]
  - Discourse Relation Prediction: predict rhetorical relations between sentences - better semantics ERNIE [Sun et al., 2020]
  - Image Region Prediction: predict the masked regions of an image - for better visual-linguistic tasks VL-BERT [Su et al., 2020]

Noising Functions

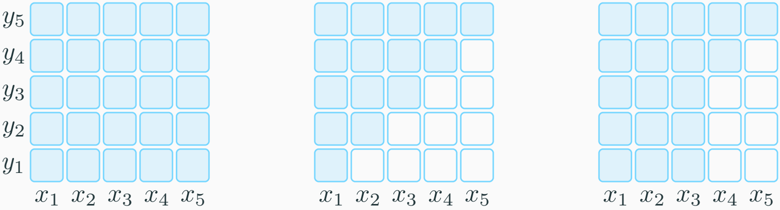
For training objectives based on reconstruction apply noise either over tokens (sub-words), whole words or spans

Original Text: Biden approved measures .

Operation	Corrupted Text
Mask	Biden **[MASK]** measures .
Replace	Biden **ate** measures .
Delete	Biden ~~approved~~ measures .
Permute	approved measures . Biden

Directionality

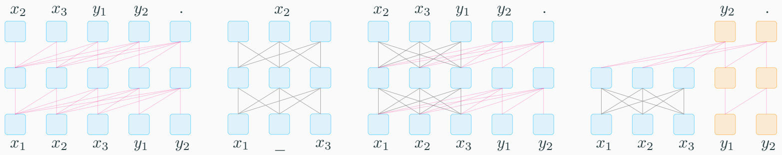
- Bidirectional: full attention no masking
- Left-to-right: diagonal attention masking
- Mix the two strategies



(a) Full. (b) Diagonal. (c) Mixture.

From [Liu et al., 2021a]

Paradigms



(a) Left-to-right LM. (b) Masked LM. (c) Prefix LM. (d) Encoder-Decoder.

From [Liu et al., 2021a]

Model	Arch	PreTrainObj	AuxObj	Mask/Repl/Del/Perm	Applic.
GPT-2/3	L2R	LM	-	-	NLU/NLG
BERT	Mask	CorruptText	NSP	Tok/-/-/-	NLU
	Enc only				
BART	Enc-Dec	FullText	-	Tok/Span/Tok/Sent	NLU/NLG
T5	Enc-Dec	CorruptText	-	-/Span/-/-	NLU/NLG

T5: Text-to-Text Transfer Transformer

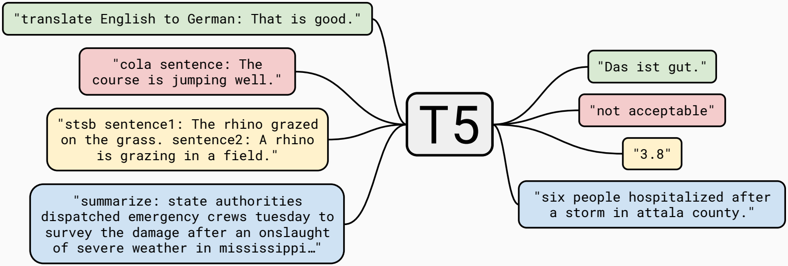


Figure from [Raffel et al., 2020]

T5

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [Raffel et al., 2020]

- Model Size: up to 11B parameters - BERT-large is 330M
- Amount of training data: 120B words of data
- Domain/Cleanness of training data
- Model used is almost the same as the original encoder-decoder model in the Vaswani et al. paper
- Conclusion: Scaling up model size and training data really helps

Really easy to use pretrained model for multiple tasks using prompts!

Common Crawl in T5

The Common Crawl produces 20TB of data every month, but much of it is not usable

Raffel et al. clean the common crawl to get "Colossal Clean Crawled Corpus" (C4):

- Discard pages with fewer than 3 sentences
- Filtered out pages with bad words
- Removed any pages with curly brackets
- Removed citation markers
- Removed boilerplate text

A few more other tricks...

T5 pre-training

- Uses span corruption pre-training objective
- Masking (encoder): "The teacher continued the lecture." $\rightarrow$ "[X] continued [Y]"
- Output (decoder): "[X] The teacher [Y] the lecture [EOS]"
- X and Y are sentinel tokens
- (there is a maximum on the number of words that can be masked)

Finetuning T5

- "This book is a fun read" $\rightarrow$ positive
- We finetune T5 for the decoder to generate the label text
- This stands in contrast to BERT, where we have a fixed set of classes chosen through a softmax
- This finetuning is what allows to work with text for specific tasks

T5 Tasks

Cast all the tasks considered (GLUE / SuperGLUE and others) into text-to-text format

Model	SQuAD EM	SQuAD F1	SuperGLUE Average
Previous best	90.1[a]	95.5[a]	84.6[d]
T5-Small	79.10	87.24	63.3
T5-Base	85.44	92.08	76.2
T5-Large	86.66	93.79	82.3
T5-3B	88.53	94.95	86.4
T5-11B	**91.26**	**96.22**	**88.9**

Example data: CoLA

Original input:

Sentence: `John made Bill master of himself.`

Processed input: `cola sentence: John made Bill master of himself.`

Original target: `1`

Processed target: `acceptable`

Example data: RTE

Original input:

Sentence 1: `A smaller proportion of Yugoslavia's Italians were settled in Slovenia (at the 1991 national census, some 3000 inhabitants of Slovenia declared themselves as ethnic Italians).`

Sentence 2: `Slovenia has 3,000 inhabitants.`

Processed input: `rte sentence1: A smaller proportion of Yugoslavia's Italians were settled in Slovenia (at the 1991 national census, some 3000 inhabitants of Slovenia declared themselves as ethnic Italians). sentence2: Slovenia has 3,000 inhabitants.`

Original target: `1`

Example data: MNLI

Original input:

Hypothesis: The St. Louis Cardinals have always won.

Premise: yeah well losing is i mean i'm i'm originally from Saint Louis and Saint Louis Cardinals when they were there were uh a mostly a losing team but

Processed input: mnli hypothesis: The St. Louis Cardinals have always won. premise: yeah well losing is i mean i'm i'm originally from Saint Louis and Saint Louis Cardinals when they were there were uh a mostly a losing team but

Original target: 2

Processed target: contradiction

Example data: Machine Translation

Original input: "Luigi often said to me that he never wanted the brothers to end up in court," she wrote.

Processed input: translate English to German: "Luigi often said to me that he never wanted the brothers to end up in court," she wrote.

Original target: "Luigi sagte oft zu mir, dass er nie wollte, dass die Brüder vor Gericht landen", schrieb sie.

Processed target: "Luigi sagte oft zu mir, dass er nie wollte, dass die Brüder vor Gericht landen", schrieb sie.

Types of finetuning in T5

- Adapter layers
- Gradual unfreezing - start by finetuning only the top layers, and slowly start introducing changes to the lower layers

Finetuning Results

Fine-tuning method	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ All parameters	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Adapter layers, $d = 32$	80.52	15.08	79.32	60.40	13.84	17.88	15.54
Adapter layers, $d = 128$	81.51	16.62	79.47	63.03	19.83	27.50	22.63
Adapter layers, $d = 512$	81.54	17.78	79.18	64.30	23.45	33.98	25.81
Adapter layers, $d = 2048$	81.51	16.62	79.47	63.03	19.83	27.50	22.63
Gradual unfreezing	82.50	18.95	79.17	**70.79**	26.71	39.02	26.93

(from Raffel et al.)

Prompting

Alternative Paradigm: Prompting

- Proposed by: Language models are unsupervised multitask learners [Radford et al., 2019]
- **Pretrain - Fine tune**: adapting LMs (objectives eg. MLM or next sentence prediction) to downstream tasks
- **Pretrain - Prompt**: adapting downstream tasks to LMs.
- Appeal: zero-shot capabilities and strong few-shot performance

Goal of Prompting

- Supervised learning:
  - Model $P(y \mid x, \theta)$
  - Input $x =$ "I love this book"
  - Output $y = ++$ out of label set $\mathcal{Y} = ++, +, \ , -, --$
- Prompting:
  - Instead use a language model and a text-to-text query to predict **y**
  - Reducing the need for large supervised datasets

1. Adding a Prompt

1. Choose a *prompting function* or template $f_{prompt}(\cdot)$
2. Apply it to the *input* $\mathbf{x}$

To create a *prompt* $\mathbf{x}' = f_{prompt}(\mathbf{x})$

Name	Notation	Example
Input	$\mathbf{x}$	I love this book
Prompting Function	$f_{prompt}(\mathbf{x})$	[X] Overall, it was a [Z] book
Prompt	$\mathbf{x}'$	I love this book. Overall, it was a [Z] book

eg. of a *Cloze prompt*

Prefix prompt: [X] TLDR; [Z]

2. Answer Search

- $\mathcal{Z}$ set of permissible values for $\mathbf{z}$: answers
- $\mathcal{Z}$ could be any item in vocabulary or restricted to subset of values
- $\hat{z} = \underset{\mathbf{z} \in \mathcal{Z}}{search}\ P(f_{fill}(\mathbf{x}', \mathbf{z}); \theta)$, where $P(., \theta)$ is the pretrained LM

Name	Notation	Example
Answers	$\mathcal{Z}$	"excellent","good","OK","bad","horrible"
Output	$\mathcal{Y}$	$++,+,\sim,-,--$
Filled Prompt	$f_{fill}(\mathbf{x}', \mathbf{z}))$	I love this book. Overall it was a bad book
Answered Prompt	$f_{fill}(\mathbf{x}', \mathbf{z}^*))$	I love this book. Overall it was a good book

3. Mapping Answer to Desired Output

- Highest scoring answer $\hat{z}$ to highest scoring output $\hat{y}$
- "excellent", "fabulous" and "wonderful" $\rightarrow$ "++"

Examples

Task	Input [X]	Template	Answer [Z]
Sentiment	I love this movie.	[X] The movie is [Z].	great fantastic ...
Topics	He slammed the ball.	[X] The text is about [Z].	sports science ...
Intention	What is the taxi fare?	[X] The question is about [Z].	price city ...
NER	[X1]: Mike went to Paris. [X2]: Paris	[X1][X2] is a [Z] entity.	org loc ...
Summary	Las Vegas police...	[X] TL;DR: [Z]	The victim... A woman ...
Translation	Je vous aime	French: [X] English: [Z]	I love you. I fancy you. ...

Adapted from [Liu et al., 2021b]

Prompt Engineering

- Creating a promtping function $f_{prompt}(\mathbf{x})$
- Manual template engineering
- Automated template learning of discrete prompts:
  - Prompt mining "[X] middle words [Z]"
  - Paraphrase existing prompts - select the ones with highest accuracy
- Continuous prompts: perform prompting directly in the embedding space of the model
  - Initialise with discrete prompt, fine tune on task
  - Template embeddings have their own parameters that can be tuned

Prompt Tuning (Lester et al., 2021)

A way of using continuous prompting by prepending the input text with some embeddings that are tailored to a specific task [Lester et al., 2021]

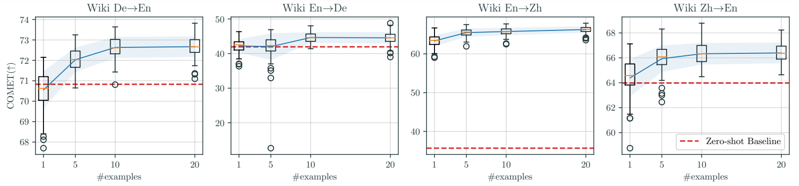
- Want to retain the performance of finetuning without having to change all parameters
- Input to LM: "p_1 p_2 p_3 **This book is really fun to read**"
- p_i are now embeddings that indicate the task of sentiment analysis
- When finetuning, change p_i based on backpropagation while freezing the model

Prompting for Zero-shot Machine Translation

ID	Template (in English)	English		German		Chinese	
		w/o	w/	w/o	w/	w/o	w/
A	`[src]: [input]` ◇ `[tgt]:`	**38.78**	**31.17**	-26.15	-16.48	**14.82**	**-1.08**
B	`[input]` ◇ `[tgt]:`	-88.62	-85.35	-135.97	-99.65	-66.55	-85.84
C	`[input]` ◇ Translate to `[tgt]:`	-87.63	-68.75	-106.30	-73.23	-63.38	-70.91
D	`[input]` ◇ Translate from `[src]` to `[tgt]:`	-113.80	-89.16	-153.80	-130.65	-76.79	-67.71
E	`[src]: [input]` ◇ Translate to `[tgt]:`	20.81	16.69	**-24.33**	**-5.68**	-8.61	-30.38
F	`[src]: [input]` ◇ Translate from `[src]` to `[tgt]:`	-27.14	-6.88	-34.36	-9.22	-32.22	-44.95

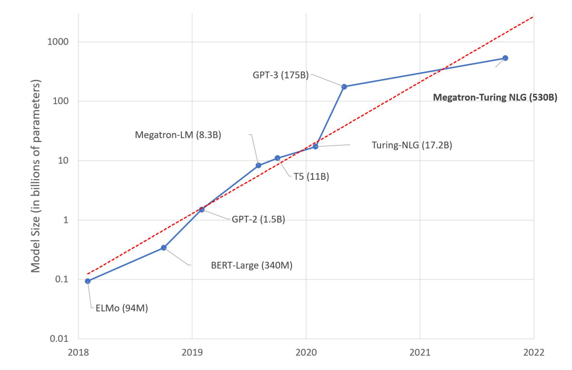
Results from [Zhang et al., 2023]

Demonstrations for Few-shot Machine Translation



Results from [Zhang et al., 2023]

Challenge: Size of Models



from Julien Simon `https://huggingface.co/blog/large-language-models`

- Cost of training: $ and carbon footprint
- Difficulty in deploying these systems
- Downsizing with: knowledge distillation, model pruning, quantization

In-context Learning

- For a given task, show some examples with their output to the language model
- Then, ask it to solve the problem on new examples
- Highly related to prompting

Summary

- New paradigm in ML: Pretraining + Finetuning
- Axes: Objective functions, noising functions, directionality
- Alternative paradigm: Pretrain + Prompt
- Good zero-shot, few-shot performance
- Prompt engineering
- Challenges

Next: Look at evaluation of natural language processing models, and in particular evaluation of machine translation.

References

Bloem, P. (2019).
Transformers from Scratch.
http://www.peterbloem.nl/blog/transformers.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019).
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
In In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics.

Lester, B., Al-Rfou, R., and Constant, N. (2021).
The power of scale for parameter-efficient prompt tuning.
arXiv preprint arXiv:2104.08691.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2021a).
Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.
CoRR, abs/2107.13586.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2021b).
Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.
arXiv preprint arXiv:2107.13586.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018).
Deep contextualized word representations.
In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2227–2237.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019).
Language models are unsupervised multitask learners.
OpenAI blog, 1(8):9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020).
Exploring the limits of transfer learning with a unified text-to-text transformer.
Journal of Machine Learning Research, 21:1–67.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. (2020).
Vl-bert: Pre-training of generic visual-linguistic representations.
In International Conference on Learning Representations.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., and Wang, H. (2020).
Ernie 2.0: A continual pre-training framework for language understanding.
Proceedings of the AAAI Conference on Artificial Intelligence, 34(05):8968–8975.

Zhang, B., Haddow, B., and Birch, A. (2023).
Prompting large language model for machine translation: A case study.