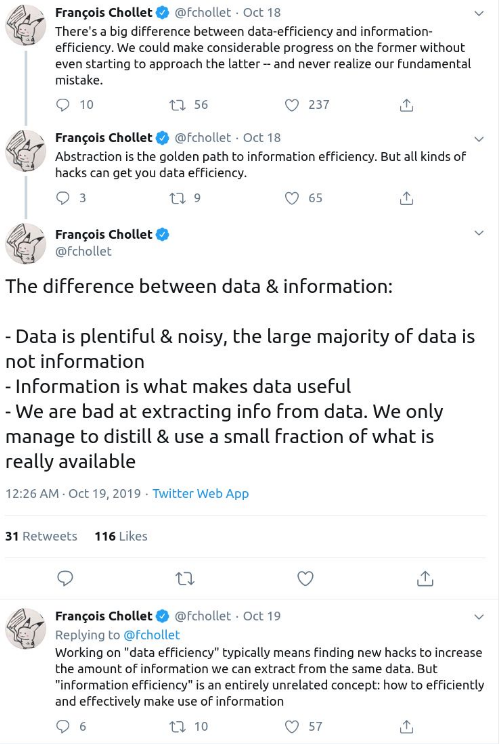
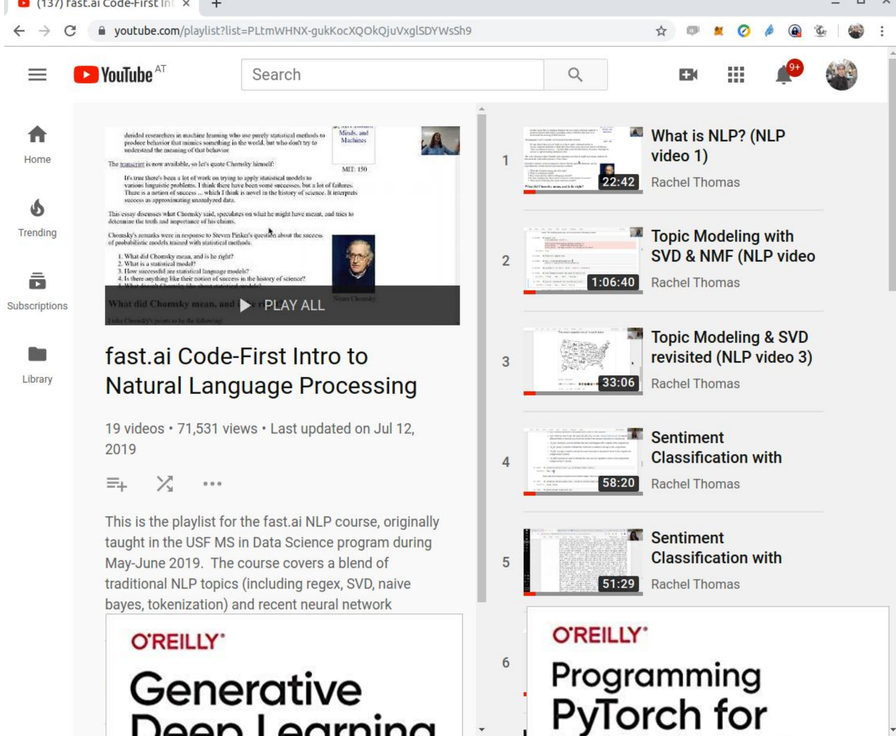
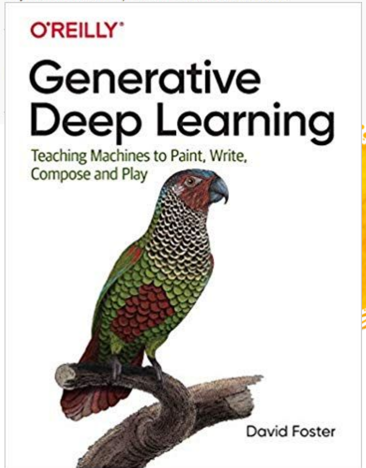
NLP News Oct 2019

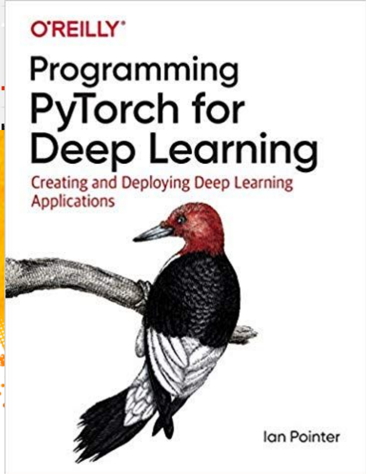
Adrian M.P. Brasoveanu, MT

Introduction



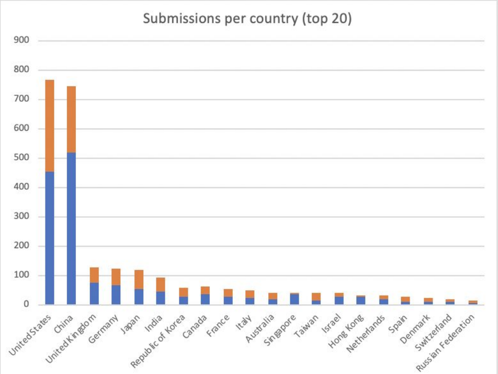


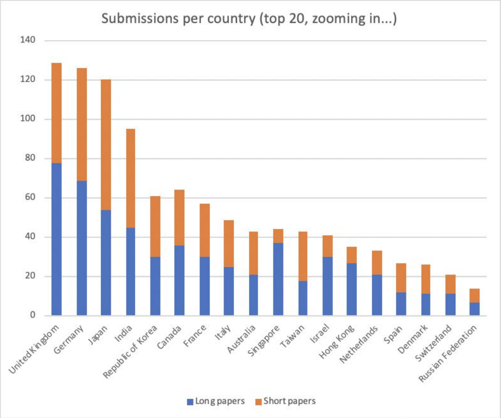


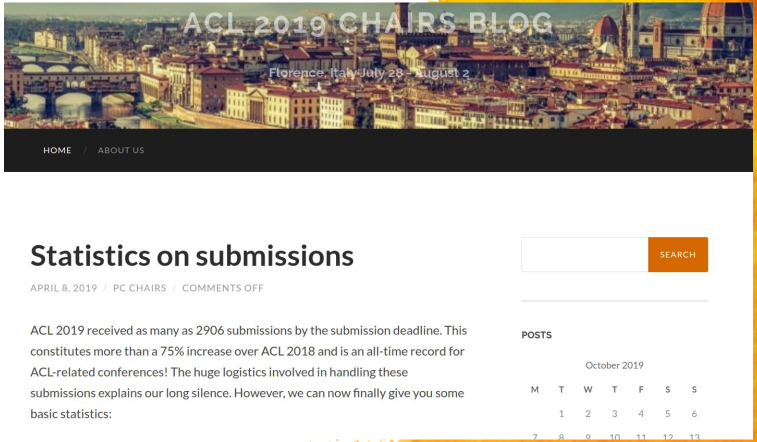


ACL 2019 Stats

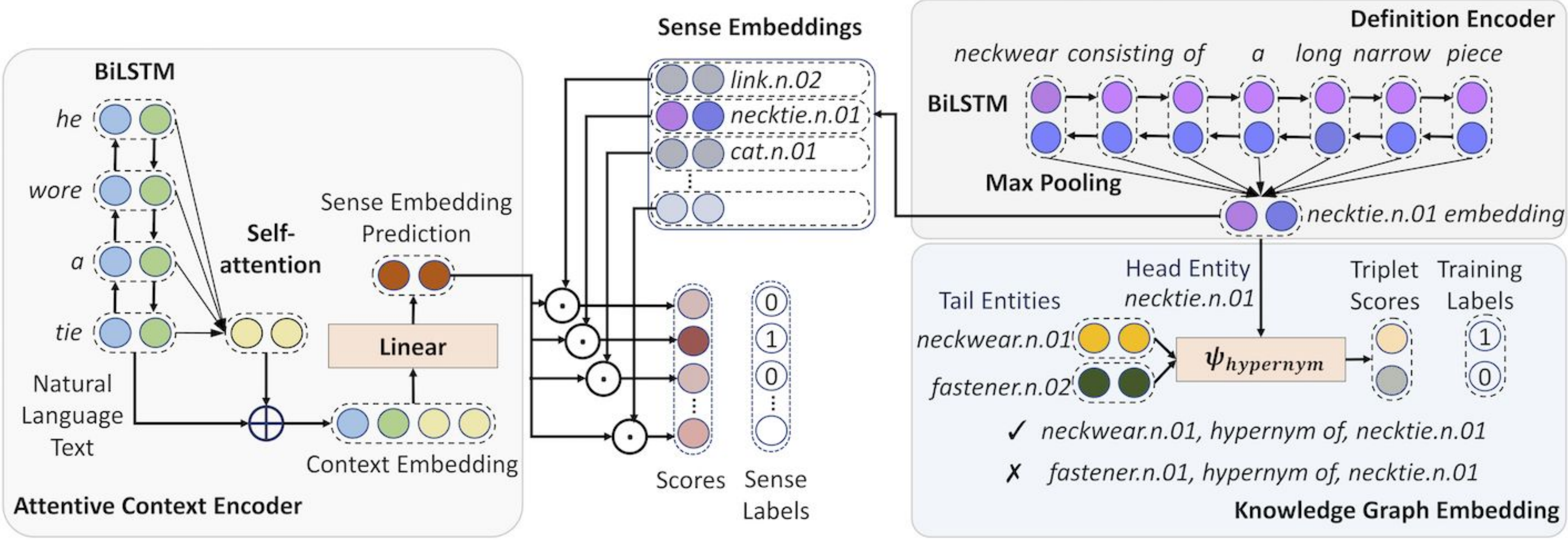
	Area	Long	Short	Total
1.	Information Extraction, Text Mining	156	93	249
2.	Machine Learning	148	73	221
3.	Machine Translation	102	105	207
4.	Dialogue and Interactive Systems	125	57	182
5.	Generation	97	58	155
6.	Question Answering	99	55	154
7.	Sentiment Analysis, Argument Mining	91	60	151
8.	Word-level Semantics	78	59	137
9.	Applications	65	72	137
10.	Resources and Evaluation	70	60	130
11.	Multidisciplinary, AC COI	70	44	114
12.	Sentence-level Semantics	70	42	112
13.	Tagging, Chunking, Syntax, Parsing	50	49	99
14.	Social Media	51	42	93
15.	Summarization	48	35	83
16.	Document Analysis	48	33	81
17.	Vision, Robotics Multimodal Grounding, Speech	56	23	79
18.	Multilinguality	43	32	75
19.	Textual Inference, Other Areas of Semantics	44	30	74
20.	Linguistic Theories, Cognitive, Psycholinguistics	39	21	60
21.	Discourse and Pragmatics	33	24	57
22.	Phonology, Morphology, Word Segmentation	26	18	44
	Total	**1609**	**1085**	**2694**







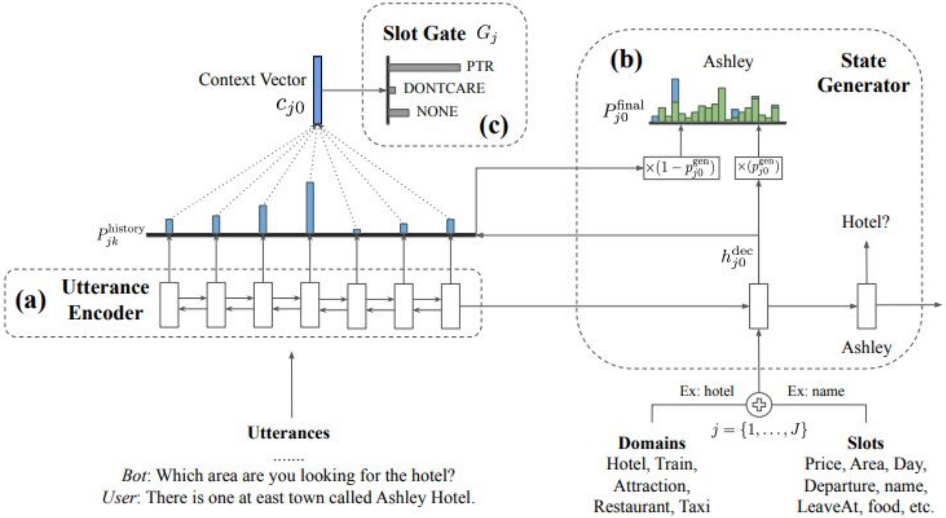
Zero-shot WSD with Sense Definition Embeddings



https://www.aclweb.org/anthology/P19-1568.pdf and https://github.com/malllabiisc/EWISE

Transferable Multi-Domain State Generator for TODS

- **Model**
  - utterance encoder
  - slot gate
  - state generator
  - all shared across domains
- **Knowledge Transfer for predicting**
  - **domain, slot, value**
- Madotto et al, ACL 2019
  - https://arxiv.org/pdf/1905.08743.pdf
  - https://github.com/jasonwu0731/trade-dst



Energy and Policy Considerations for NLP

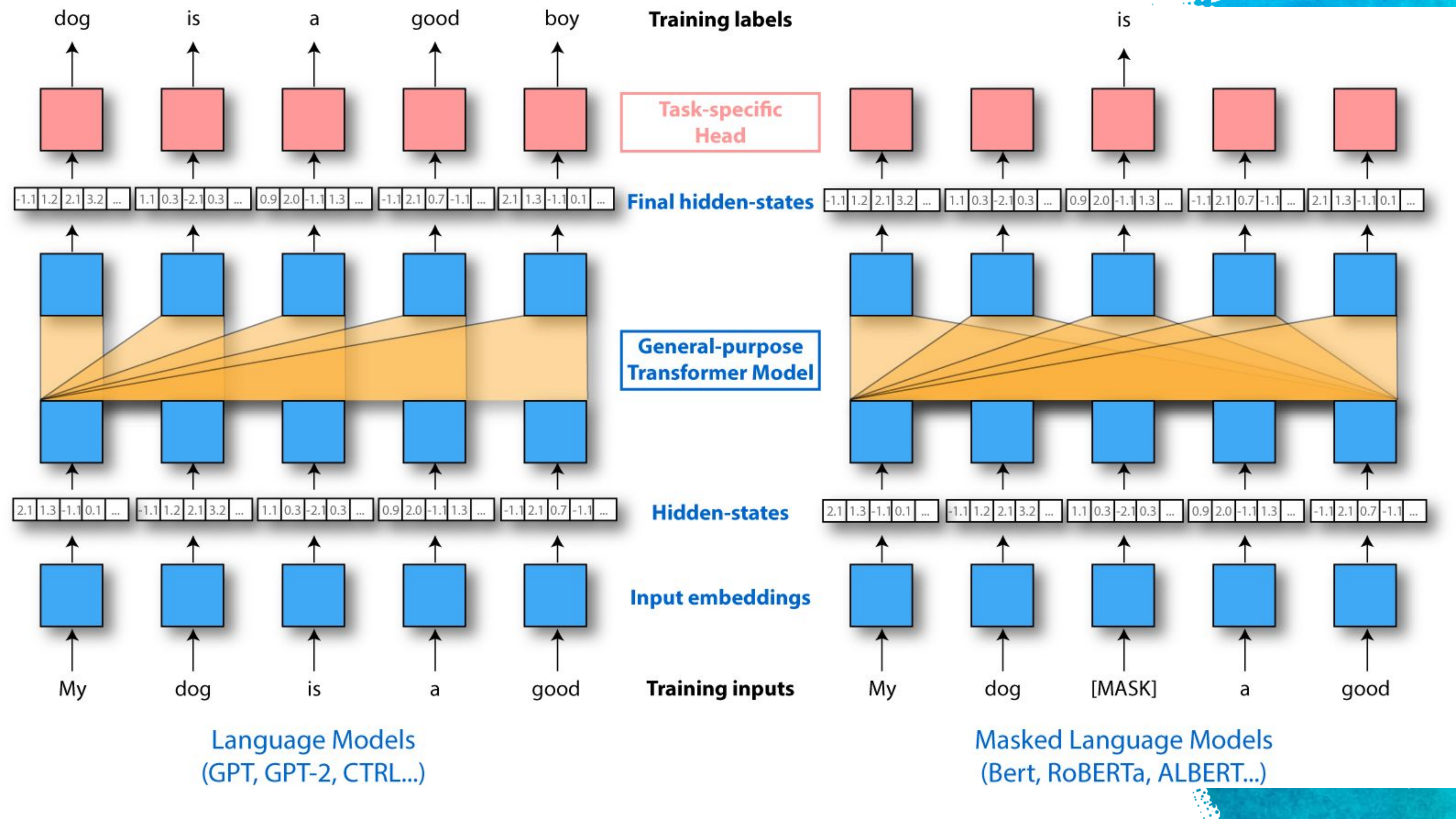
- Cost of hardware
- Cost of electricity
- Cloud compute time
- Carbon footprint
- Cost of training
- Cost of development
- Emma Strubell et al
  - ACL 2019 paper
  - https://arxiv.org/pdf/1906.02243.pdf

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
$\text{Transformer}_{base}$	P100x8	1415.78	12	27	26	\$41–\$140
Transformer_{big}	P100x8	1515.43	84	201	192	\$289–\$981
ELMo	P100x3	517.66	336	275	262	\$433–\$1472
BERT_{base}	V100x64	12,041.51	79	1507	1438	\$3751–\$12,571
BERT_{base}	TPUv2x16	—	96	—	—	\$2074–\$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	\$942,973–\$3,201,722
NAS	TPUv2x1	—	32,623	—	—	\$44,055–\$146,848
GPT-2	TPUv3x32	—	168	—	—	\$12,902–\$43,008

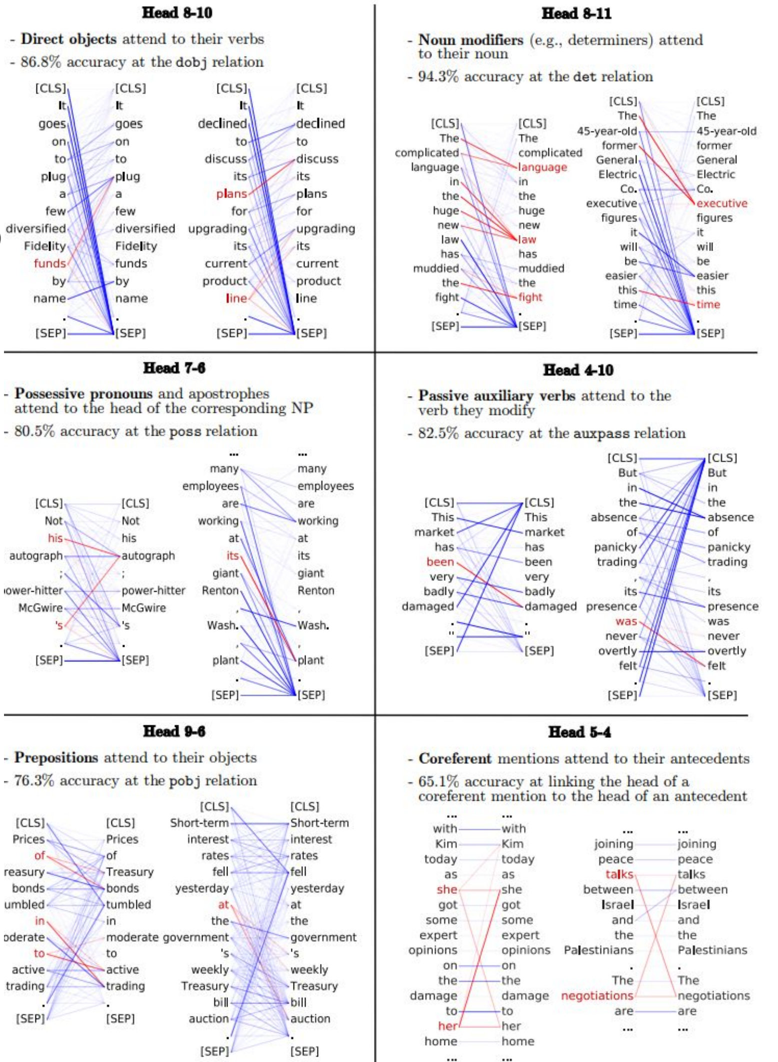
Table 3: Estimated cost of training a model in terms of CO_2 emissions (lbs) and cloud compute cost (USD).[7] Power and carbon footprint are omitted for TPUs due to lack of public information on power draw for this hardware.



Training labels
dog
is
a
good
boy
is
Task-specific Head
Final hidden-states
General-purpose Transformer Model
Hidden-states
Input embeddings
Training inputs
My
dog
is
a
good
My
dog
[MASK]
a
good
Language Models
(GPT, GPT-2, CTRL...)
Masked Language Models
(Bert, RoBERTa, ALBERT...)

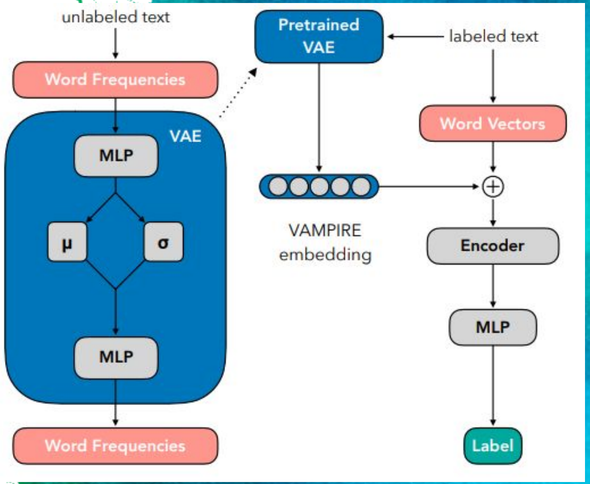
What does BERT Look at?

- Attention heads correspond to
  - **syntaxs**
  - **coreference**
  - **direct objects of verbs or prepositions**
  - **determiners of nouns**
- Clark, Khandelwal, Levy, Manning
  - https://arxiv.org/pdf/1906.04341.pdf
  - https://github.com/clarkkev/attention-analysis
- Alternatively see
  - Attention is not Explanation
  - Attention is not not Explanation



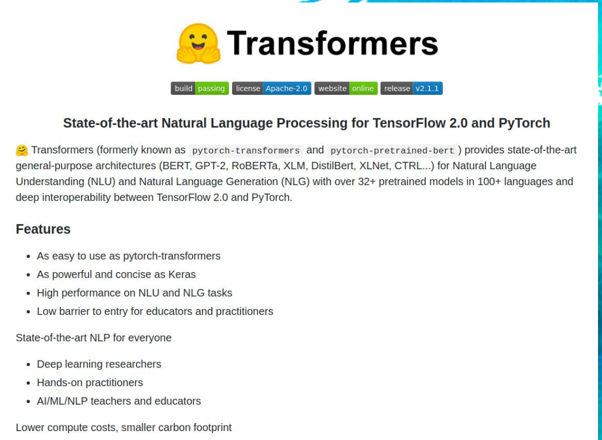
Transformer Updates

- **RoBERTa** performs better than BERT
  - https://arxiv.org/pdf/1907.11692.pdf
- **ALBERT** shows best results
  - https://arxiv.org/pdf/1909.11942.pdf
- **DistilBERT - less parameters**, 95% performance
  - https://arxiv.org/abs/1910.01108
- What about **lower resources** (e.g., less GPUs)?
  - **VAMPIRE** architecture is promising
  - https://arxiv.org/pdf/1906.02242.pdf



Unified Transformers

- Unified APIs for TF and PyTorch Transformers
- Includes most well-known Transformers
- Generally faster than original libraries
- SoA
- Focus on pre-trained models
- 2 libraries
  - https://github.com/huggingface/transformers
  - https://github.com/explosion/spacy-transformers



🤗 Transformers

build passing | license Apache-2.0 | website online | release v2.1.1

State-of-the-art Natural Language Processing for TensorFlow 2.0 and PyTorch

🤗 Transformers (formerly known as `pytorch-transformers` and `pytorch-pretrained-bert`) provides state-of-the-art general-purpose architectures (BERT, GPT-2, RoBERTa, XLM, DistilBert, XLNet, CTRL...) for Natural Language Understanding (NLU) and Natural Language Generation (NLG) with over 32+ pretrained models in 100+ languages and deep interoperability between TensorFlow 2.0 and PyTorch.

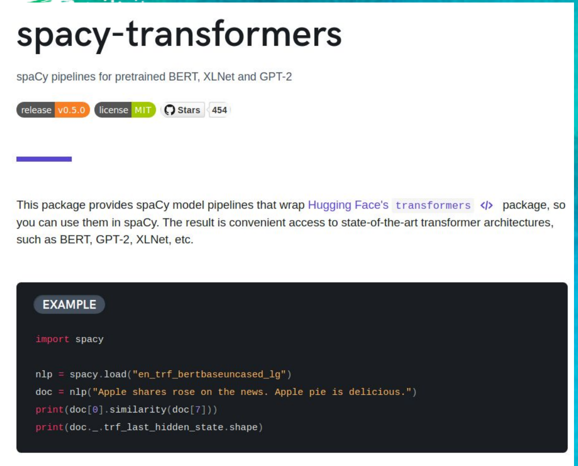
Features

- As easy to use as pytorch-transformers
- As powerful and concise as Keras
- High performance on NLU and NLG tasks
- Low barrier to entry for educators and practitioners

State-of-the-art NLP for everyone

- Deep learning researchers
- Hands-on practitioners
- AI/ML/NLP teachers and educators

Lower compute costs, smaller carbon footprint



spacy-transformers

spaCy pipelines for pretrained BERT, XLNet and GPT-2

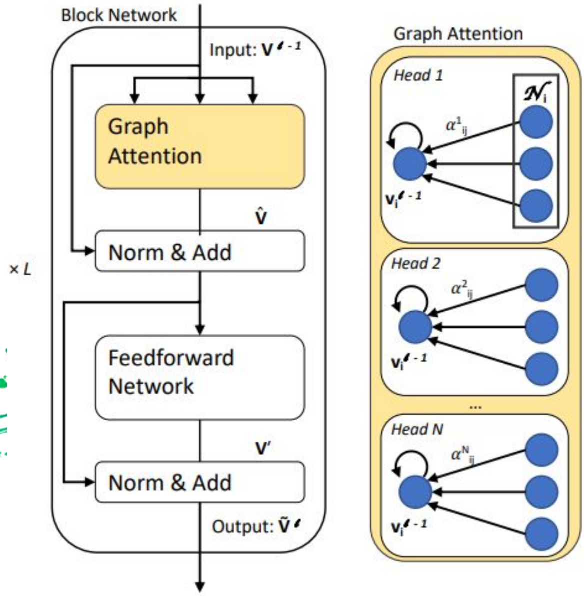
release v0.5.0 | license MIT | Stars 454

This package provides spaCy model pipelines that wrap Hugging Face's `transformers` package, so you can use them in spaCy. The result is convenient access to state-of-the-art transformer architectures, such as BERT, GPT-2, XLNet, etc.

EXAMPLE

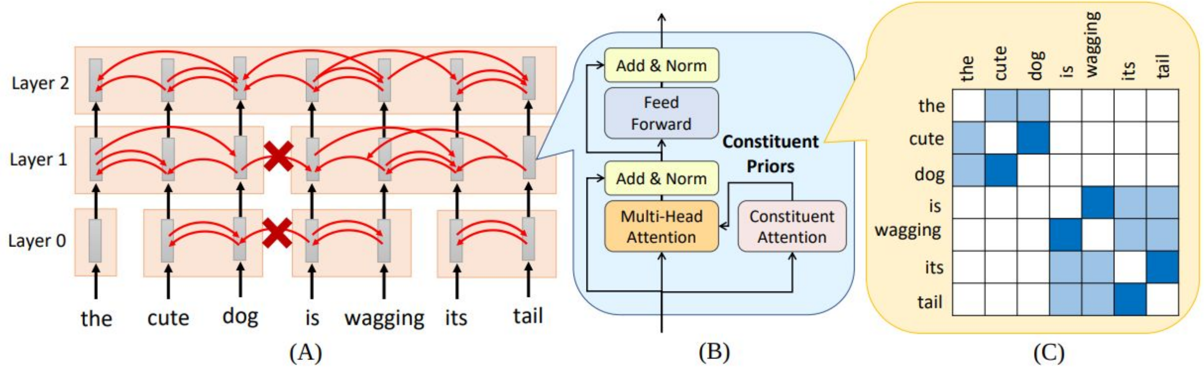
```
import spacy

nlp = spacy.load("en_trf_bertbaseuncased_lg")
doc = nlp("Apple shares rose on the news. Apple pie is delicious.")
print(doc[0].similarity(doc[7]))
print(doc._.trf_last_hidden_state.shape)
```

https://www.aclweb.org/anthology/N19-1238.pdf



Cross Transformers

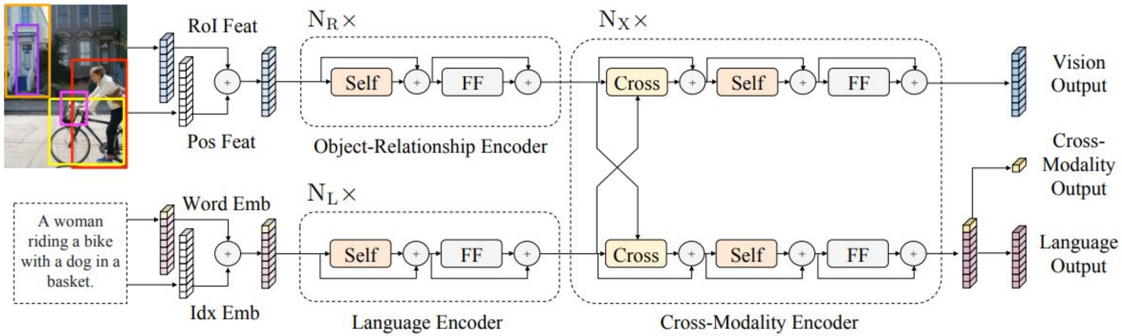
- Why should we use different Transformers?
- Cross-pollination
  - **Graph (NAACL)** - text generation
  - **Tree (EMNLP)** - composition, interpretation
  - **Capsule (WS)** - prediction

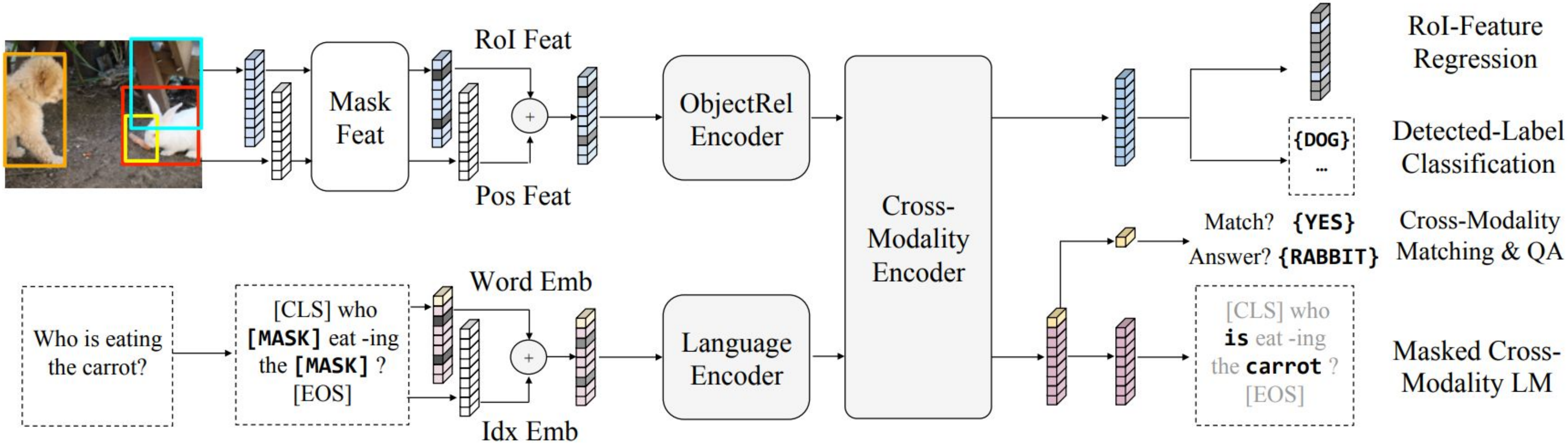


https://arxiv.org/pdf/1909.06639.pdf

https://arxiv.org/pdf/1908.07490.pdf

LXMERT Cross-Modality





Thanks!

Any questions?

You can find me at:

@AdrianB82