Machine Learning for Natural Language Processing

Guillaume Gravier

`guillaume.gravier@irisa.fr`

ENSAI

École nationale
de la statistique
et de l'analyse
de l'information

Outline of the course

Lectures (6 x 3h)

- Lecture #1: Introduction and representation of words

  Notions: morphology, tokens, lemmas, POS, word net, word embedding

  Hands-on: manipulate basic pipelines and visualize word embeddings

- Lecture #2: Representation of documents

  Notions: vocabulary, Zipf’s curse, bag of words, Bayes, RNN, BERT

  Hands-on: basic neural network classifiers

- Lecture #3: Language models

  Notions: ngrams, LSTM, bi-LSTM, language generation

  Hands-on: train a small LM and generate text

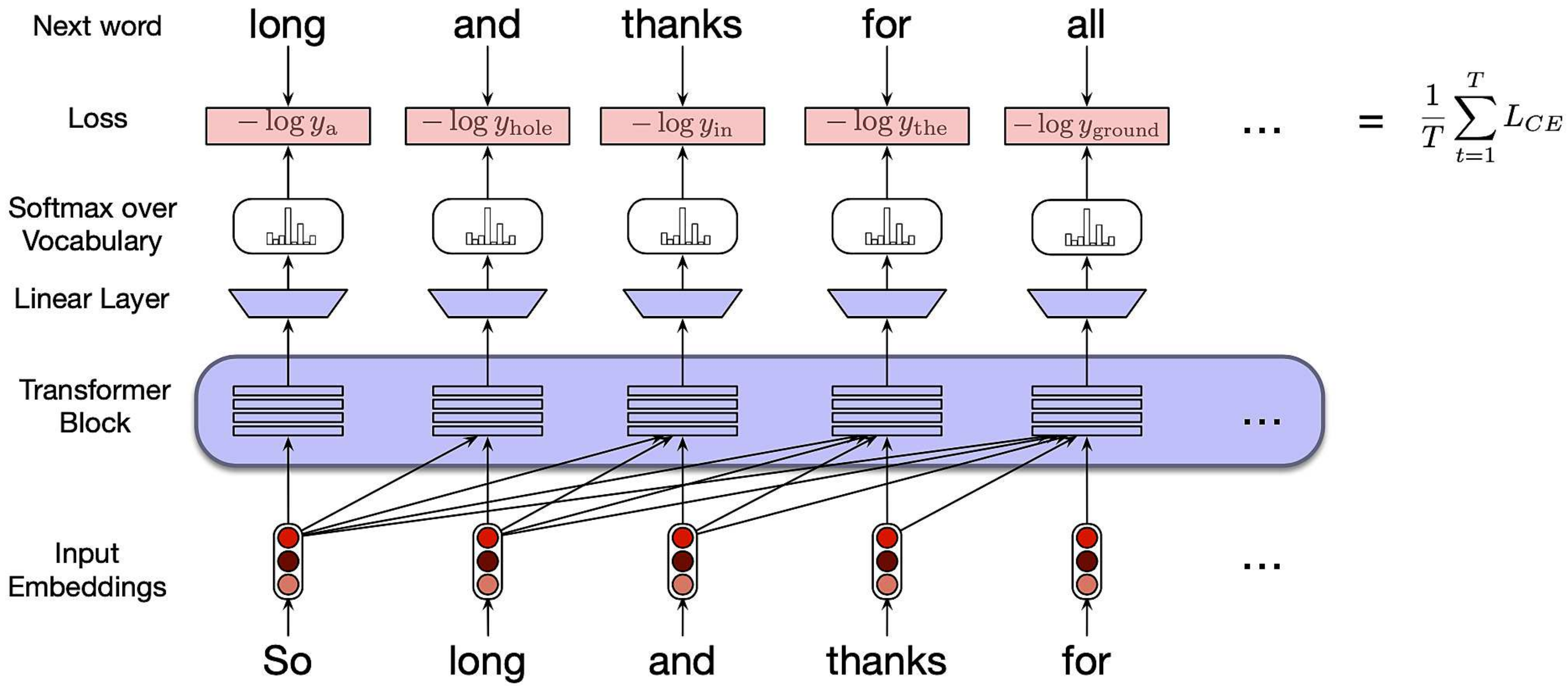
- Lecture #4: Transformers and large language models

  Notions: encoder/decoder, transformers, fine-tuning

  Hands-on: visualize embeddings, fine-tune a LLM

Large language models

Causal transformers as (generative) language models (aka GPT)

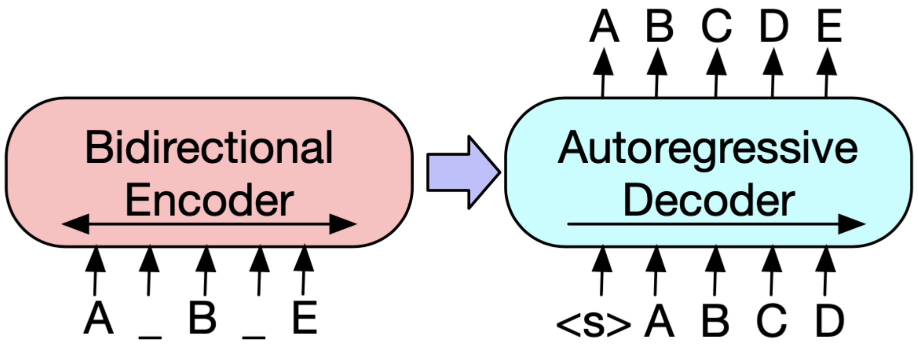


borrowed again from Jurafsky and Martin's book

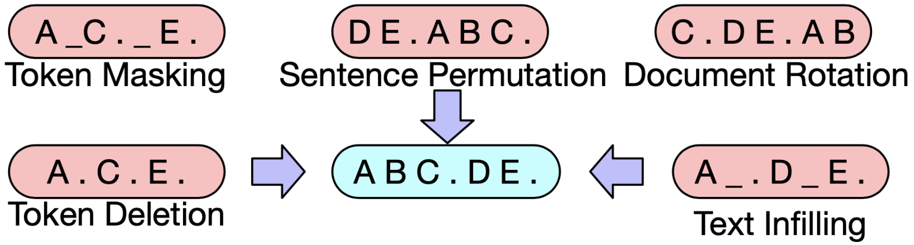
⇒ can be used for language generation exactly as recurrent networks

Encoder/decoder as a denoising auto-encoder

Pre-train a model that encodes a noisy version of the input and decodes the correct output



combining various input corruption mechanisms...



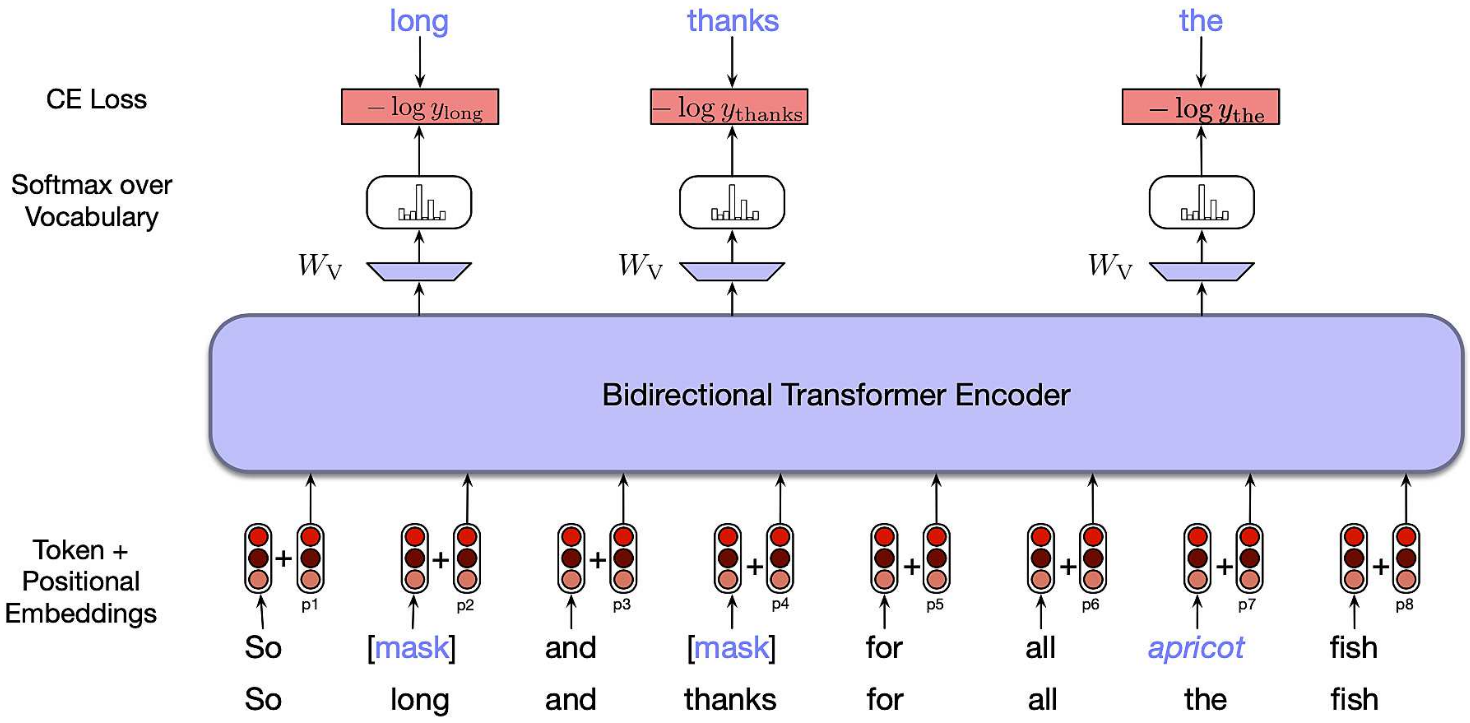
Lewis et al., 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.

See `https://huggingface.co/docs/transformers/en/model_doc/bart`

Turning the encoder into a language models (aka BERT)

Randomly mask or modify words and train networks to precit the actual word from the contextual embedding of the modified tokens:

→ Original BERT trained on 3.3B tokens, 15 % are modified: 80 % are actually masked, 10 % randomly replaced, 10 % left unchanged; training the LM objective function only on the 15 % modified tokens.

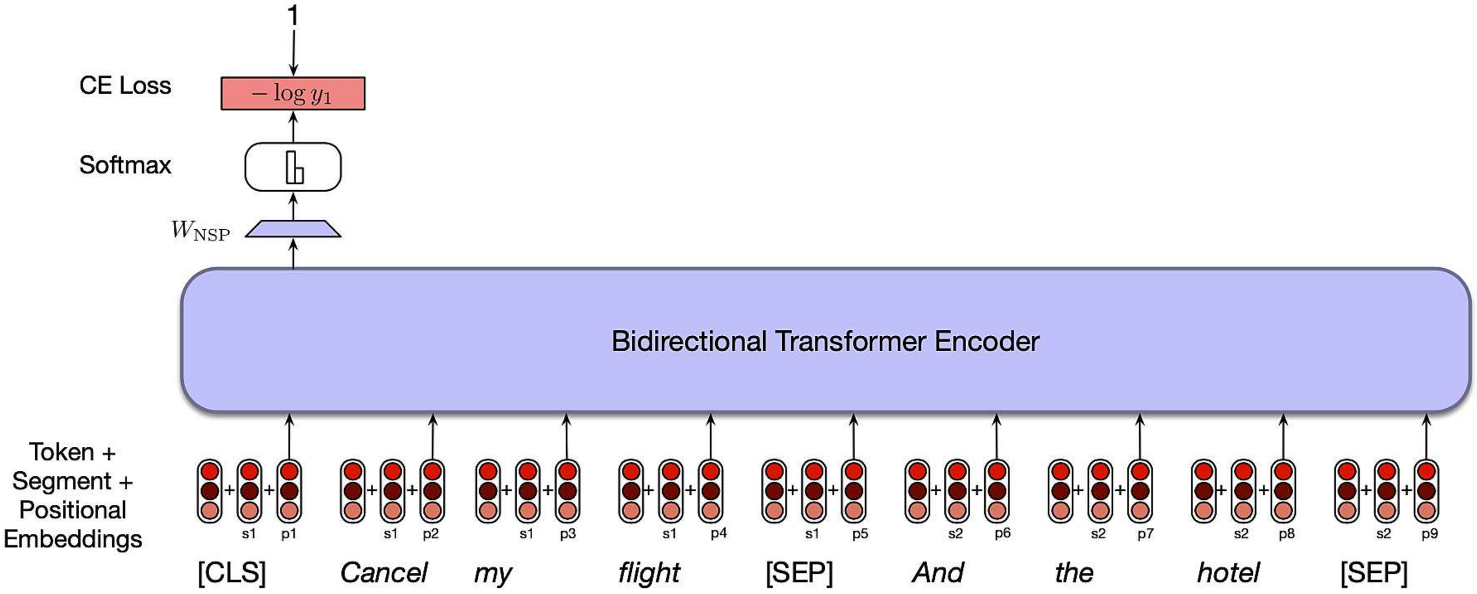


borrowed again and again from Jurafsky and Martin’s book

Pushing BERT one-step further with next sentence prediction

Many tasks look at two sentences (pieces of text as input), e.g., for natural language inference (entailment), measuring semantic proximity, summarization, etc.

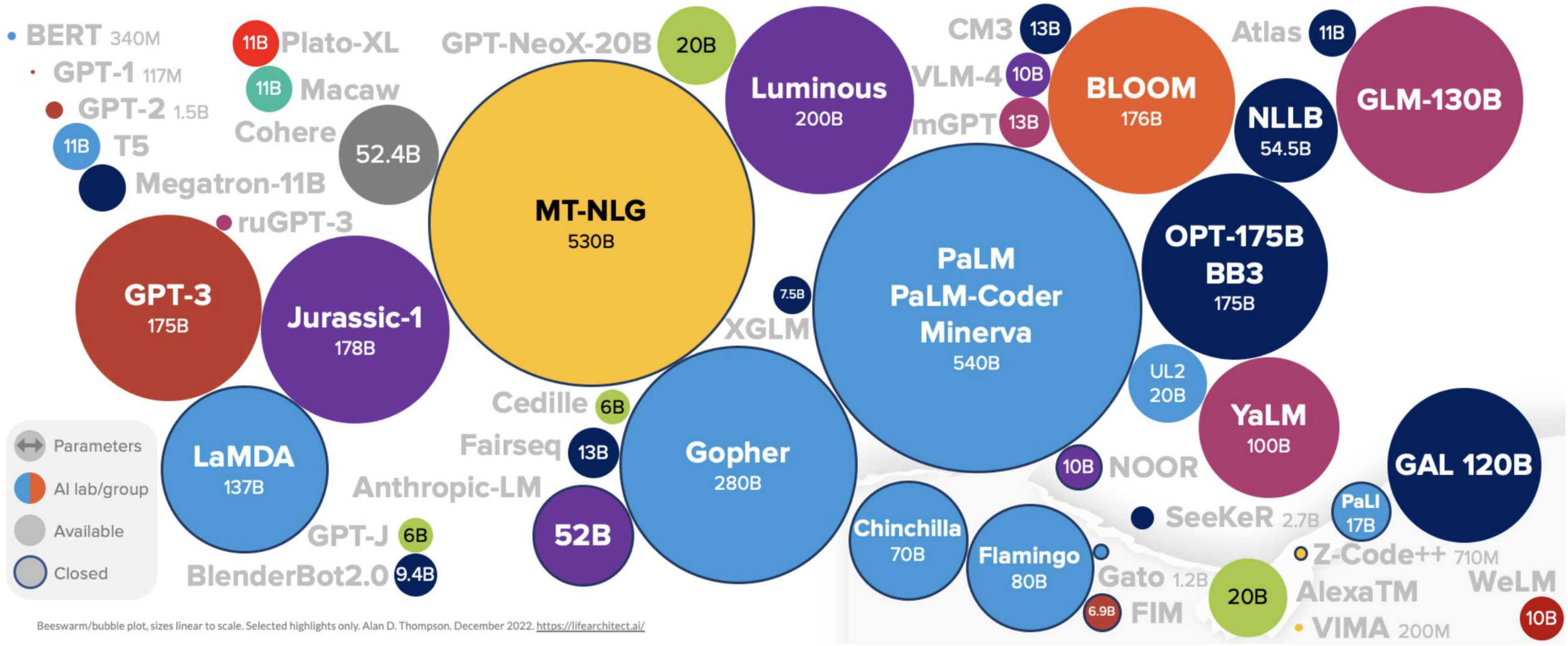
Train BERT with two sentences to predict whether one follows the other or not



borrowed again from Jurafsky and Martin's book

Full BERT training combines masked LM and next sentence prediction (from modified sentences)

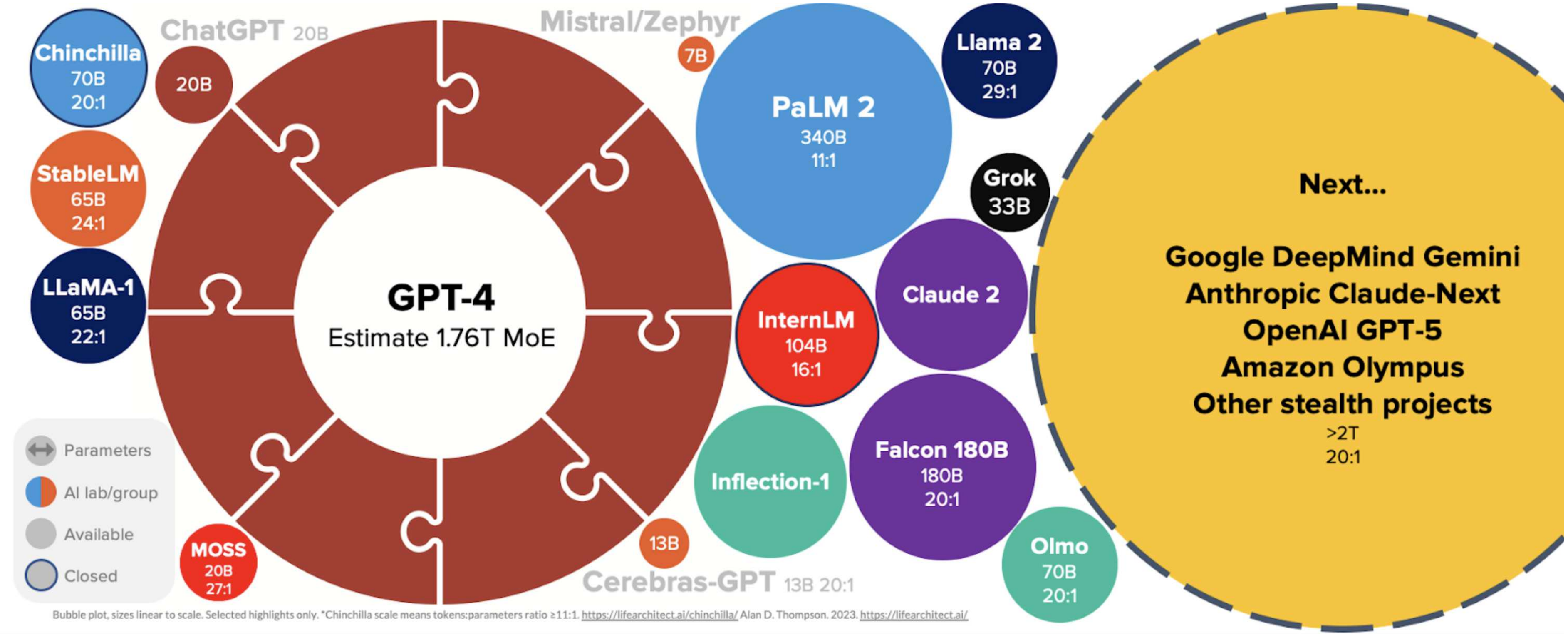
The transformer pre-trained model frenzy



Beeswarm/bubble plot, sizes linear to scale. Selected highlights only. Alan D. Thompson. December 2022. https://lifearchitect.ai/

copied from https://lifearchitect.ai/timeline

The transformer pre-trained model frenzy (more)

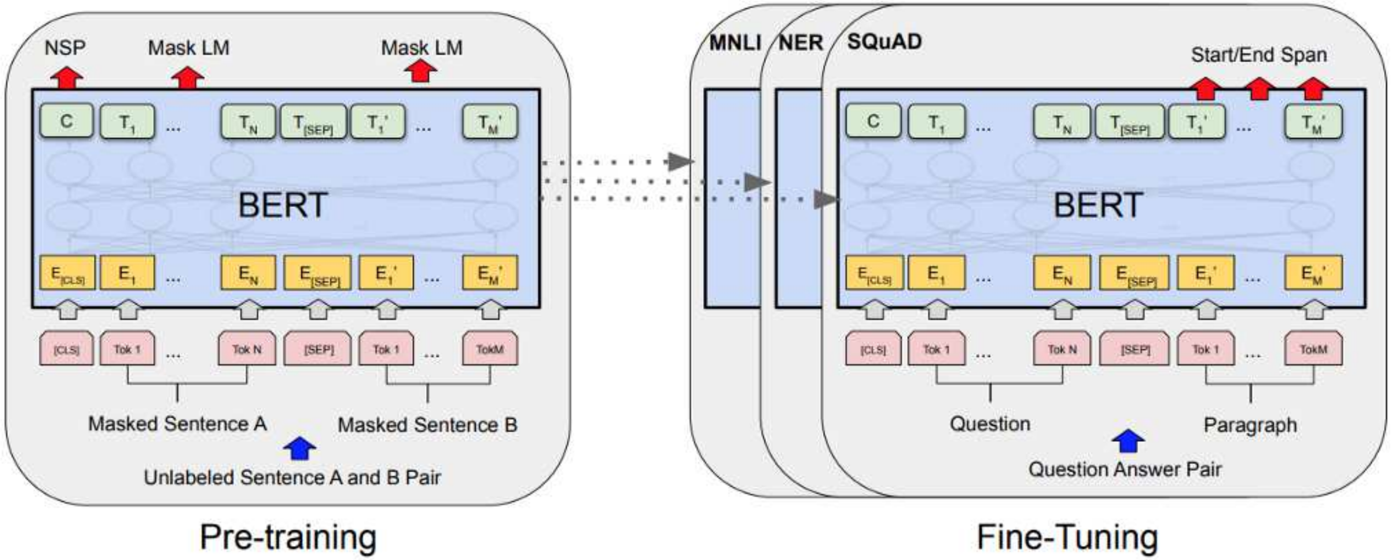


Bubble plot, sizes linear to scale. Selected highlights only. *Chinchilla scale means tokens:parameters ratio ≥11:1. https://lifearchitect.ai/chinchilla/ Alan D. Thompson. 2023. https://lifearchitect.ai/

copied from `https://lifearchitect.ai/timeline`

Transfer learning: the BERT case

Add a classifier on top of a pre-trained transformer language model and retrain the whole stuff

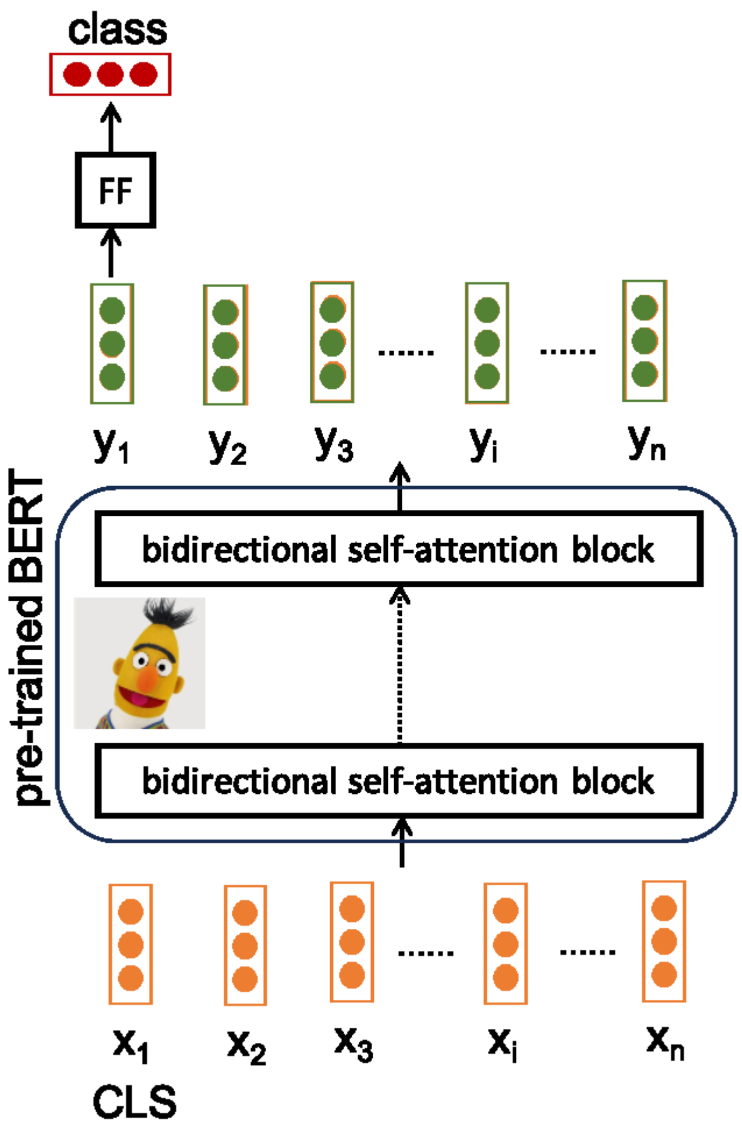


Jacob Devlin et al. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding

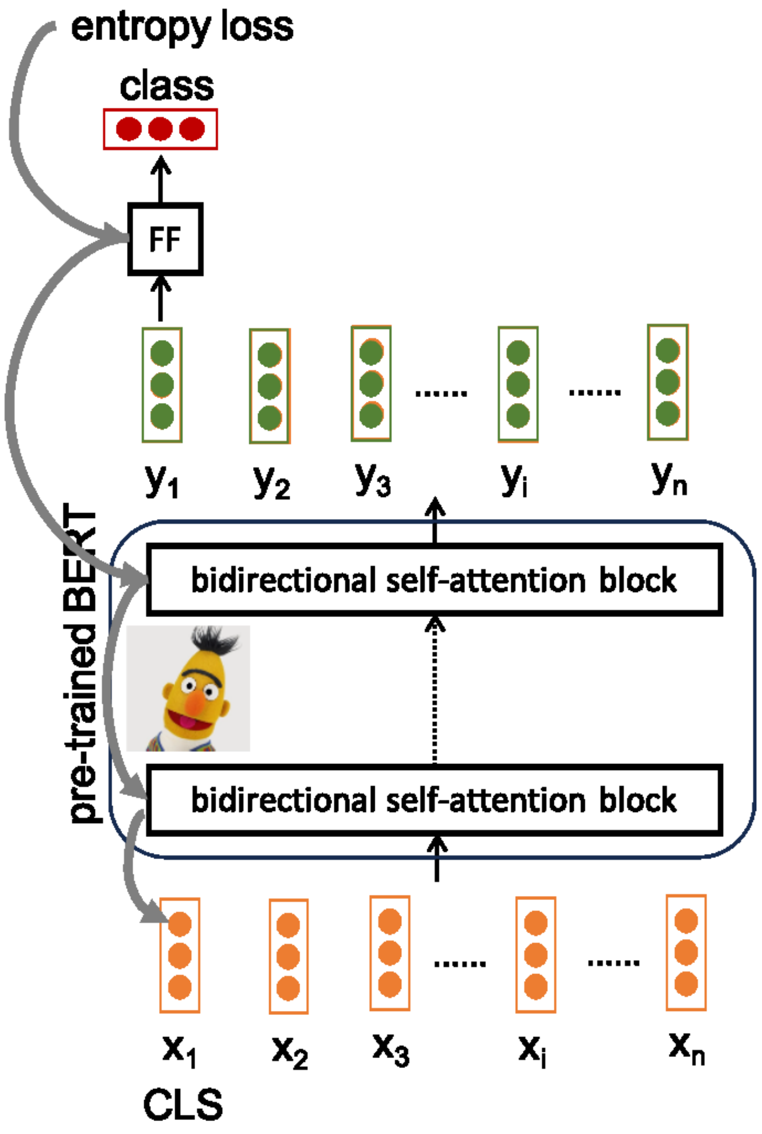
BERT's classification *head* depending on the task:

Classification	one sentence, embedding of the [CLS] token for classification
Comparison	two sentences, embedding of the [CLS] token for classification
Tagging	one sentence, contextual embeddings of tokens used in tagger

BERT transfer learning for document classification

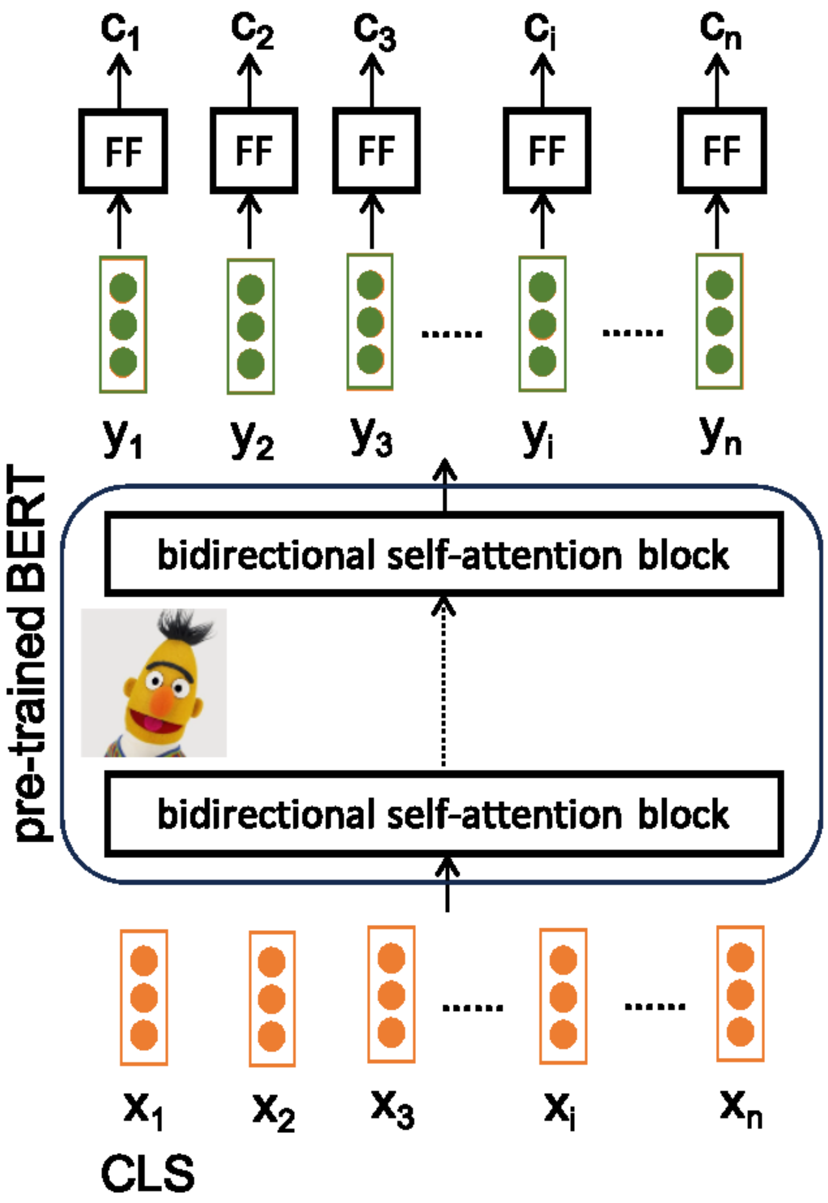

class
FF
y_1 y_2 y_3 ...... y_i ...... y_n
pre-trained BERT
bidirectional self-attention block
bidirectional self-attention block
x_1 x_2 x_3 ...... x_i ...... x_n
CLS

inference

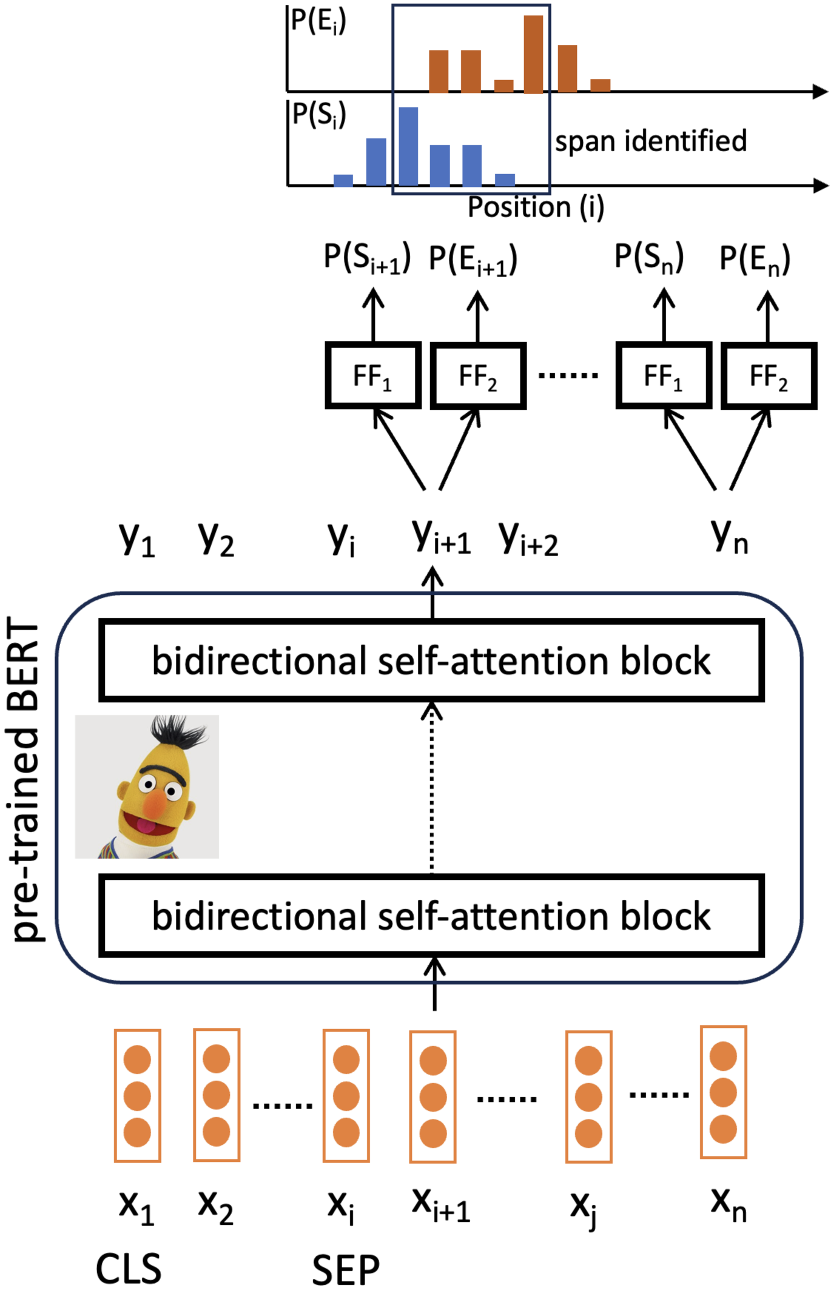

entropy loss
class
FF
y_1 y_2 y_3 ...... y_i ...... y_n
pre-trained BERT
bidirectional self-attention block
bidirectional self-attention block
x_1 x_2 x_3 ...... x_i ...... x_n
CLS

training

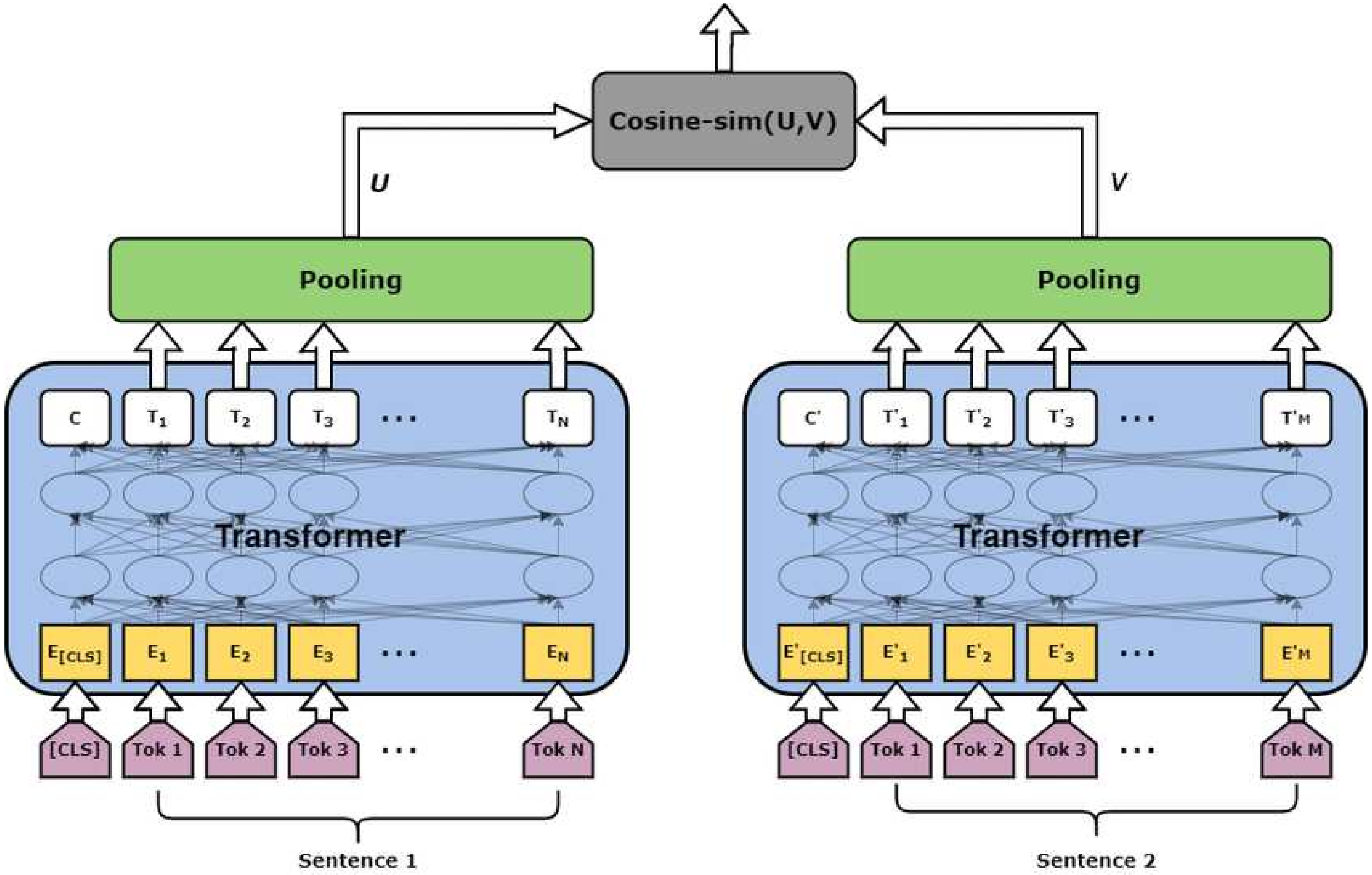
BERT transfer learning for token classification



BERT transfer learning for span detection (QA)



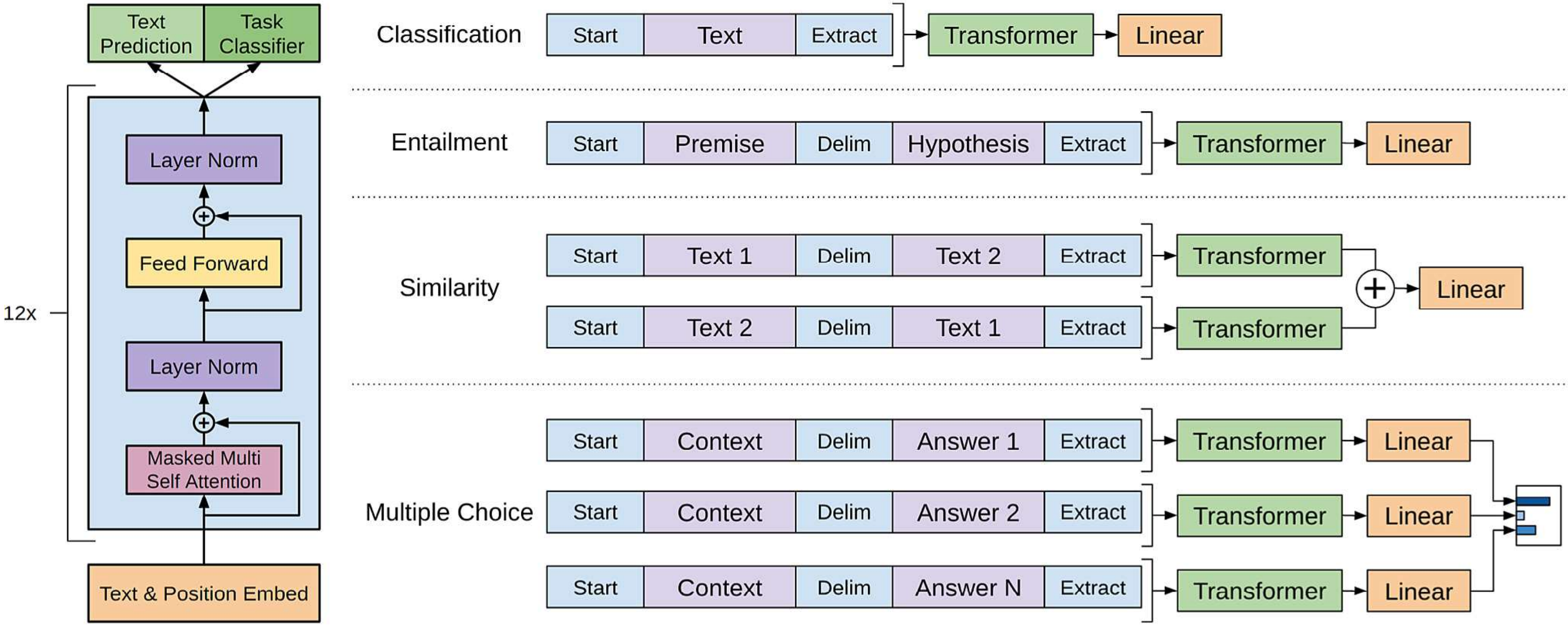
Sentence BERT: transfer learning for document representation



Borrowed from Hettiarachchi et al. (2021). Transformer-based Multilingual Socio-political and Crisis Event Detection.

Transfer learning: the GPT case

Add a classifier on top of a pre-trained transformer language model and retrain the whole stuff



Alex Radford et al. 2018. Improving Language Understanding by Generative Pre-Training

Other line of thoughts

- embedding spans rather than tokens
  - SpanBERT, GliNER, etc.
- retrieval augmented generation (RAG)
- prompt engineering has become increasingly popular
  - prompting models to classify, extract knowledge, etc.

This is just the begining of the NLP journey!!!!

That's all Folks!