

# Transformers

[https://github.com/TheHackerLlama/charlas/tree/main/r1iaa\\_2021](https://github.com/TheHackerLlama/charlas/tree/main/r1iaa_2021)



# Sobre mí - Omar Sanseviero

- <https://www.linkedin.com/in/omarsanseviero>
- <https://twitter.com/osanseviero>
- <https://osanseviero.github.io/hackerllama/>

TensorFlow  
Graphics



**HUGGING FACE**



Hi, how can I help?



# Agenda del día

1. Introducción a Transformers (35 minutos)
2. Clasificación de Texto (35 minutos)
3. Tendencias y nuevas direcciones (30 minutos)

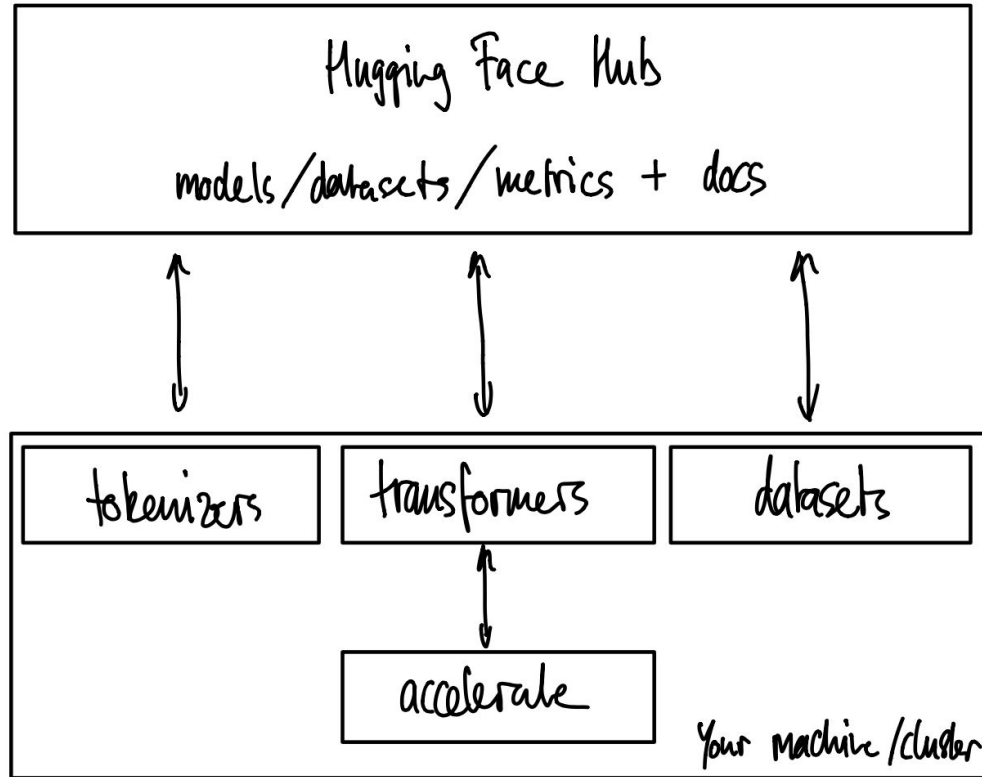


# Parte 1. Introducción a Transformers

1. Tareas comunes de NLP
2. ¿Qué son los Transformers?
3. Encoder, decoder y seq2seq
4. Limitaciones y sesgos



# Ecosistema de bibliotecas Hugging Face



[https://github.com/TheHackerLlama/charlas/tree/main/r1iaa\\_2021](https://github.com/TheHackerLlama/charlas/tree/main/r1iaa_2021)



# Pipelines

text-classification

summarization

zero-shot-answering

ner

text-generation

automatic-speech-recognition

question-answering

fill-mask

image-classification

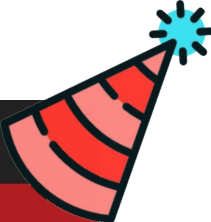
translation


conversational

table-question-answering



# Breve historia



 Cornell University

arXiv.org > cs > arXiv:1706.03762

**Computer Science > Computation and Language**

*[Submitted on 12 Jun 2017 (v1), last revised 6 Dec 2017 (this version, v5)]*

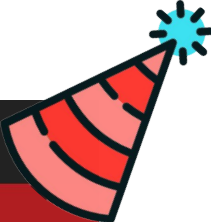
**Attention Is All You Need**


Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin





# Breve historia



 Cornell University

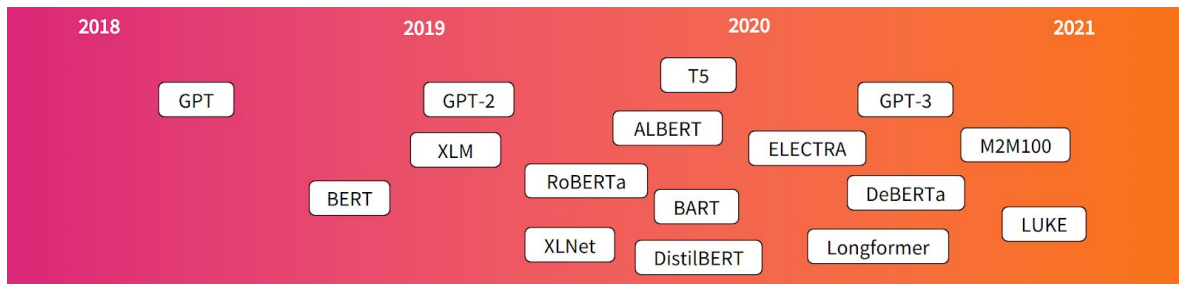
arXiv.org > cs > arXiv:1706.03762

**Computer Science > Computation and Language**

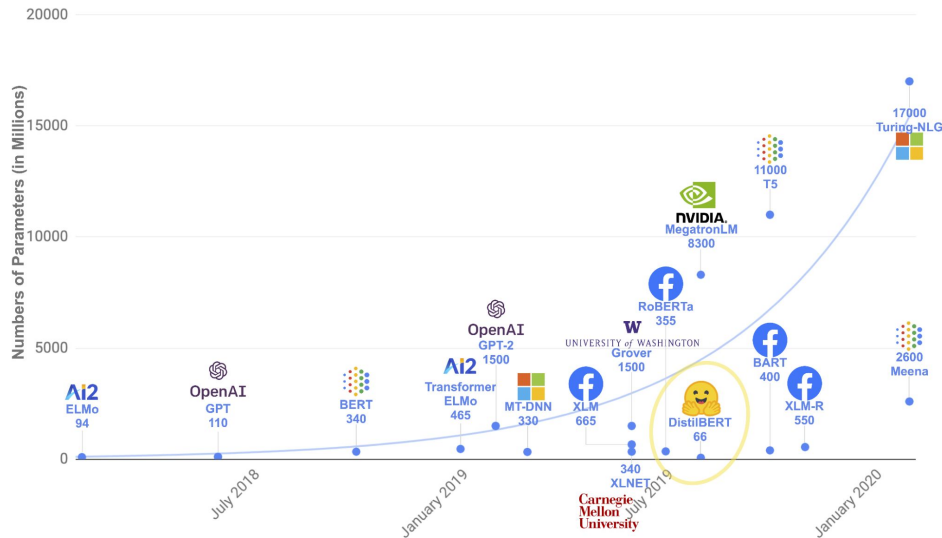
*[Submitted on 12 Jun 2017 (v1), last revised 6 Dec 2017 (this version, v5)]*

**Attention Is All You Need**

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin



# ¡Los Transformers son grandes!



Digital Journal

[PangyoTechnovalley] Naver to open South Korea's first ...

Korean data which HyperCLOVA learned is 6,500 times more than that of GPT-

3. The South Korea's AI model is the largest hyperscale ...

4 days ago



MarkTechPost

EleutherAI Develops GPT-3's Free Alternative: GPT-Neo

In terms of model size and computing power, the largest GPT-Neo model consists of 2.7 billion parameters. The GPT-3 API offers four models, ...

2 weeks ago



TNW

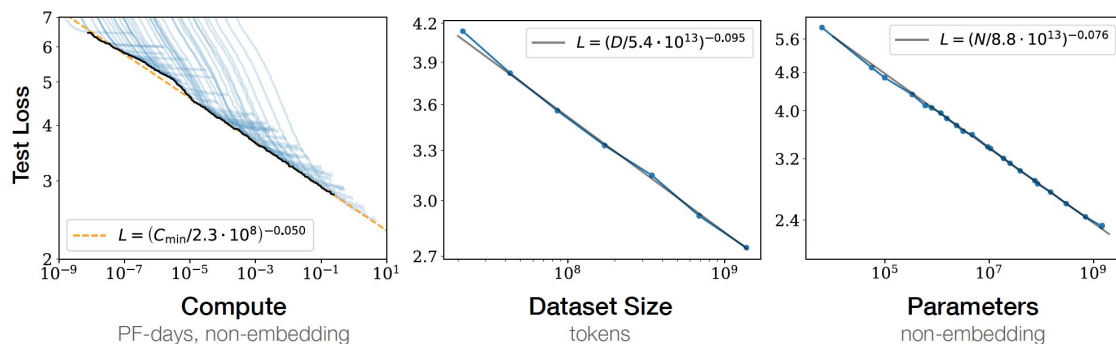
China's 'Wu Dao' AI is 10X bigger than GPT-3, and it can sing

China's going all in on deep learning. The Beijing Academy of Artificial Intelligence (BAAI) recently released details concerning its "Wu Dao" AI ...

5 days ago



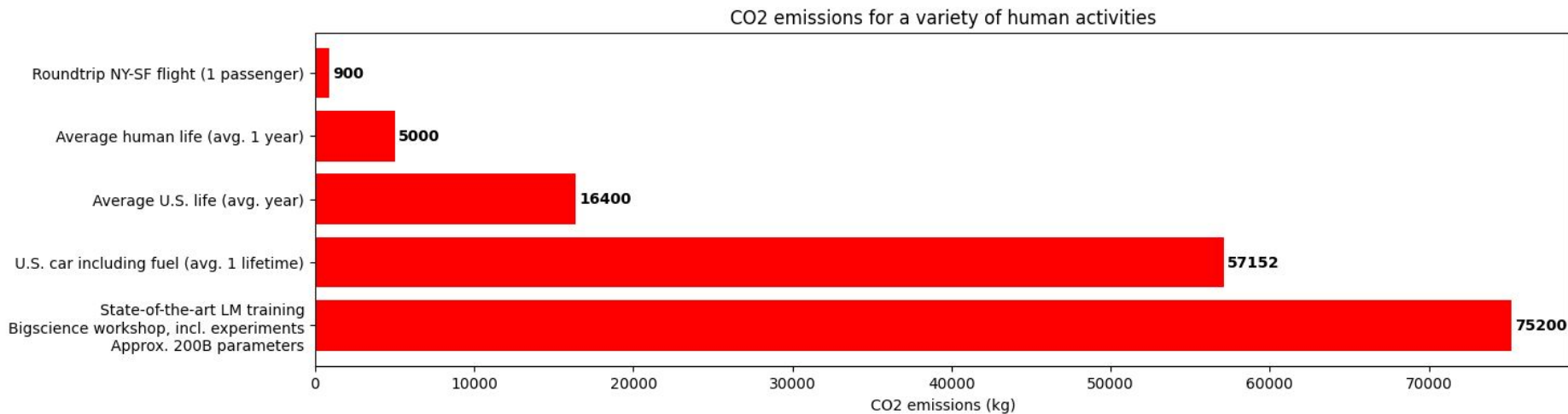
# ...y requieren muchos datos



**Figure 1** Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute<sup>2</sup> used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

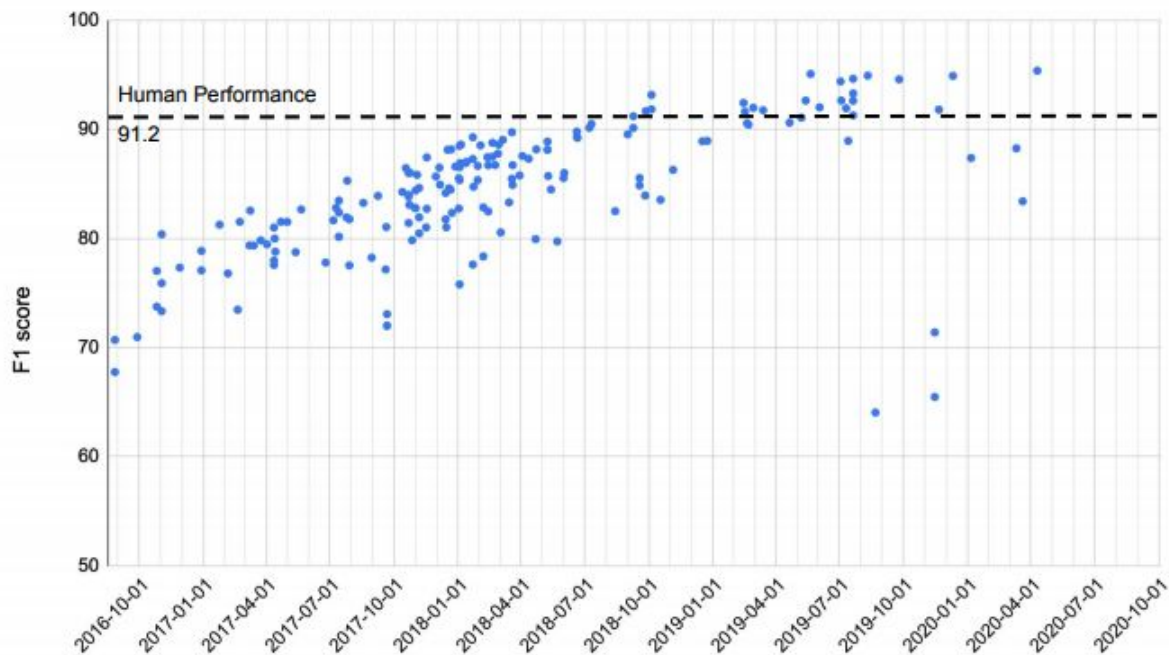


...y pueden tener alto impacto ambiental



... pero funcionan muy bien!

SQuAD1.1 F1 score vs. date



The Stanford Question Answering Dataset, <https://rajpurkar.github.io/SQuAD-explorer/>



# Problemas con Transformers

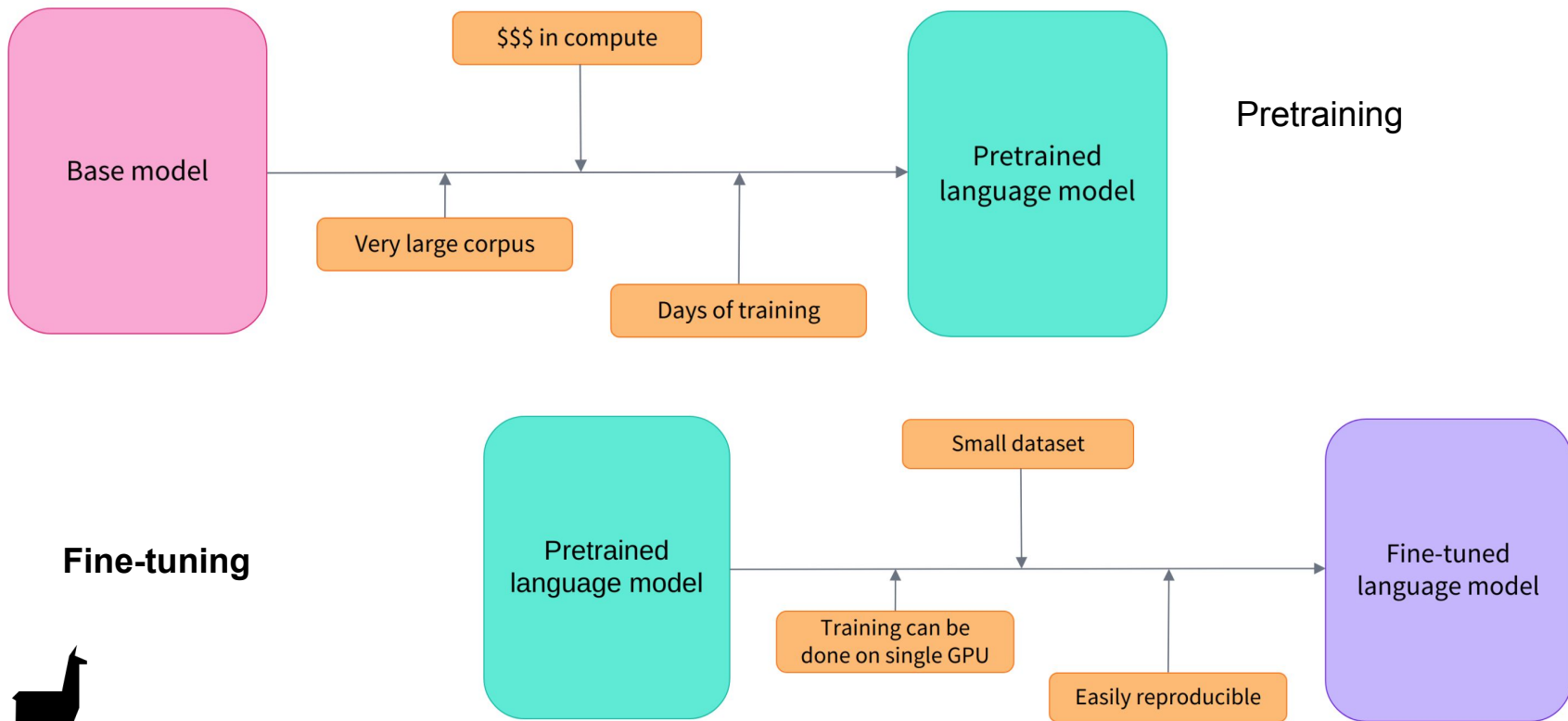
- Requieren muchos datos
- Requieren mucho poder computacional
- No funcionan con documentos largos
- Muchas veces funcionan como cajas negras
- Adoptan sesgos



# Transformers

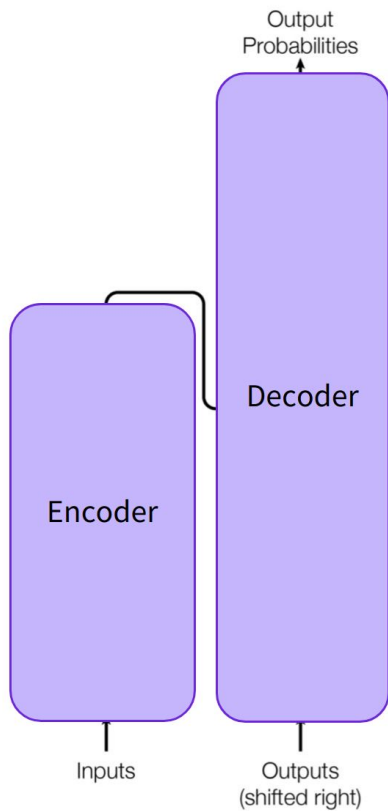


# Transfer Learning

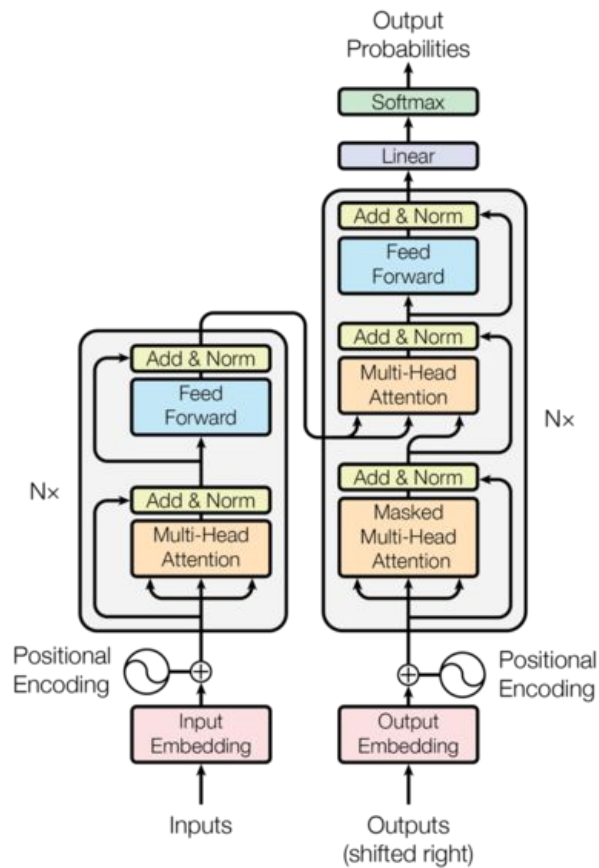




# Transformers



# Transformers



# Mecanismo de atención



# Encoders

- Aprenden a representar la entrada
- Crea un conocimiento estadístico del lenguaje

# Decoders

- Genera salidas a partir de una entrada y la representación del encoder



# Encoders

- Clasificación de texto
- NER
- POS

# Decoders

- Generación de texto

## Encoder-decoder (seq2seq)

- Generar texto a partir de una entrada
- Traducir
- Resumir



# Tareas con Transformers

Modelo	Ejemplos	Tareas
Encoder	ALBERT, <b>BERT</b> , DistilBERT, ELECTRA, RoBERTa	Clasificación de oraciones, NER, POS, extraer respuestas de un contexto
Decoder	CTRL, GPT, GPT-2, Transformer XL	Generación de Texto
Encoder Decoder	BART, T5, Marian, mBART	Resumir, traducción, generar respuestas a preguntas



# Los Transformers están en todos lados

NLP

Voz

Visión

BioChem

Series



RL



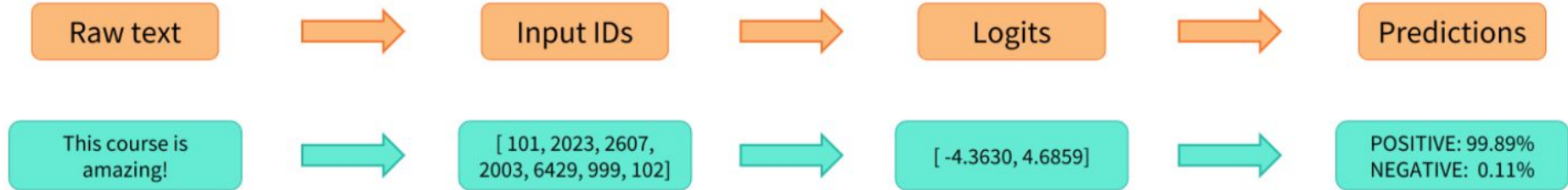
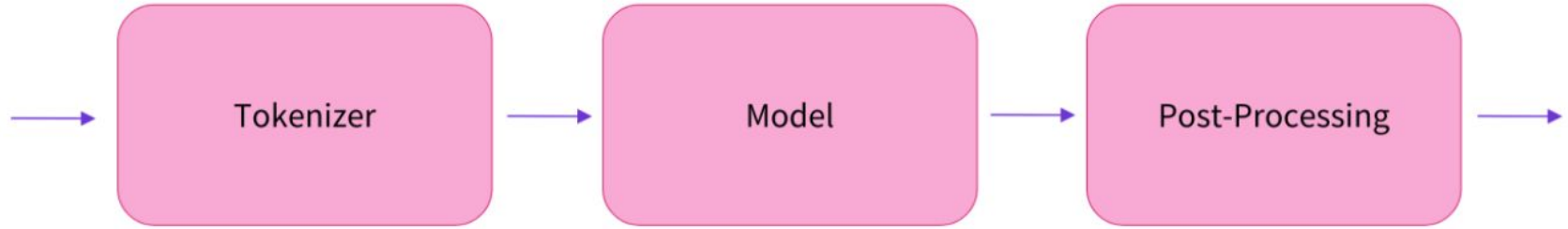
## Parte 2. Clasificación de Texto

1. ¿Cómo funciona un pipeline?
2. ¿Qué es un tokenizer?
3. ¿Cómo hacer clasificación de texto con fine-tuning?



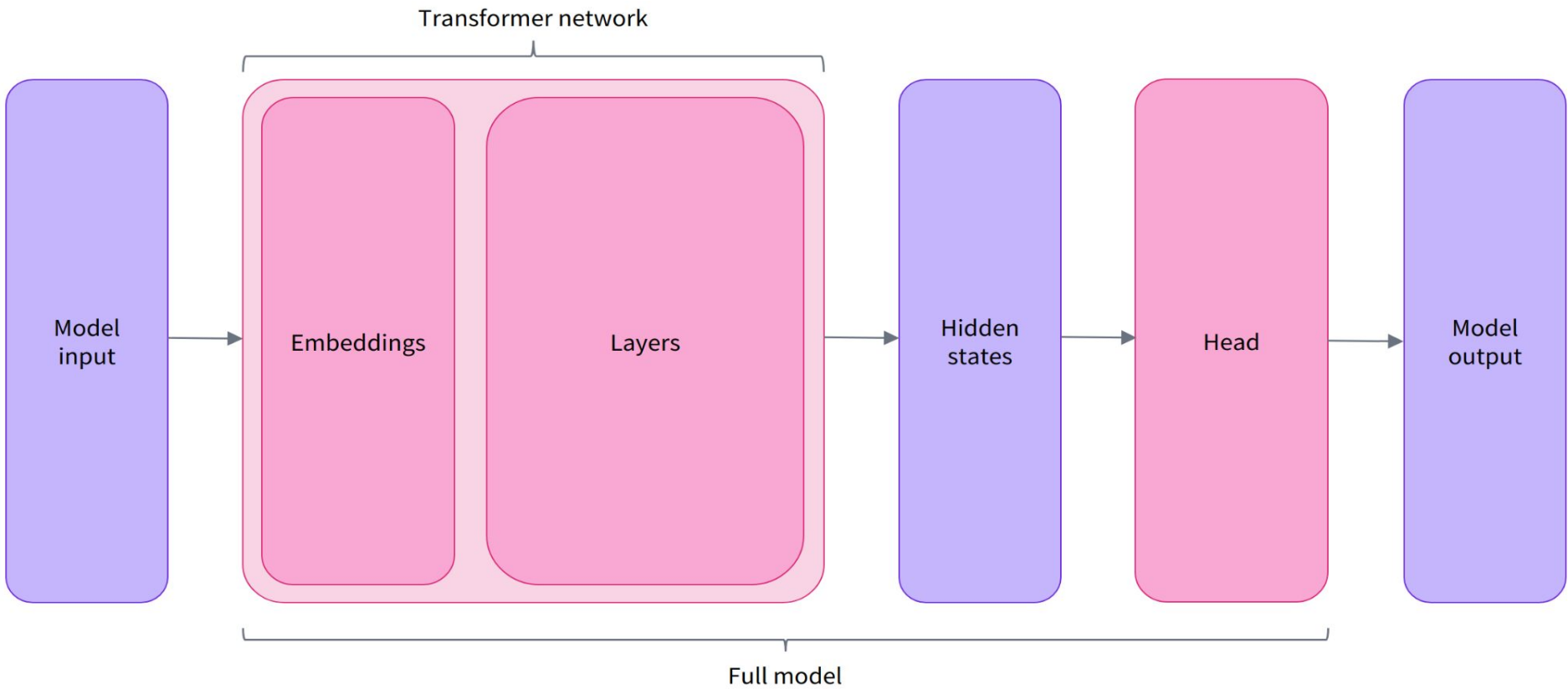


# ¿Cómo funciona un pipeline?



[https://github.com/TheHackerLlama/charlas/tree/main/r1iaa\\_2021](https://github.com/TheHackerLlama/charlas/tree/main/r1iaa_2021)





[https://github.com/TheHackerLlama/charlas/tree/main/r1iaa\\_2021](https://github.com/TheHackerLlama/charlas/tree/main/r1iaa_2021)



# Tokenizers

- Word-based
- Character-based
- Subword-based
- muchos otros



# Word-based

Split on spaces

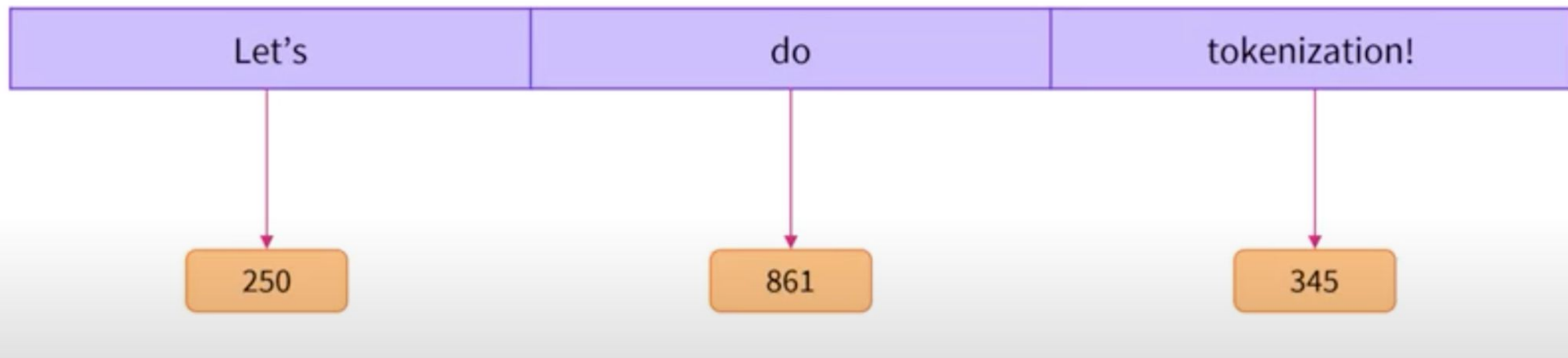
Let's	do	tokenization!
-------	----	---------------

Split on punctuation

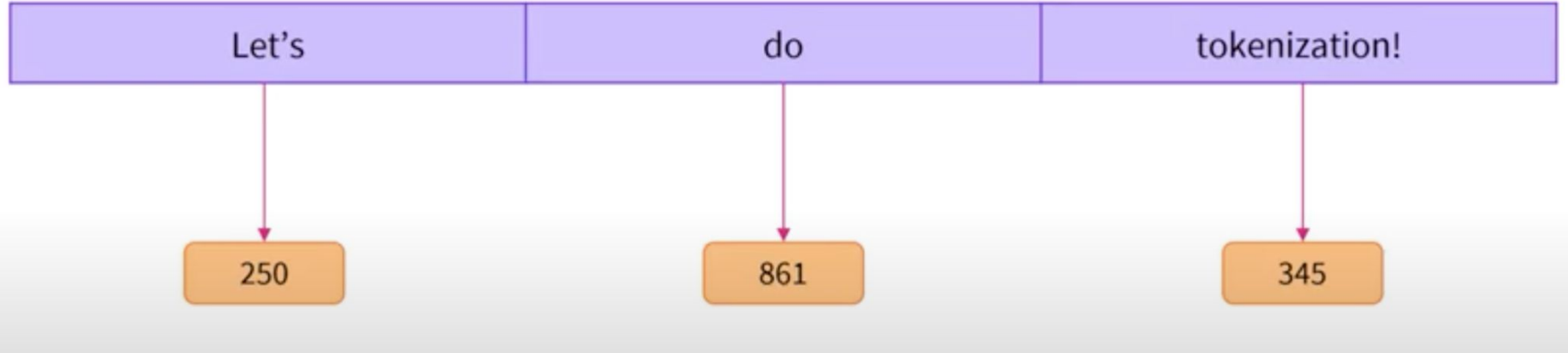
Let	's	do	tokenization	!
-----	----	----	--------------	---



# Word-based



# Word-based



¿Qué problemas puede causar esto?





# Word-based

**¿Qué problemas puede causar esto?**  
Palabras similares, significados diferentes

the	→	1
of	→	2
and	→	3
to	→	4
in	→	5
was	→	6
the	→	7
is	→	8
for	→	9
as	→	10
on	→	11
with	→	12
that	→	13
dog	→	14
dogs	→	15



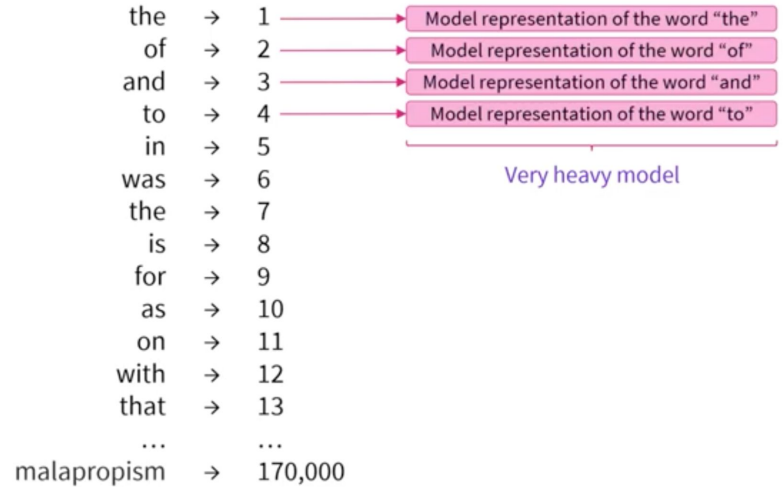
# Word-based

**¿Qué problemas puede causar esto?**  
Vocabularios muy grandes

the	→	1
of	→	2
and	→	3
to	→	4
in	→	5
was	→	6
the	→	7
is	→	8
for	→	9
as	→	10
on	→	11
with	→	12
that	→	13
...		...
malapropism	→	170,000



# Word-based



**¿Qué problemas puede causar esto?**  
Vocabularios muy grandes



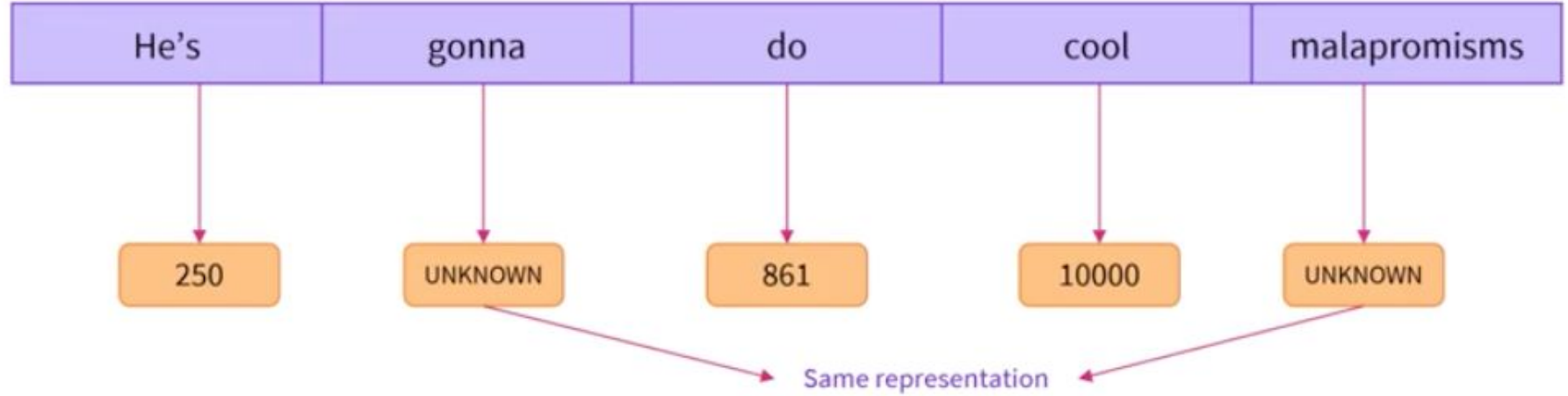
# Word-based

**¿Qué problemas puede causar esto?**  
Vocabularios muy grandes

the	→	1
of	→	2
and	→	3
to	→	4
in	→	5
was	→	6
the	→	7
is	→	8
for	→	9
as	→	10
on	→	11
with	→	12
that	→	13
...		...
hug	→	10,000



# Word-based



**¿Qué problemas puede causar esto?**  
Vocabularios muy grandes



# Character-based

L	e	t	'	s	d	o	t	o	k	e	n	i	z	a	t	i	o	n	!
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---



# Character-based

- ¡256 caracteres vs cientos de miles de palabras!
- Casi no hay OOV/UNK tokens
- Menos informativos
  - p, e, r, r, o vs perro
  - pero más útil para ciertos lenguajes en los que un caracter representa una palabra
- Secuencias muy largas



# Subword-based

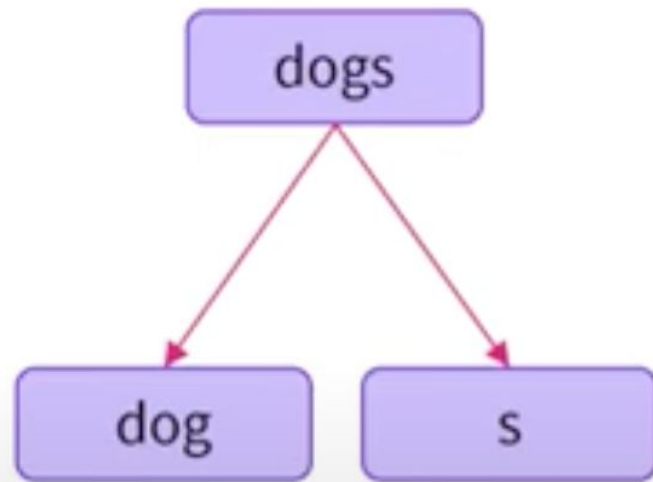
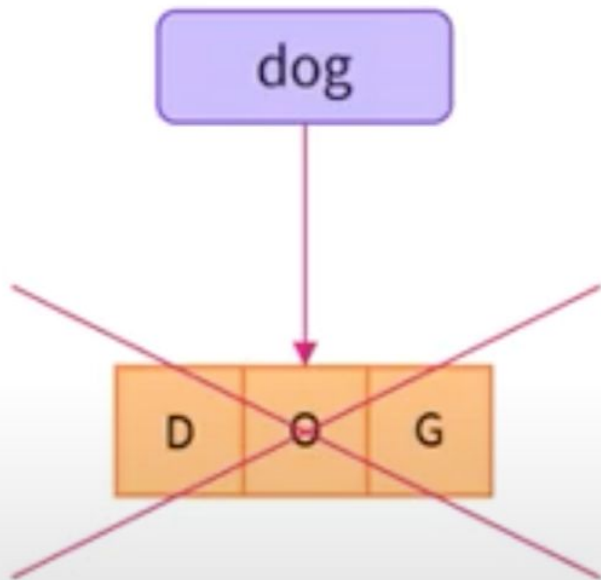
Let's</w>	do</w>	token	ization</w>	!</w>
-----------	--------	-------	-------------	-------

- Palabras frecuentes no se separan
- Palabras raras se descomponen en palabras significantivas

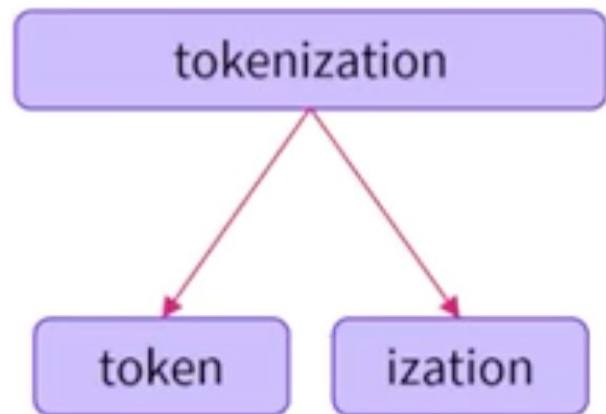




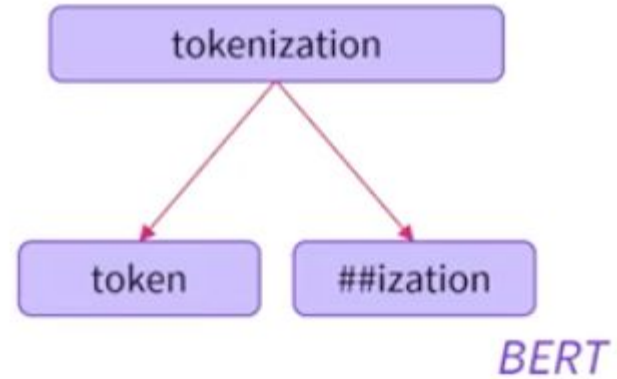
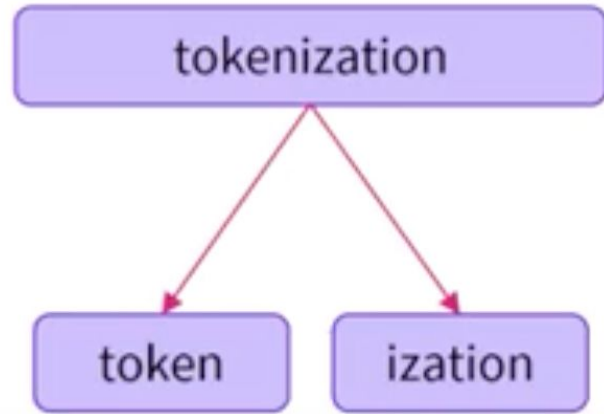
# Subword-based



# Subword-based



# Subword-based



# Otros métodos

- Byte-level BPE
- WordPiece
- SentencePiece
- Unigram

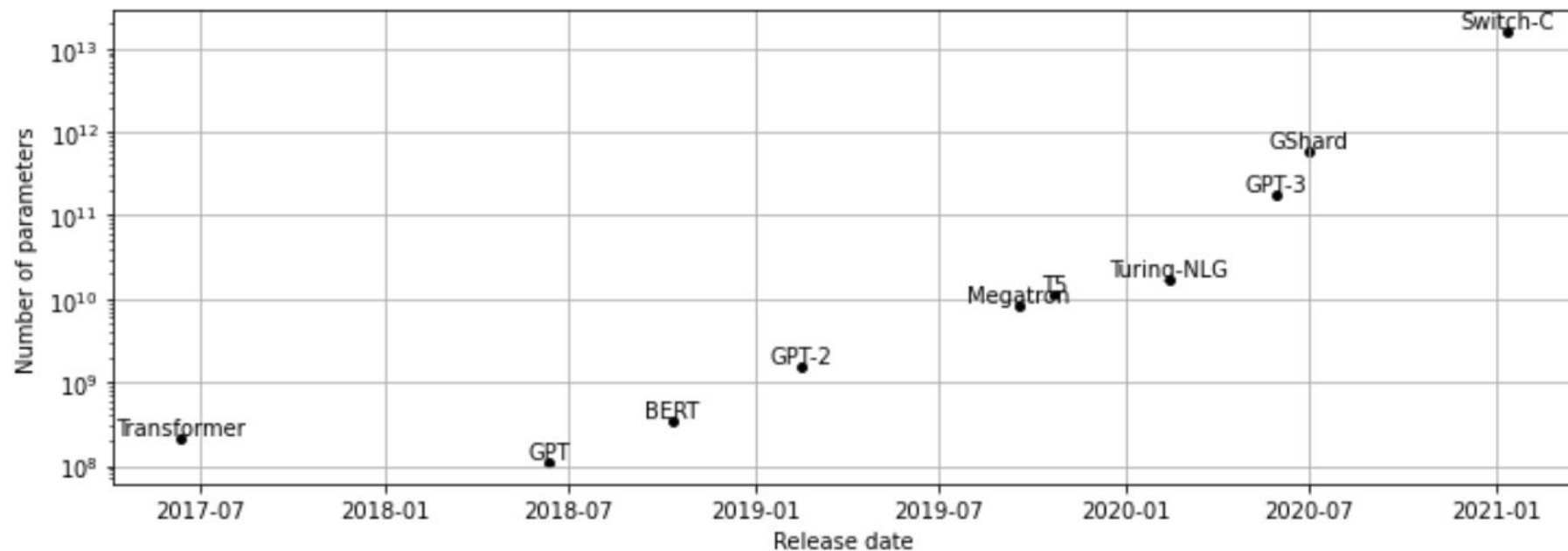


## Parte 3. Futuras direcciones

1. Escalando transformers
2. Iniciativas de comunidad
3. Otras aplicaciones



# Escalando Transformers



# Escalando Transformers

- Evidencia empírica que modelos grandes se desempeñan mejor en tareas **downstream**
- Propiedades interesantes que emergen en 10-100 mil millones de parámetros
  - Aprendizaje zero-shot
  - Aprendizaje few-shot
- ¿Sólo importa el número de parámetros?



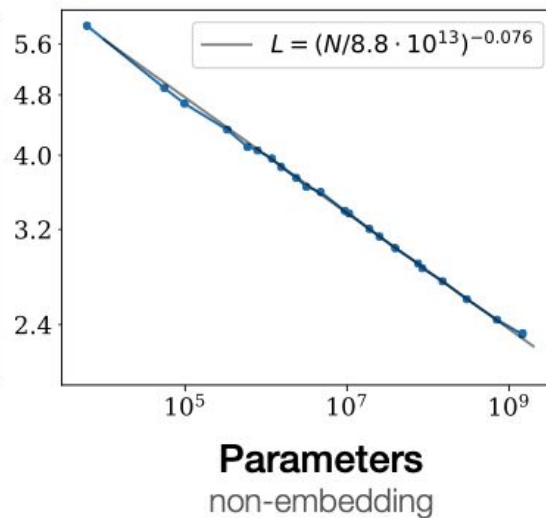
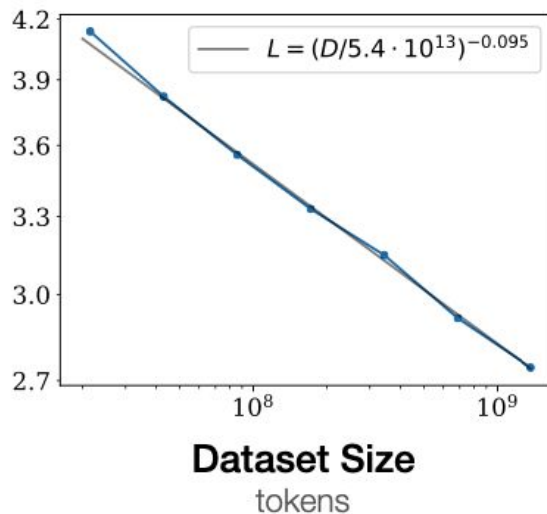
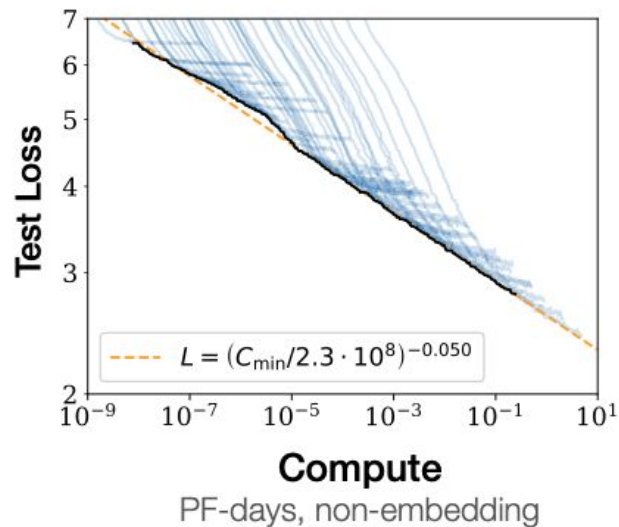
# Escalando Transformers

- Se estima que GPT-3 costó \$4.6 millones para entrenar (OpenAI's GPT-3 Language Model: A Technical Overview, C. Li (2020))
- Leyes de la escalabilidad (Scaling Laws for Neural Language Models, J. Kaplan et al. (2020))





# Leyes de escalabilidad



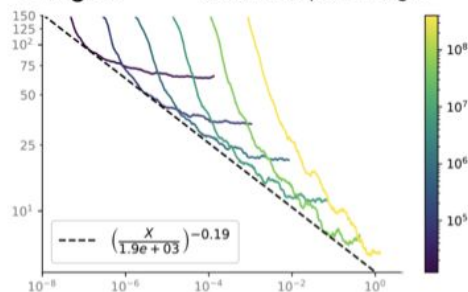
# Leyes de escalabilidad

- Cuantifica empíricamente el paradigma que modelos grandes son mejores
  - C: compute power
  - D: tamaño del dataset
  - N: tamaño del modelo
- Sample efficiency:
  - Modelos grandes tienen mismo desempeño que modelos chicos con menos entrenamiento

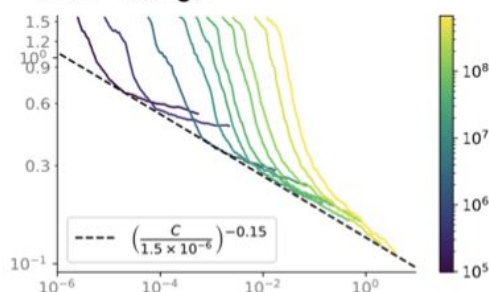


# Leyes de escalabilidad no están limitadas a texto

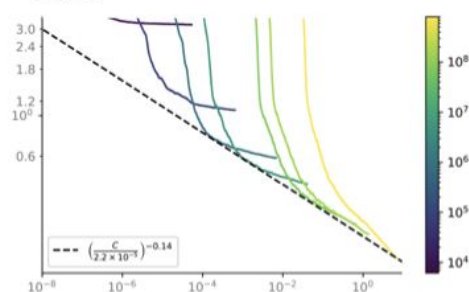
Images 8x8, loss per image



Text → Image



Video



Math

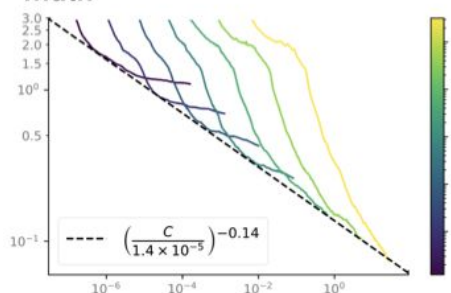
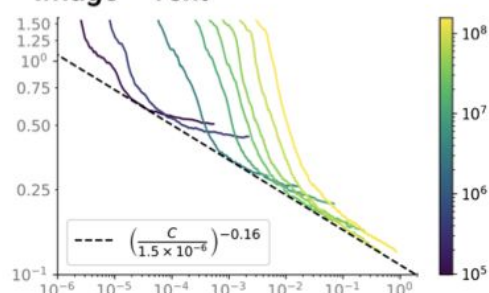
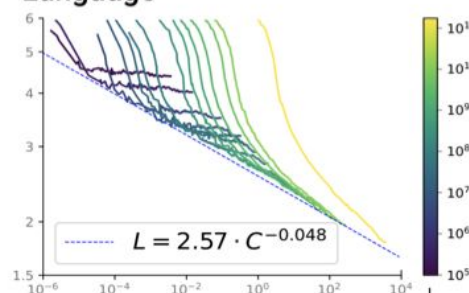


Image → Text



Language



Compute (PF-days)

Line color denotes model size



# Retos de escalar

- Infraestructura
- Costo
- Curación de datasets
- Evaluación de modelos
- Despliegue



# Infraestructura

- 100s a 1000s de nodos con GPU
- Algunos modelos no entran en GPU de consumidores
- Muchos problemas que no suelen ser parte del skillset de equipos normales de data science
- Se requieren ingenieras familiares con experimentos distribuidos de larga escala



# Costo



# Costo

Distributed Deep Learning in Open Collaborations (Diskin et. al., 2021)



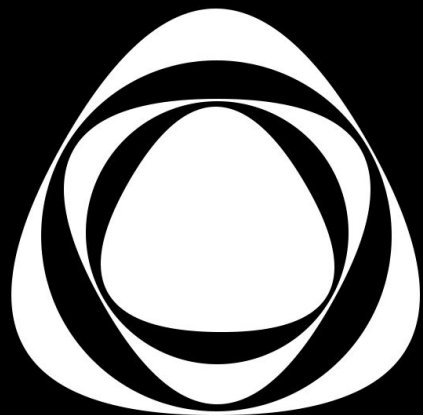
# Curación de datasets

- Garbage in, garbage out
- Se necesitan datasets grandes de “buena” calidad
- TBs de datos hacen problemas retadores
  - ¿Cómo procesar teras de datos?
  - ¿Cómo controlar sesgos como sexismo o racismo en los datos?
  - ¿Cómo se manejan temas de licenciamiento?





## Iniciativas de comunidad



EleutherAI

BigScience 🌸



# Otras aplicaciones

- Visión
  - iGPT
  - ViT
  - ¿Qué podemos hacer para videos?
    - Is Space-Time Attention All You Need for Video Understanding?, G. Bertasius et al. (2021)
- Tablas
  - TAPAS
- Multimodal
  - Speech to text



[https://github.com/TheHackerLlama/charlas/tree/main/r1iaa\\_2021](https://github.com/TheHackerLlama/charlas/tree/main/r1iaa_2021)



# Multimodal

- Visión con texto
  - VQA
  - LayoutLM
  - DALL-E
  - CLIP



¡Muchas gracias!



## HUGGING FACE

- <https://www.linkedin.com/in/omarsanseviero>
- <https://twitter.com/osanseviero>
- <https://osanseviero.github.io/hackerllama/>



# Recursos adicionales

- <https://huggingface.co/course>
- <https://jalammar.github.io/illustrated-transformer/>
- <https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/>
- <https://jalammar.github.io/how-gpt3-works-visualizations-animations/>
- [https://lena-voita.github.io/nlp\\_course.html](https://lena-voita.github.io/nlp_course.html)
- <https://bigscience.huggingface.co/en/#!/index.md>
- <https://eleuther.ai/>

