MiniLM: Deep Self-attention Distillation For Task-agnostic Compression Of Pre-trained Transformers

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, Ming Zhou

Microsoft Research 2020

Miguel Angel Ruiz Ortiz

Maestría en Matemáticas Aplicadas

miguel.ruiz@cimat.mx

LIMITACIONES DE MODELOS GRANDES

- Modelos de lenguaje pre-entrenados (e.g., BERT) → Éxito en NLP □

- Tamaño masivo! → $\text{BERT}_{\text{BASE}}$ (110M parámetros)
  $\text{BERT}_{\text{LARGE}}$ (340M parámetros)

#Layers	Hidden Size	#Param (Emd)	#Param (Trm)	Inference Time
12	768	23.4M	85.1M	93.1s (1.0×)
6	768	23.4M	42.5M	46.9s (2.0×)
12	384	11.7M	21.3M	34.8s (2.7×)
6	384	11.7M	10.6M	17.7s (5.3×)
4	384	11.7M	7.1M	12.0s (7.8×)
3	384	11.7M	5.3M	9.2s (10.1×)

- Emd: Embedding -Trm: Transformer

Confidential - Oracle Restricted

LIMITACIONES DE MODELOS GRANDES

- Modelos de lenguaje pre-entrenados (e.g., BERT) → Éxito en NLP □
- Tamaño masivo! → $\text{BERT}_{\text{BASE}}$ (110M parámetros)
  $\text{BERT}_{\text{LARGE}}$ (340M parámetros)

- Complica el fine-tuning
- Tiempos de inferencia lentos
- Límites en latencia en modelos en producción.

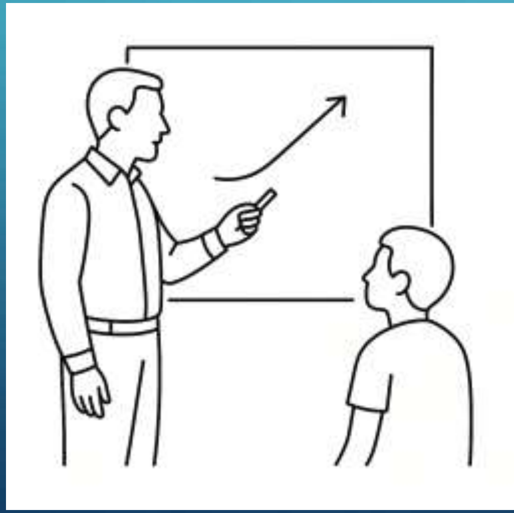
#Layers	Hidden Size	#Param (Emd)	#Param (Trm)	Inference Time
12	768	23.4M	85.1M	93.1s (1.0×)
6	768	23.4M	42.5M	46.9s (2.0×)
12	384	11.7M	21.3M	34.8s (2.7×)
6	384	11.7M	10.6M	17.7s (5.3×)
4	384	11.7M	7.1M	12.0s (7.8×)
3	384	11.7M	5.3M	9.2s (10.1×)

- Emd: Embedding -Trm: Transformer

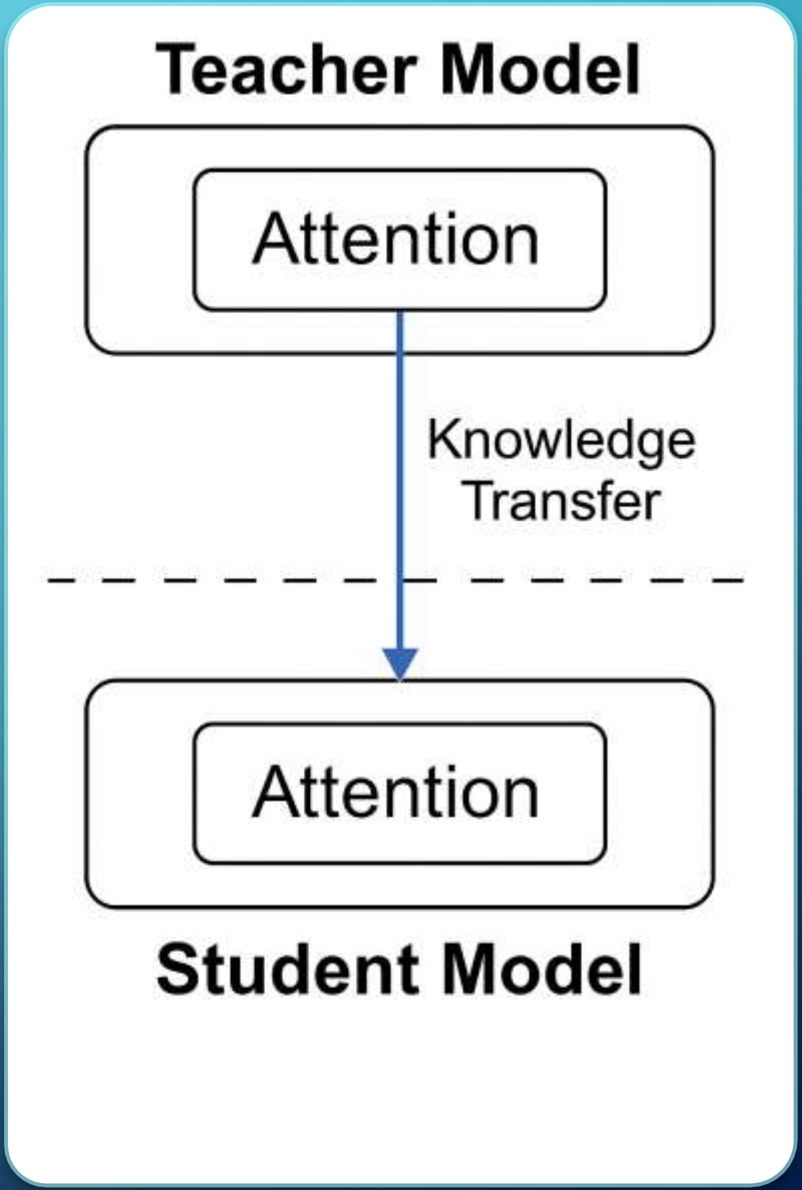
DESTILACIÓN DE CONOCIMIENTO

- Objetivo: Comprimir Transformers pre-entrenados
- Deep self-attention distillation

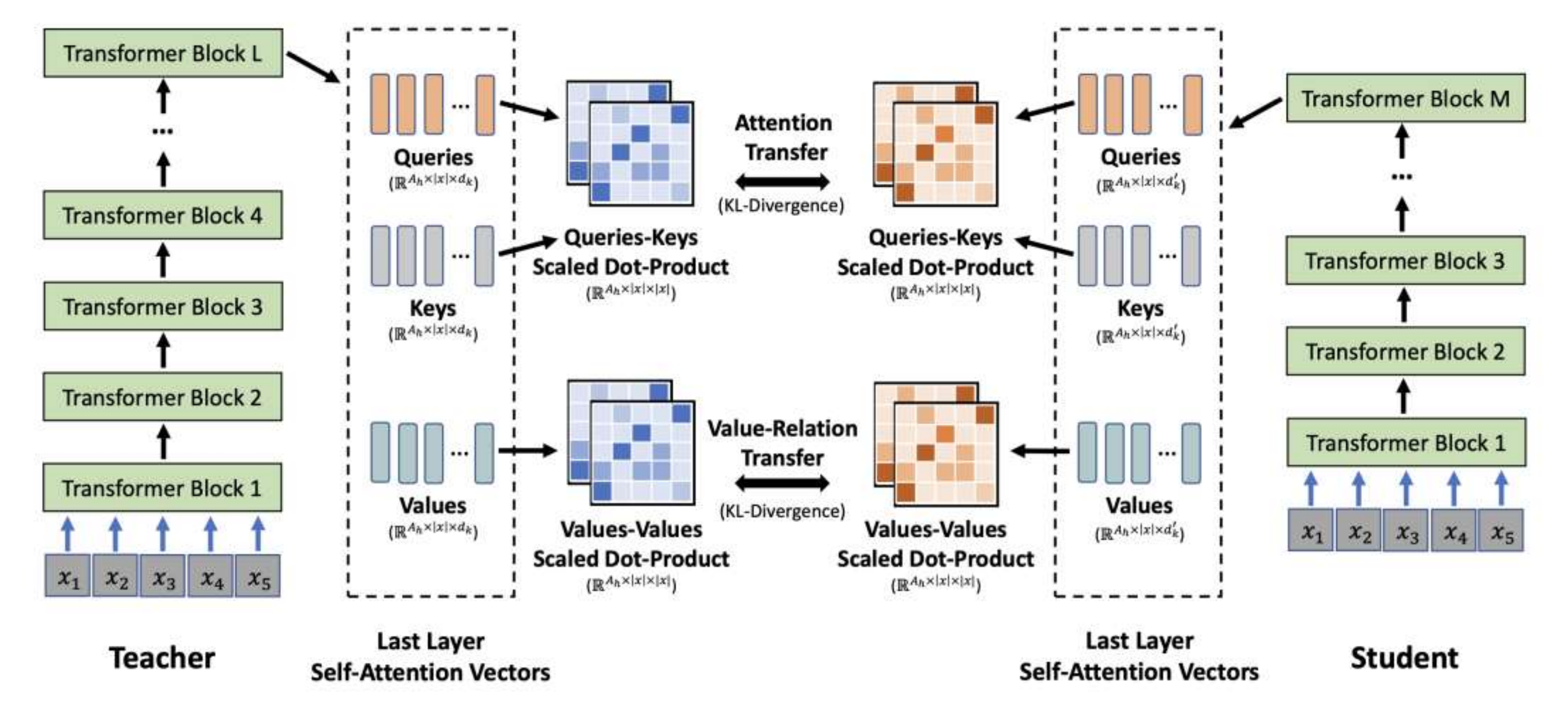
Teacher



Student

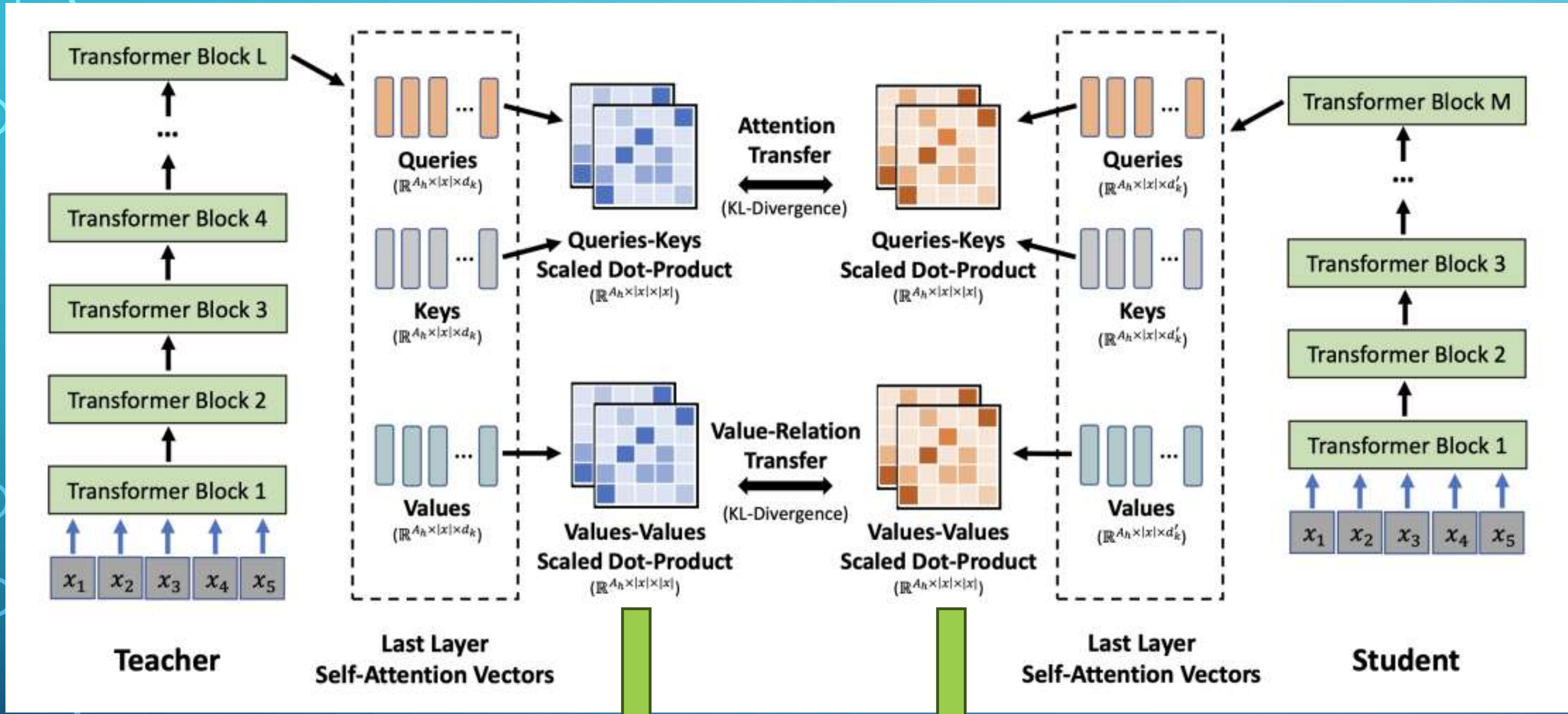


DESTILACIÓN DE LA ATENCIÓN



Confidential - Oracle Restricted

DESTILACIÓN DE LA ATENCIÓN



$$\mathbf{VR}^T_{L,a} = \mathrm{softmax}\left(\frac{\mathbf{V}^T_{L,a}\mathbf{V}^{T\intercal}_{L,a}}{\sqrt{d_k}}\right)$$

$$\mathbf{VR}^S_{M,a} = \mathrm{softmax}\left(\frac{\mathbf{V}^S_{M,a}\mathbf{V}^{S\intercal}_{M,a}}{\sqrt{d'_k}}\right)$$

Función de costo:

1) Divergencia KL entre distribuciones de self-attention entre maestro y estudiante.

$$\mathcal{L}_{\text{AT}} = \frac{1}{A_h|x|}\sum_{a=1}^{A_h}\sum_{t=1}^{|x|} D_{KL}(\mathbf{A}^T_{L,a,t} \parallel \mathbf{A}^S_{M,a,t})$$

2) Divergencia KL de self-attention values-values

$$\mathcal{L}_{\text{VR}} = \frac{1}{A_h|x|}\sum_{a=1}^{A_h}\sum_{t=1}^{|x|} D_{KL}(\mathbf{VR}^T_{L,a,t} \parallel \mathbf{VR}^S_{M,a,t})$$

Confidential - Oracle Restricted

COMPARACIÓN CON PROPUESTAS ANTERIORES

Approach	Teacher Model	Distilled Knowledge	Layer-to-Layer Distillation	Requirements on the number of layers of students	Requirements on the hidden size of students
DistillBERT	$\text{BERT}_{\text{BASE}}$	Soft target probabilities Embedding outputs			✓
TinyBERT	$\text{BERT}_{\text{BASE}}$	Embedding outputs Hidden states Self-Attention distributions	✓		
MobileBERT	$\text{IB-BERT}_{\text{LARGE}}$	Soft target probabilities Hidden states Self-Attention distributions	✓	✓	✓
MiniLM	$\text{BERT}_{\text{BASE}}$	Self-Attention distributions Self-Attention value relation			

Confidential - Oracle Restricted

FINE TUNING EN SQUAD2 Y GLUE

Model	#Param	SQuAD2	MNLI-m	SST-2	QNLI	CoLA	RTE	MRPC	QQP	Average
$\text{BERT}_{\text{BASE}}$	109M	76.8	84.5	93.2	91.7	58.9	68.6	87.3	91.3	81.5
DistillBERT	66M	70.7	79.0	90.7	85.3	43.6	59.9	87.5	84.9	75.2
TinyBERT	66M	73.1	83.5	91.6	90.5	42.8	**72.2**	**88.4**	90.6	79.1
MINILM	66M	**76.4**	**84.0**	**92.0**	**91.0**	**49.2**	71.5	**88.4**	**91.0**	**80.4**

- Destilaciones con 6 layers y 768 hidden dim
- MiniLM: 50% más ligero, retiene 99% del accuracy en diferentes tasks

Confidential - Oracle Restricted

TEACHER ASSISTANT

- Modelo de tamaño intermedio entre el maestro y el estudiante.
- Guía el entrenamiento del estudiante.

Architecture	#Param	Model	SQuAD 2.0	MNLI-m	SST-2	Average
M=6;d'_h=384	22M	MLM-KD (Soft-Label Distillation)	67.9	79.6	89.8	79.1
		TinyBERT	71.6	81.4	90.2	81.1
		MiniLM	72.4	82.2	91.0	81.9
		MiniLM (w/ TA)	**72.7**	**82.4**	**91.2**	**82.1**
M=4;d'_h=384	19M	MLM-KD (Soft-Label Distillation)	65.3	77.7	88.8	77.3
		TinyBERT	66.7	79.2	88.5	78.1
		MiniLM	69.4	80.3	90.2	80.0
		MiniLM (w/ TA)	**69.7**	**80.6**	**90.6**	**80.3**
M=3;d'_h=384	17M	MLM-KD (Soft-Label Distillation)	59.9	75.2	88.0	74.4
		TinyBERT	63.6	77.4	88.4	76.5
		MiniLM	66.2	78.8	89.3	78.1
		MiniLM (w/ TA)	**66.9**	**79.1**	**89.7**	**78.6**

TA: Teacher Assistant

KEYTAKE AWAYS

- **Distilación centrada en atención:** MiniLM solo destila las distribuciones de self-attention de la última capa y las relaciones entre vectores *value*.
- **Flexibilidad** en la arquitectura del estudiante.
- **Teacher assistant intermedio**
- < ½ de parámetros, y buena precisión.

GRACIAS

- Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, *33*, 5776-5788.