A Light Transformer-Based Architecture for Handwritten Text Recognition

Killian Barrere Yann Soullard Aurélie Lemaitre Bertrand Coüasnon

Document Analysis Systems (DAS), La Rochelle 24<sup>th</sup> May 2022



IntuiDoc research team, Univ. Rennes, CNRS, IRISA, France

Introd ●0	uction Our Archi 000		Experiments 0000000	Conclusion 0
Sta	ate of the Art: From RN	N to Transformer		
	Usual approaches: Convolutior	al Recurrent Neural Net	works (CRNN)	
	• Convolutional layers + ree	current layers		
	$\Rightarrow$ Lack of parallelism / slow	raining speed		

	$\Rightarrow$ Lack of parallelism / slow training s	peed	
	• Convolutional layers + recurrent lay	rers	
	Usual approaches: Convolutional Recurre	ent Neural Networks (CRNN)	
S	ate of the Art: From RNN to Tra	ansformer	
•0	000 Our Architecture	0000000	o

Fully Convolutional Networks [Ingle et al. 2019, Yousef et al. 2020, Coquenet et al. 2021]

• Composed of convolutional layers, no recurrent layers

 $\Rightarrow$  Faster training speed, but might be hard to learn long-range contexts

	$\Rightarrow$ Lack of parallelism / slow training sp	eed	
	• Convolutional layers + recurrent laye		
	Usual approaches: Convolutional Recurren	nt Neural Networks (CRNN)	
S	ate of the Art: From RNN to Tra	nsformer	
•0	000	0000000	0

Fully Convolutional Networks [Ingle et al. 2019, Yousef et al. 2020, Coquenet et al. 2021]

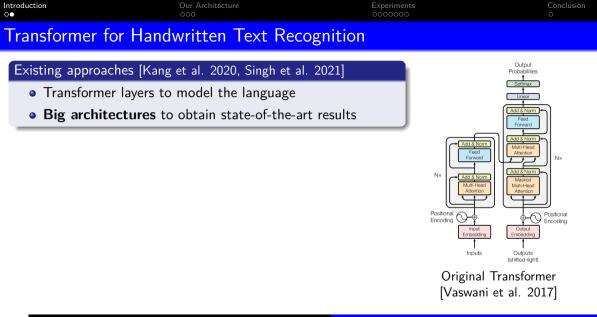
• Composed of convolutional layers, no recurrent layers

 $\Rightarrow$  Faster training speed, but might be hard to learn long-range contexts

#### Multi-Head Attention (Transformer layers) [Vaswani et al. 2017]

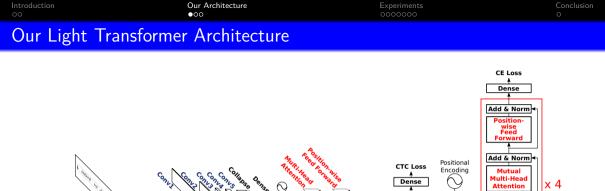
- Able to learn long-range context
- Strong parallelism
- $\Rightarrow$  Good alternative but require a lot of training data

Introduction



Introduction ⊙●	Our Architecture	Experiments 0000000	Conclus O	
Transformer for H	andwritten Text R	ecognition		
• Transformer laye	Kang et al. 2020, Singh e rs to model the languag <b>es</b> to obtain state-of-the	e	Output Probabilities Softmax Linear Add 5 Norm Food	
• Few annotated d	<b>data</b> to be trained lata in handwritten reco <b>ata</b> to perform well	gnition (10k lines)	Nx Positional Encoding Prot	
			Criginal Transformer [Vaswani et al. 2017]	

Introduction ⊙●	Our Architecture 000	Experimen 0000000		Conclusion O
Transformer for Hand	lwritten Text Re	cognition		
<ul><li>Existing approaches [Kang</li><li>Transformer layers to</li><li>Big architectures to</li></ul>	o model the language	2	Out; Probab Softmer Lister Foo	
<ul><li>Problem</li><li>Require a lot of date</li></ul>		witting (101, lines)	Add & Norm         Add & Norm           Feed         Forward           Nx         Add & Norm           Multi-Med Add & Norm         Multi-Med Multi-Med Add & Norm	Nx Norm
<ul> <li>Few annotated data</li> <li>⇒ Additional data</li> </ul>	-	nition (10k lines)	Positional Encoding	Positional Encoding
Our proposition			Inputs Outputs (shifted	
<ul><li>Light architecture</li><li>Hybrid loss to ease t</li></ul>	-	h few data	Original Transfo [Vaswani et al.	



128

"<sos>A MOVE to stop Mr. Gaitskell from" -

Target Transcription (shifted right)

when training

with teacher forcing

Positional

Encoding

Characters already predicted

when testing / training

without teacher forcing

Transformer

**Encoder Lavers** 

x 4

Character

Embedding

Add & Norm

Self

Multi-Head

Attention

Positional

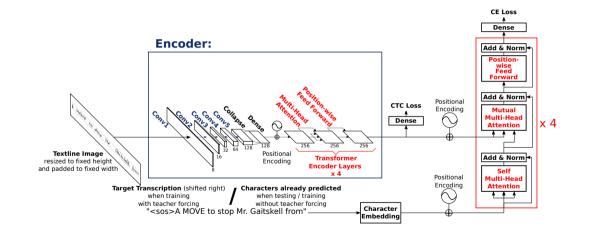
Encoding

Textline Image

resized to fixed height

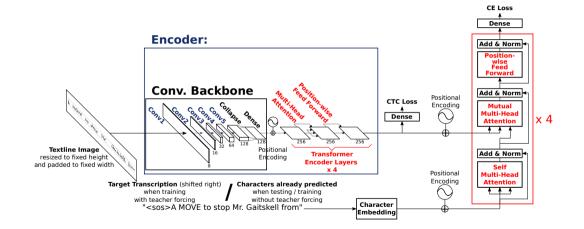
and padded to fixed width





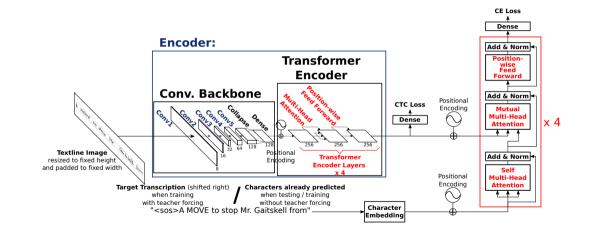
Our Light Transformer	· • • • • • • • • • • • • • • • • • • •		
	•00		
Introduction	Our Architecture	Experiments	Conclusion



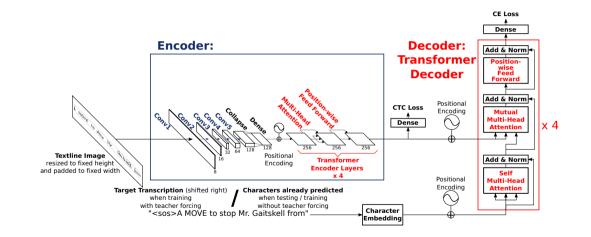


Introduction	Our Architecture	Experiments	Conclusion
00	•00	0000000	0

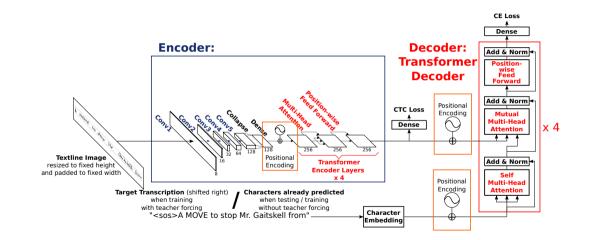




Introduction	Our Architecture	Experiments	Conclusion
00	●00	0000000	O
Our Light Tra	nsformer Architecture		

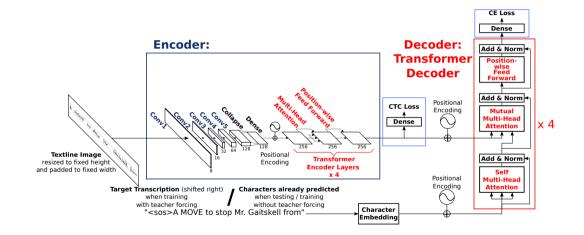


Introduction	Our Architecture	Experiments	Conclusion
00	●00	0000000	O
Our Light Tra	nsformer Architecture		



	<b>●</b> 00		
Introduction	Our Architecture	Experiments	Conclusion



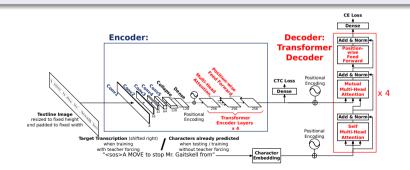


A Light Architecture	Introduction	Our Architecture	Experiments	Conclusion
	00	○●○	0000000	O
	A Light Archite	ecture		

#### How to make a smaller Transformer

## Convolutional backbone

## Big backbone (i.e. ResNet18)

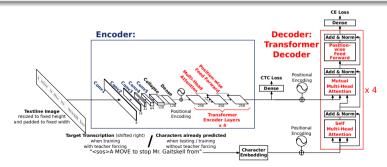


Introduction 00	Our Architecture 000	Experiments 0000000	Conclusion O
A Light Architecture			

#### How to make a smaller Transformer

#### **Convolutional backbone** Big backbone (i.e. ResNet18)

 $\Rightarrow$  Only 5 convolutional layers



Introdu 00		ur Architecture ●0		xperiments 200000	Conclusion O
Αl	ight Architecture				
	How to make a smaller T	ransformer			
	Convolutional backbon	е	Neurons i	n Transformer layer	s
	Big backbone (i.e. ResN	<del>et18)</del>	Up to 1,02	24 neurons	
	$\Rightarrow$ Only 5 convolutional	layers			
	-	ranscription (shifted right) Charact when training wh	stional reading training recording training recording training recording training recording training recording training recording training	Encoding	

00	000	experiments 0000000	o
A Ligh	nt Architecture		
How	w to make a smaller Transformer		
Big	p <b>nvolutional backbone</b> <del>g backbone (i.e. ResNet18)</del> Only 5 convolutional layers	<b>Neurons in Transformer layers</b> Up to 1,024 neurons $\Rightarrow$ Only 256 neurons	
	Excision langer rescale to fixed width and padded to fixed width Target Transcription Lichter right) Character a loc when training with teator forcing "< sos>A MOVE to stop Mr. Gaitskell fro	g / training ther forcing	

luction	000 Architecture 000	experiments	o
Light Archited	ture		
How to make a s	maller Transformer		
Convolutional	backbone	Neurons in Transformer layers	
<del>Big backbone (</del> i	.e. ResNet18)	Up to 1,024 neurons	
$\Rightarrow$ Only 5 convo	olutional layers	$\Rightarrow$ Only 256 neurons	
In total			
100M parameter	s		

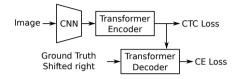
o
J

00		0000000	0
A Light Architecture			
	How to make a smaller Transformer		
	Convolutional backbone	Neurons in Transformer layers	
	Big backbone (i.e. ResNet18)	Up to 1,024 neurons	
	$\Rightarrow$ Only 5 convolutional layers	$\Rightarrow$ Only 256 neurons	
	In total	Potential benefits	
	100M parameters	• Faster to train compared to other	

 $\Rightarrow$  6.9M parameters

- Faster to train compared to other Transformer-based architecture
- Does not require additional data to be trained efficiently

Introduction	Our Architecture	Experiments	Conclusion
00		0000000	O
Hybrid Loss			

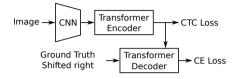


$$\mathcal{L} = \lambda \cdot \mathcal{L}_{\textit{CTC}} + (1 - \lambda) \cdot \mathcal{L}_{\textit{CE}}$$

## Hybrid loss [Michael et al. 2019]

- Connectionist Temporal Classification (CTC) for the Encoder
- Cross Entropy (CE) for the Decoder

Introduction	Our Architecture	Experiments	Conclusion
00	000		
Hybrid Loss			
TYDNU LOSS			



$$\mathcal{L} = \lambda \cdot \mathcal{L}_{\textit{CTC}} + (1 - \lambda) \cdot \mathcal{L}_{\textit{CE}}$$

## Hybrid loss [Michael et al. 2019]

- Connectionist Temporal Classification (CTC) for the Encoder
- Cross Entropy (CE) for the Decoder

#### Potential benefits

- Help to train deep layers with gradients from both losses
- Faster convergence

Introduction	Our Architecture	Experiments	Conclusion
00	000	●000000	O
Outline of the E	xperiments		

#### Experiments presented

- Ablation Study
  - Transformer layers
  - Decoder
- Architecture Size
- Hybrid loss
- Omparison with state-of-the-art methods

Introduction	Our Architecture	Experiments	Conclusion
00	000	000000	
Outline of the	Experiments		

#### Experiments presented

- Ablation Study
  - Transformer layers
  - Decoder
- Architecture Size
- O Hybrid loss
- Omparison with state-of-the-art methods

• Results with and without synthetic data (to compare fairly with others)

Introduction	Our Architecture	Experiments	Conclusion
00	000	o●ooooo	0
Data used			

#### Real data, without additional data

- IAM dataset (modern English, 10,363 lines, 76k words)
- Data augmentation techniques

a sit earthbourd foright, Front. I've got a queer there have been only two occasions on

OO	Our Architecture	Experiments 000000	O
Data used			
Destate the State	an a data ta sa babara		

#### Real data, without additional data

G

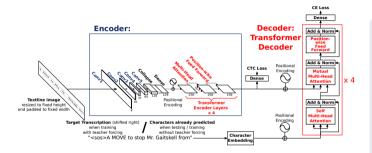
- IAM dataset (modern English, 10,363 lines, 76k words)
- Data augmentation techniques 5:4

Our synthetic data (to compare with other transformers)

- Articles from Wikipedia (21,350 articles, 66M words)
- Handwritten fonts (32 fonts)
- Random deformations / augmentations

OU

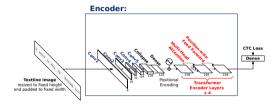
Introduction 00	Our Architecture 000	Experiments 000000	Conclusion O
Ablation Study:	Transformer Layers i	nstead of Recurrent Layers	5



#### Without the decoder

Introduction	Our Architecture	Experiments	Conclusion

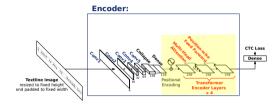




#### Without the decoder

		estand of Document Low	
Introduction	Our Architecture	Experiments	Conclusion

## Ablation Study: Transformer Layers instead of Recurrent Layers

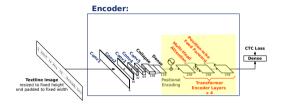


#### Without the decoder

 CRNN with Transformer instead of recurrent layers

	<b>— — — — — —</b>	stand of Decument Low	
Introduction	Our Architecture	Experiments	Conclusion

## Ablation Study: Transformer Layers instead of Recurrent Layers



Architecture	# params CE	14	١M	IAM + S	ynth. Data
		CER (%)	WER (%)	CER (%)	WER (%)
CRNN (Baseline) Our Encoder only	1.7M 3.2M	6.14 5.93	23.26 22.82		

#### Without the decoder

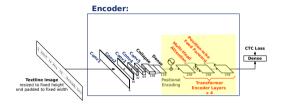
 CRNN with Transformer instead of recurrent layers

#### $\textbf{Recurrent} \Rightarrow \textbf{Transformer}$

- More parameters
- Lower error rates
  - Better context

	<b>— — — — — —</b>	stand of Decument Low	
Introduction	Our Architecture	Experiments	Conclusion

## Ablation Study: Transformer Layers instead of Recurrent Layers



Architecture	# params Cl	IAM		IAM+Synth. Data	
		CER (%)	WER (%)	CER (%)	WER (%)
CRNN (Baseline) Our Encoder only	1.7M 3.2M	6.14 5.93	23.26 22.82	5.66 6.15	21.62 24.02

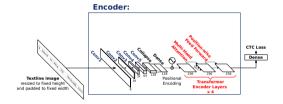
#### Without the decoder

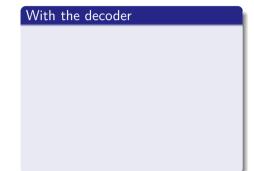
 CRNN with Transformer instead of recurrent layers

#### $Recurrent \Rightarrow Transformer$

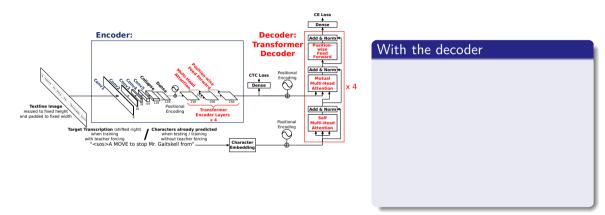
- More parameters
- Lower error rates
  - Better context
- Worse with synthetic data (may not generalize well)

Introduction	Our Architecture	Experiments	Conclusion
00	000	0000000	O
Ablation Study	: Decoder		

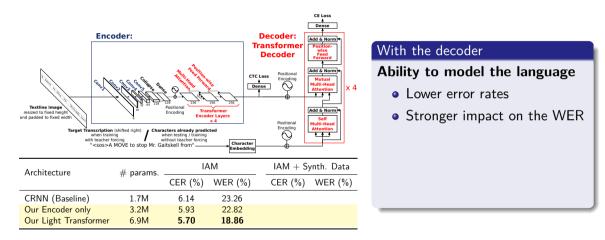




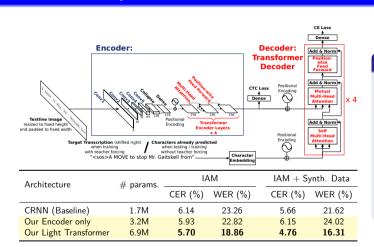
Introduction 00	Our Architecture	Experiments	Conclusion
Ablation Study:			



Introduction	Our Architecture	Experiments	Conclusion
00	000	000●000	O
Ablation Study	: Decoder		



Introduction	Our Architecture	Experiments	Conclusion
00		0000000	O
Ablation Study:	Decoder		



#### With the decoder

#### Ability to model the language

- Lower error rates
- Stronger impact on the WER

# Benefits more from synthetic data

• More data to learn the language

Densfire of Hoten of Links Analytic struct		
Introduction     Our Architecture       00     000	Experiments 0000000	Conclusion O

## Benefits of Using a Light Architecture

Different sizes of our architecture

- Light Transformer: 6.9M params.
- Large Transformer: 28M params.

Introduction	Our Architecture	Experiments	Conclusion

# Benefits of Using a Light Architecture

## Different sizes of our architecture

- Light Transformer: 6.9M params.
- Large Transformer: 28M params.

Architecture	IAM		IAM + Synth. Data	
, a cintecture	CER (%) WER (%)		CER (%)	WER (%)
Our Light Transformer	<b>5.70</b>	<b>18.86</b>	<b>4.76</b>	<b>16.31</b>
Our Large Transformer	5.79	19.67	4.87	17.67

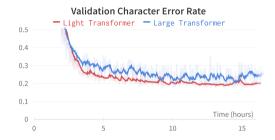
# • Our light architecture is **competitive**

Ponofite of Using a Light Architecture	
Introduction         Our Architecture         Experiments           oo         ooo         oooo●oo	Conclusion O

# Benefits of Using a Light Architecture

## Different sizes of our architecture

- Light Transformer: 6.9M params.
- Large Transformer: 28M params.



Architecture	IAM		IAM + Synth. Data	
/ i entecture	CER (%) WER (%)		CER (%)	WER (%)
Our Light Transformer	5.70	18.86	4.76	16.31
Our Large Transformer	5.79	19.67	4.87	17.67

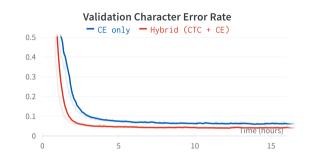
- Our light architecture is **competitive**
- Our light architecture might be **trained faster**

Introduction	Our Architecture	Experiments	Conclusion
00	000	00000€0	O
Interest of the Hybr	rid Loss		

- **CE only**: Cross-Entropy loss after the decoder
- Hybrid: CTC after the encoder and CE after the decoder

Introduction	Our Architecture	Experiments	Conclusion
00	000	00000€0	O
Interest of the H	lybrid Loss		

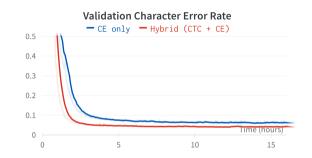
- **CE only**: Cross-Entropy loss after the decoder
- Hybrid: CTC after the encoder and CE after the decoder



• Faster convergence

Introduction	Our Architecture	Experiments	Conclusion
00		00000€0	O
Interest of the I	Hybrid Loss		

- **CE only**: Cross-Entropy loss after the decoder
- Hybrid: CTC after the encoder and CE after the decoder



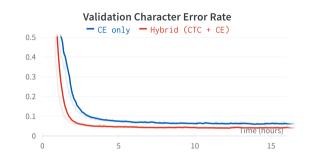
Loss	IAM		$IAM + S_{2}$	/nth. Data
Function(s)	CER (%)	WER (%)	CER (%)	WER (%)
CE only	10.29	26.36		
Hybrid (CTC $+$ CE)	5.70	18.86		

- Faster convergence
- Crucial with few data

Introduction	Our Architecture	Experiments	Conclusion
00		00000€0	O
Interest of the I	Hybrid Loss		

Killian Barrere

- **CE only**: Cross-Entropy loss after the decoder
- Hybrid: CTC after the encoder and CE after the decoder



Loss	IAM CER (%) WER (%)		IAM + Synth. Data	
Function(s)			CER (%)	WER (%)
CE only	10.29	26.36	6.76	19.62
Hybrid (CTC + CE)	5.70	18.86	4.76	16.31

- Faster convergence
- Crucial with few data
- Important with synthetic data

Introdu 00	ction Our Architecture		Experiments 000000●	Conclusion O
Cor	nparison with the state of the ar	t		
	Model Encoder	# params.	IAM CER (%)	IAM + Synth. Data CER (%)
	CRNN + LSTM [Michael et al. 2019]		5.24	
	FCN [Yousef et al. 2020]	3.4M	4.9	
	VAN (line level) [Coquenet et al. 2022]	1.7M	4.95	
	Transformer [Kang et al. 2020]	100M	7.62	4.67
	FPHR Transformer [Singh et al. 2021]	28M		6.5
	Forward Transformer [Wick et al. 2021]	13M	6.03	
	Bidi. Transformer [Wick et al. 2021]	27M	5.67	
	Our Light Transformer-based	6.9M	5.70	4.76

Introdu 00	ction Our Architecture		Experiments 000000●	Conclusion O
Cor	nparison with the state of the ar	t		
	Model Encoder	# params.	IAM CER (%)	IAM + Synth. Data CER (%)
	CPNN - LETM [Michael at al 2019]		5.24	
	Compared with other Transformers:	3.4M	4.9	
		1.7M	4.95	
	Transformer [Kang et al. 2020]	100M	7.62	4.67
	FPHR Transformer [Singh et al. 2021]	28M		6.5
	Forward Transformer [Wick et al. 2021] Bidi. Transformer [Wick et al. 2021]	13M 27M	6.03 5.67	
	Our Light Transformer-based	6.9M	5.70	4.76

Introdu 00	ction Our Architecture		Experiments 000000●	Conclusion O
Cor	mparison with the state of the ar	t		
	Model Encoder	# params.	IAM CER (%)	IAM + Synth. Data CER (%)
	CRNN + LSTM [Michael et al. 2019]		F 04	
	FCN [Yousef et al. 2020]	3 Low error rates without additional data		
	VAN (line level) [Coquenet et al. 2022]	1.		uutu
	Transformer [Kang et al. 2020]	100M	7.62	4.67
	FPHR Transformer [Singh et al. 2021]	28M		6.5
	Forward Transformer [Wick et al. 2021] Bidi. Transformer [Wick et al. 2021]	13M 27M	6.03 5.67	
	Our Light Transformer-based	6.9M	5.70	4.76

Killian Barrere Light Transformer for Handwritten Text Recognition

Introdu 00	ction Our Architecture		Experiments 000000●	Conclusion O
Cor	mparison with the state of the ar	t		
	Model Encoder	# params.	IAM CER (%)	IAM + Synth. Data CER (%)
	CRNN + LSTM [Michael et al. 2019]		5.24	
	FCN [Yousef et al. 2020]	3.4M	4.9	ate-of-the-art results with synthetic data
	VAN (line level) [Coquenet et al. 2022]	1.7M	4.95	June 1 Synthetic Lata
	Transformer [Kang et al. 2020]	100M	7.62	4.67
	FPHR Transformer [Singh et al. 2021]	28M		6.5
	Forward Transformer [Wick et al. 2021] Bidi. Transformer [Wick et al. 2021]	13M 27M	6.03 5.67	
	Our Light Transformer-based	6.9M	5.70	4.76

Killian Barrere Light Transformer for Handwritten Text Recognition

Introdu 00	ction Our Architecture		Experiments 000000●	Conclusion O
Cor	nparison with the state of the art	t		
	Model Encoder	# params.	IAM CER (%)	IAM + Synth. Data CER (%)
	CRNN + LSTM [Michael et al. 2019]		5.24	
	EVIN LYOUSEL ET AL ZUZUI	While being <b>ht Transforn</b>	4.9	
	VAN (line level) [Coquenet et al. 20		4.95	
	Transformer [Kang et al. 2020]	100M	7.62	4.67
	FPHR Transformer [Singh et al. 2021]	28M		6.5
	Forward Transformer [Wick et al. 2021] Bidi. Transformer [Wick et al. 2021]	13M 27M	6.03 5.67	
	Our Light Transformer-based	6.9M	5.70	4.76

Killian Barrere Light Transformer for Handwritten Text Recognition

Introdu 00	ction Our Architecture		Experiments 000000●	Conclusion O
Cor	mparison with the state of the ar	t		
	Model Encoder	# params.	IAM CER (%)	IAM + Synth. Data CER (%)
	CRNN + LSTM [Michael et al. 2019]		5.24	
	FCN [Yousef et al. 2020]	3.4M	4.9	
	VAN (line level) [Coquenet et al. 2022]	1.7M	4.95	
	Transformer [Kang et al. 2020]	100M	7.62	4.67
	FPHR Transformer [Singh et al. 2021]	28M		6.5
	Forward Transformer [Wick et al. 2021]	13M	6.03	
	Bidi. Transformer [Wick et al. 2021]	27M	5.67	
	Our Light Transformer-based	6.9M	5.70	4.76

Introduction	Our Architecture	Experiments	Conclusion
			•
Conclusion			

## Our Contribution

A light Transformer architecture, trained with a hybrid loss

- Faster to train than other Transformers
- Good results without additional data
- State-of-the-art results with synthetic data

Introduction	Our Architecture	Experiments	Conclusion
			•
Conclusion			

### Our Contribution

A light Transformer architecture, trained with a hybrid loss

- Faster to train than other Transformers
- Good results without additional data
- State-of-the-art results with synthetic data

#### Future Works: Historical Documents

- Ability of Transformers to model the language is crucial
- Very few annotated data  $\Rightarrow$  our light Transformer architecture