EMNLP 2022

VisToT: Vision-Augmented Table-to-Text Generation

Prajwal Gatti[1], Anand Mishra[1],
Manish Gupta[2], Mithun Das Gupta[2]

[1]IIT Jodhpur, [2]Microsoft

Can we combine the information in **tables** and **images** to generate rich text descriptions?

What is Table-to-Text?

Lough Leane	
Location	Killarney, County Kerry
Coordinates	52°2′30″N 9°33′0″W
Basin countries	Ireland
Surface area	4,700 acres (19 km^2)
Islands	Innisfallen

Lough Leane is a 4700 acre estate in Killarney, County Kerry, Ireland.

Lebret et al. EMNLP 2016; Parikh et al. EMNLP 2020

What is Vision-Augmented Table-to-Text?

Lough Leane	
Location	Killarney, County Kerry
Coordinates	52°2′30″N 9°33′0″W
Basin countries	Ireland
Surface area	4,700 acres (19 km^2)
Islands	Innisfallen

Lough Leane is a large lake in Killarney, County Kerry, Ireland.

What is **Vision-Augmented** Table-to-Text?

Given a *table **T*** and an *image **I*** about an entity

Generate a *text summary **S*** describing the entity using **T** and **I** as the source context.

We introduce **WikiLandmarks**

A new dataset containing tables and images for **73K** world landmarks

Image

Table

Name	Amitabha Drukpa
Country	Nepal
Location	Kathmandu
Dedicated To	Amitabha

Text summary

"Amitabha Monastery is a Tibetan Buddhist Monastery in Nepal"

Image

Table

Name	Michigan Stadium
Location	1201 South Main Street Ann Arbor, Michigan
Owner	University of Michigan
Nickname	The Big House

Text summary

*"Michigan Stadium, nicknamed The Big House, is **the football stadium** for the University of Michigan in Ann Arbor, Michigan"*

Image

Table

Name	Niesen
Elevation	2,632 m
Prominence	407 m
Location	Canton of Bern, Switzerland
Parent Range	Bernese Alps

Text summary

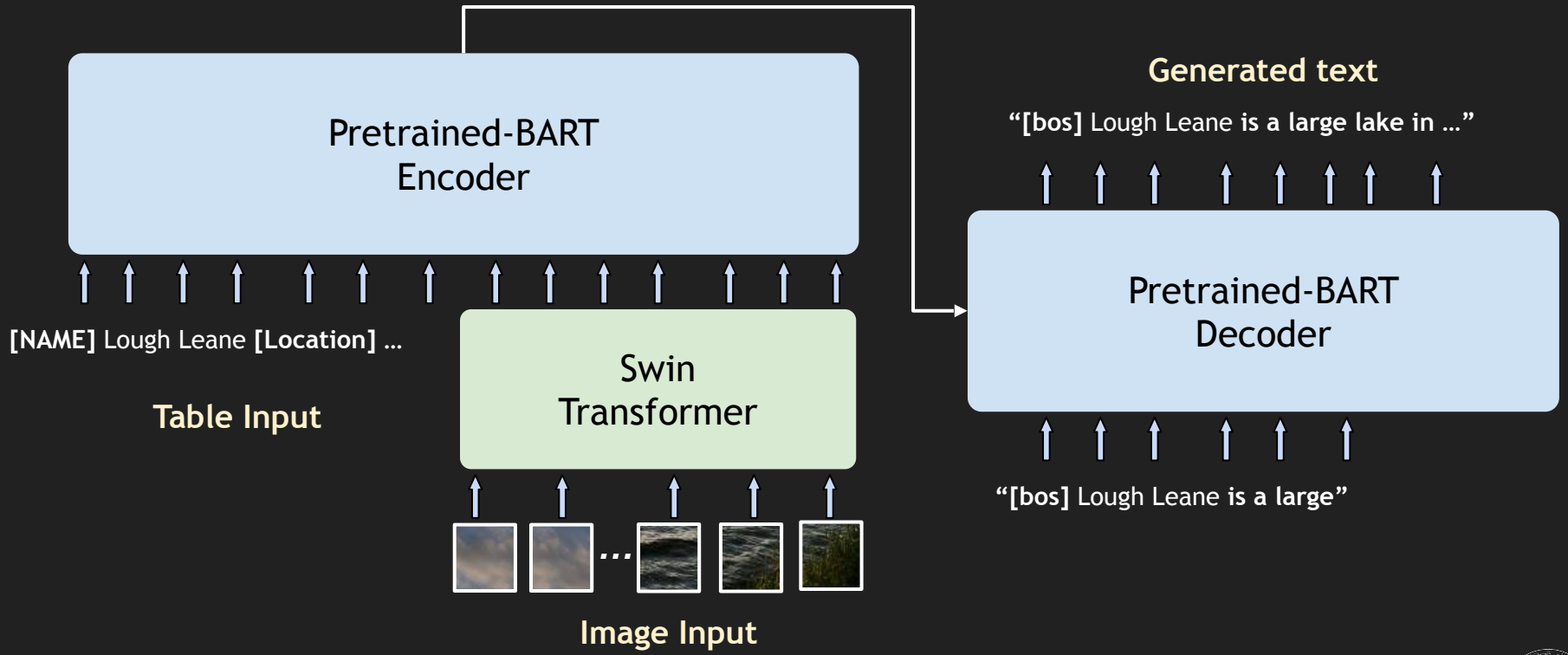
*"The Niesen is a **mountain peak** of the Bernese Alps in the Canton of Bern, Switzerland".*

Dataset Statistics

- Samples for 73K+ unique World landmarks
- ~10 natural images available per landmark
- Includes churches, mountains peaks, castles, historical sites, statues, etc.
- Curated from Wikipedia and Google Landmarks Dataset (Weyand et al. 2020)

A new multimodal transformer based approach

Our approach



VT3: Visual-Tabular Data-to-Text Transformer

Results

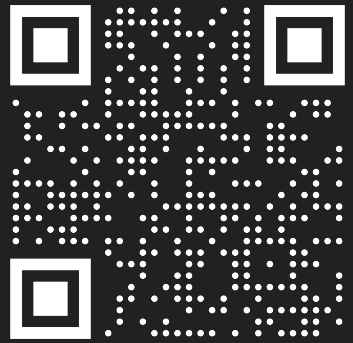
Method	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BLEURT
Image captioning-based						
PureT	6.4	26.1	33.2	12.8	31.1	0.40
Table-to-Text						
Pointer-Generator	17.8	39.2	51.6	31.7	49.2	0.50
BERT-to-BERT	22.1	43.9	55.3	35.6	53.1	0.50
T5	25.8	48.1	58.8	38.8	57.0	0.54
PlanGen	8.6	20.6	32.5	20.2	31.9	0.49
Visual-Tabular Data-to-text						
LSTM+ResNet50	6.5	19.8	31.0	19.1	30.3	0.39
VisualBERT+BERT	26.1	49.0	60.4	39.2	58.8	0.54
VT3	**30.2**	**53.5**	**62.9**	**43.4**	**60.8**	**0.56**

EMNLP 2022

Thank you!

Visit our project page for Code and Dataset

https://vl2g.github.io/projects/vistot/



॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥

Microsoft