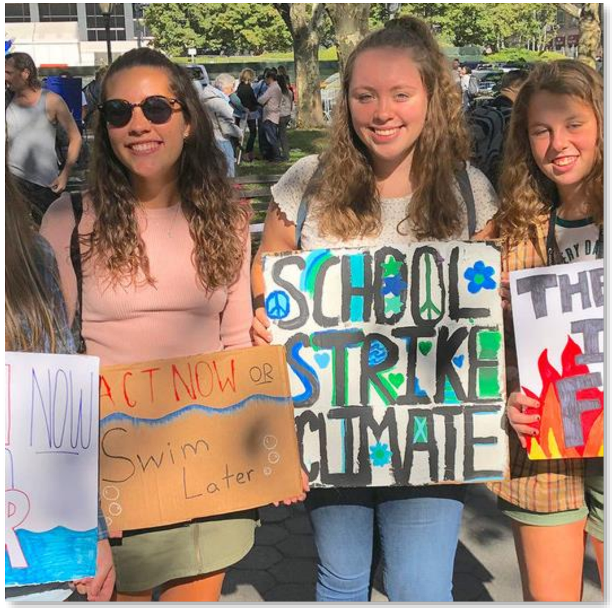
A Multilingual Approach to Scene Text Visual Question Answering

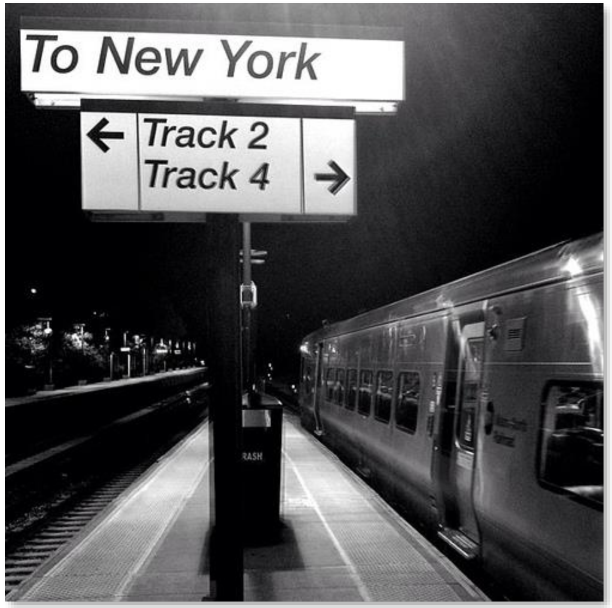
Josep Brugués i Pujolràs
Lluís Gómez i Bigordà,
Dimosthenis Karatzas

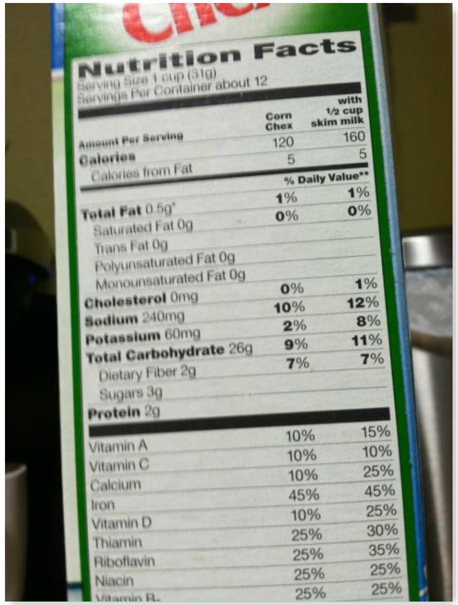
Centre de Visió per Computador

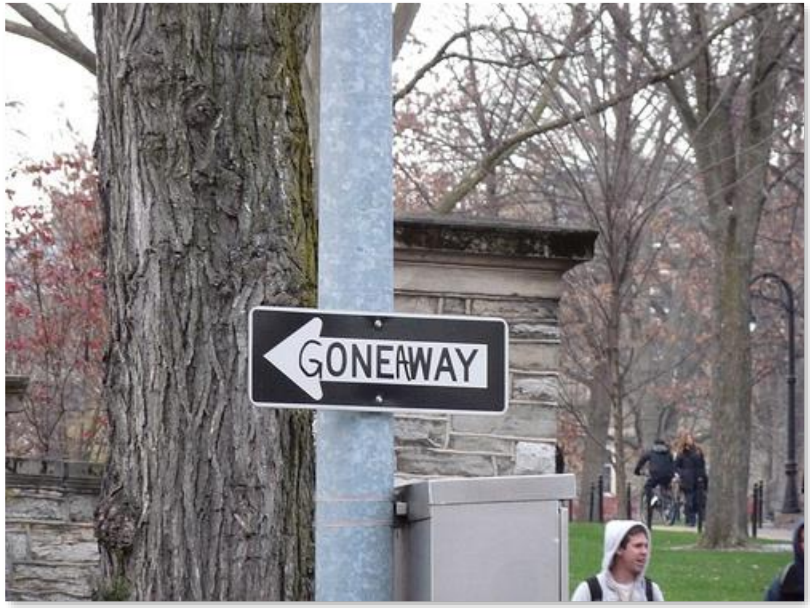
May 2022

Scene text VQA









What is written on the banner of the girl wearing sunglasses?

Act now or swim later

Where is the train going?

To New York

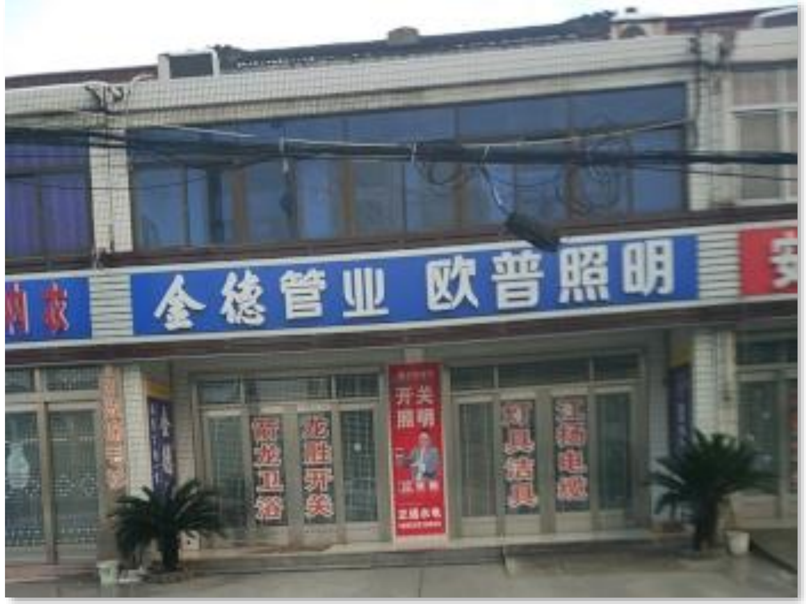
How many calories are from fat?

5

What did the sign say originally?

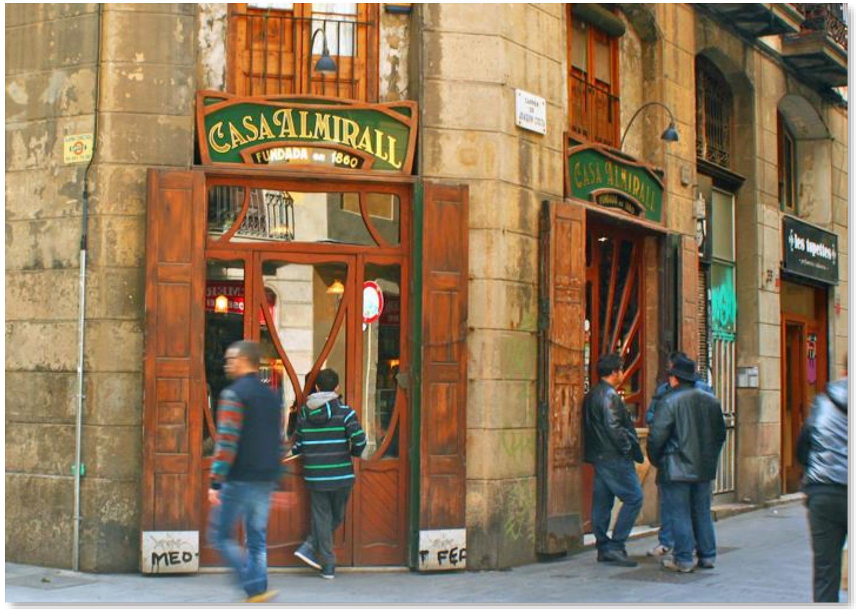
ONE WAY

Multilingual Scene Text VQA?



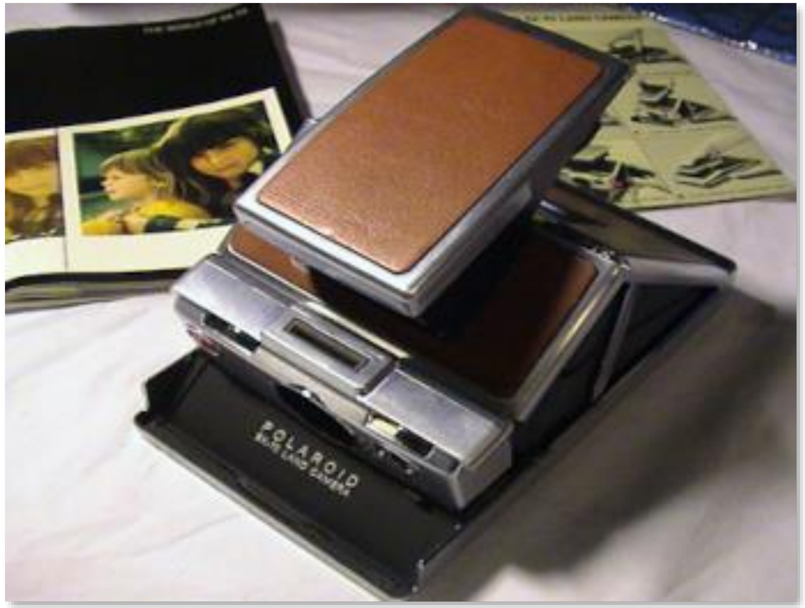
What is the name of the shop on the right?

欧普照明 (OP Lighting)



Πότε ιδρύθηκε η Casa Almirall?

1860



Quina és la marca de la càmera?

Polaroid

Scene text for VQA

	VizWiz	**ST-VQA**	**TextVQA**	**EST-VQA**
Year	2018	2019	2019	2020
Languages	English	English	English	Chinese, English
Images	31,000	23,038	28,408	25,239
Questions	70,000	31,791	45,336	28,158
Answers	58,789	31,791	453,360	28,158
Scene Text Relevance	Answer **sometimes** requires scene text	Answer **always** in the scene text	Answer **sometimes** requires scene text	Answer **always** in the scene text

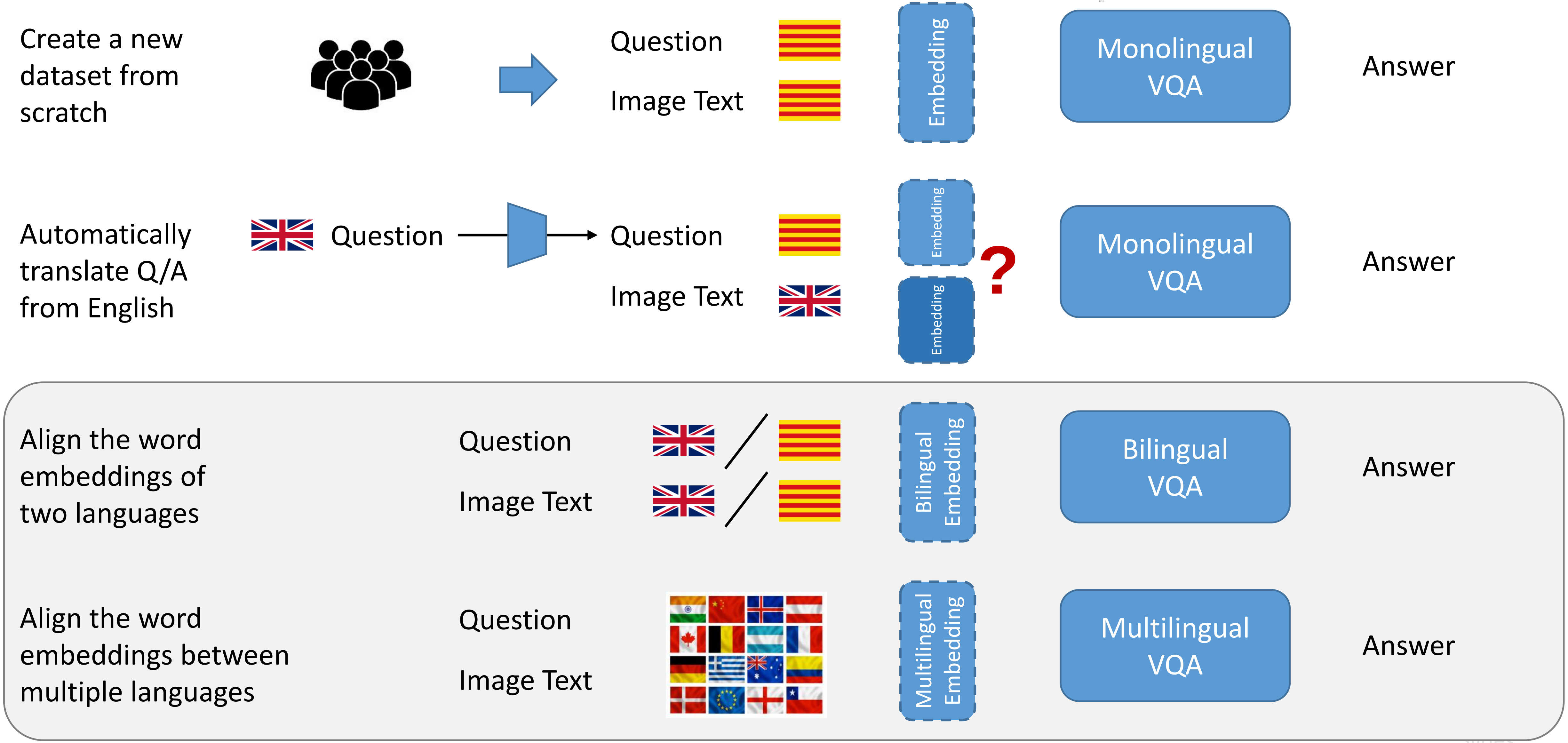
D. Gurari, Q. Li, A.J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, J.P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people" CVPR (2018)

A. Biten, R. Tito, A. Mafla, L. Gomez, M. Rusiñol, E. Valveny, C.V. Jawahar, D. Karatzas, "Scene Text Visual Question Answering", ICCV (2019)

A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, M. Rohrbach, "Towards vqa models that can read", CVPR (2019)

X. Wang, Y. Liu, C. Shen, C.C. Ng, C. Luo, L. Jin, C.S. Chan, A.v.d. Hengel, L. Wang, "On the general value of evidence, and bilingual scene-text visual question Answering" CVPR (2020)

Going Multilingual



Create a new dataset from scratch

Question

Image Text

Embedding

Monolingual VQA

Answer

Automatically translate Q/A from English

Question

Question

Image Text

Embedding

Embedding

?

Monolingual VQA

Answer

Align the word embeddings of two languages

Question

Image Text

Bilingual Embedding

Bilingual VQA

Answer

Align the word embeddings between multiple languages

Question

Image Text

Multilingual Embedding

Multilingual VQA

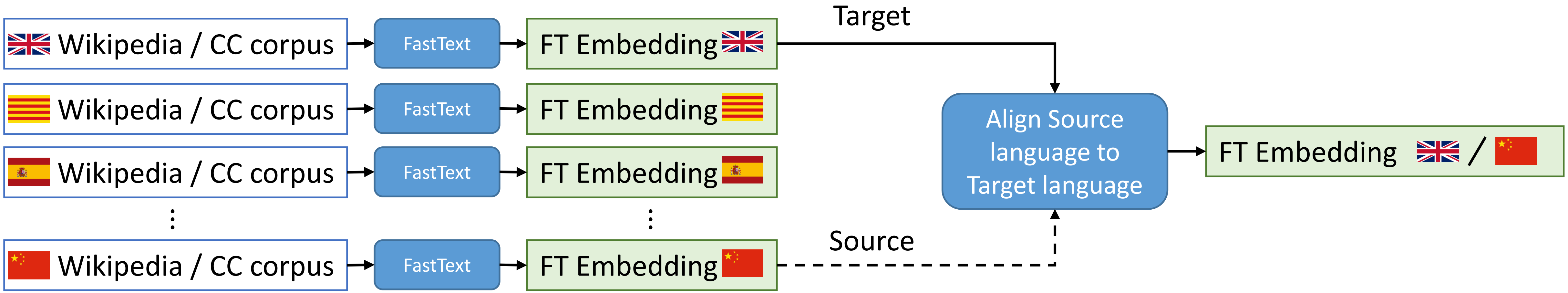
Answer

Multilingual Embeddings

Bilingual FastText Embeddings

157 × Monolingual Embeddings

Bilingual embedding



Target

Wikipedia / CC corpus → FastText → FT Embedding

Wikipedia / CC corpus → FastText → FT Embedding

Wikipedia / CC corpus → FastText → FT Embedding

⋮

Wikipedia / CC corpus → FastText → FT Embedding

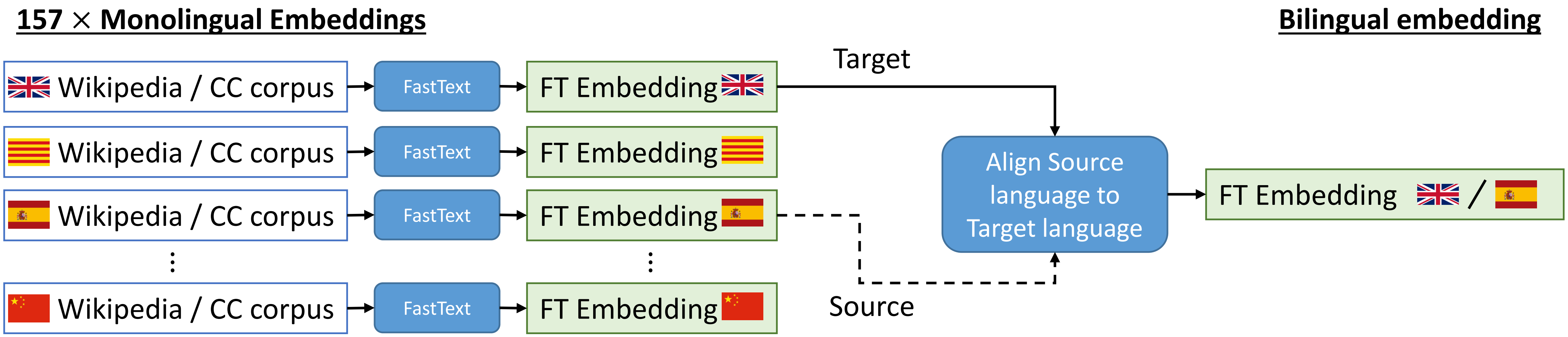
Source

Align Source language to Target language → FT Embedding /

Smith et al, “Offline bilingual word vectors, orthogonal transformations and the inverted softmax”, ICLR (2017)

Joulin et al, “Loss in translation: Learning bilingual word mapping with a retrieval criterion”, EMNLP (2018)

Bilingual FastText Embeddings



157 × Monolingual Embeddings

Bilingual embedding

Wikipedia / CC corpus → FastText → FT Embedding

Wikipedia / CC corpus → FastText → FT Embedding

Wikipedia / CC corpus → FastText → FT Embedding

⋮

Wikipedia / CC corpus → FastText → FT Embedding

Target

Source

Align Source language to Target language

FT Embedding /

Smith et al (ICLR 17):

- Maximize the cosine similarity of translation pairs, subject to the mapping between semantic spaces being orthogonal (to enforce self-consistency)
- Use an Inverted Softmax at inference time to avoid the hubness problem

Joulin et al (EMNLP 18):

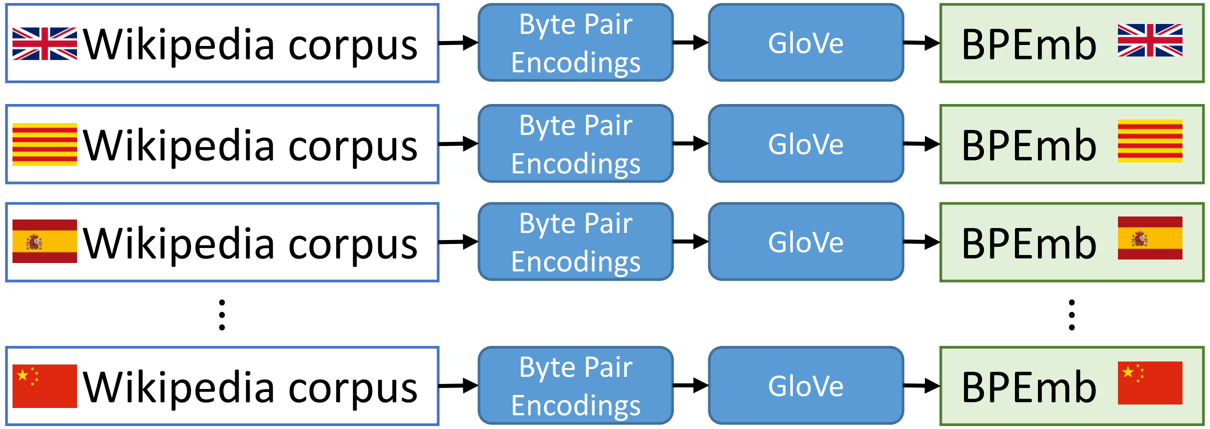
- Makes use of the CSLS criterion (cross-domain similarity local scaling)

Smith et al, “Offline bilingual word vectors, orthogonal transformations and the inverted softmax”, ICLR (2017)

Joulin et al, “Loss in translation: Learning bilingual word mapping with a retrieval criterion”, EMNLP (2018)

Multilingual Byte Pair Embeddings

275 × Monolingual Embeddings



Wikipedia corpus → Byte Pair Encodings → GloVe → BPEmb

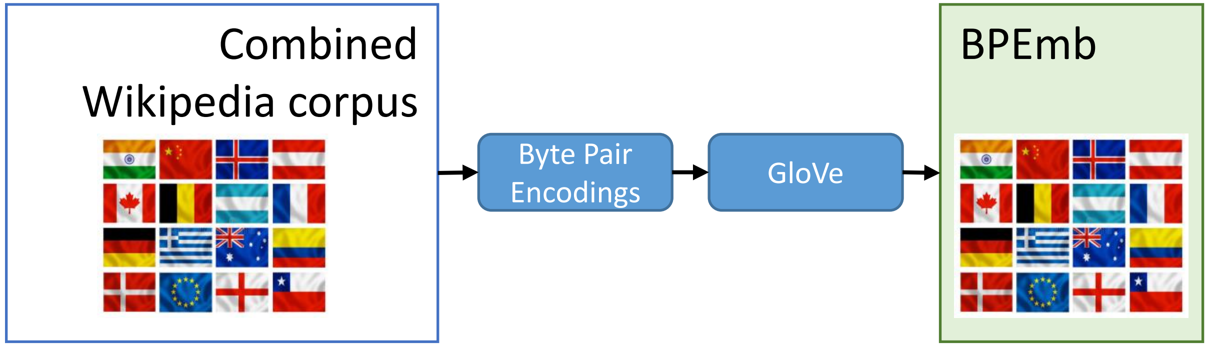
Wikipedia corpus → Byte Pair Encodings → GloVe → BPEmb

Wikipedia corpus → Byte Pair Encodings → GloVe → BPEmb

⋮

Wikipedia corpus → Byte Pair Encodings → GloVe → BPEmb

1 × Multilingual Embedding



Combined Wikipedia corpus → Byte Pair Encodings → GloVe → BPEmb

Heinzerling et al (ICLR 18):

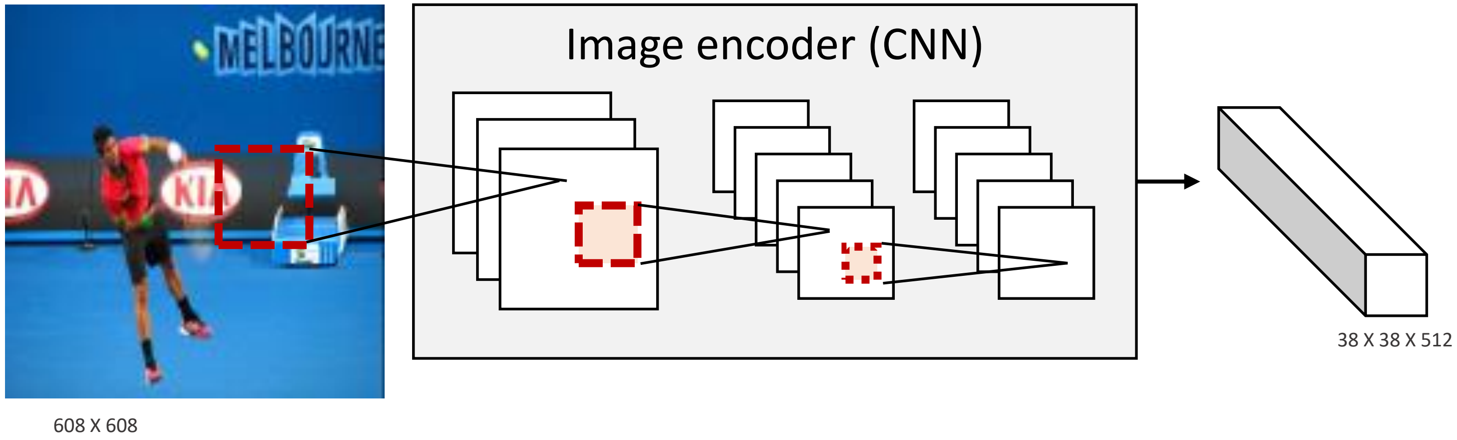
- Uses byte-pair encoding (BPE) – no need for tokenization
- Both Monolingual and Multilingual embeddings available

B. Heinzerling, M. Strube, "BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages", LREC (2018)

Centre de Visió per Computador

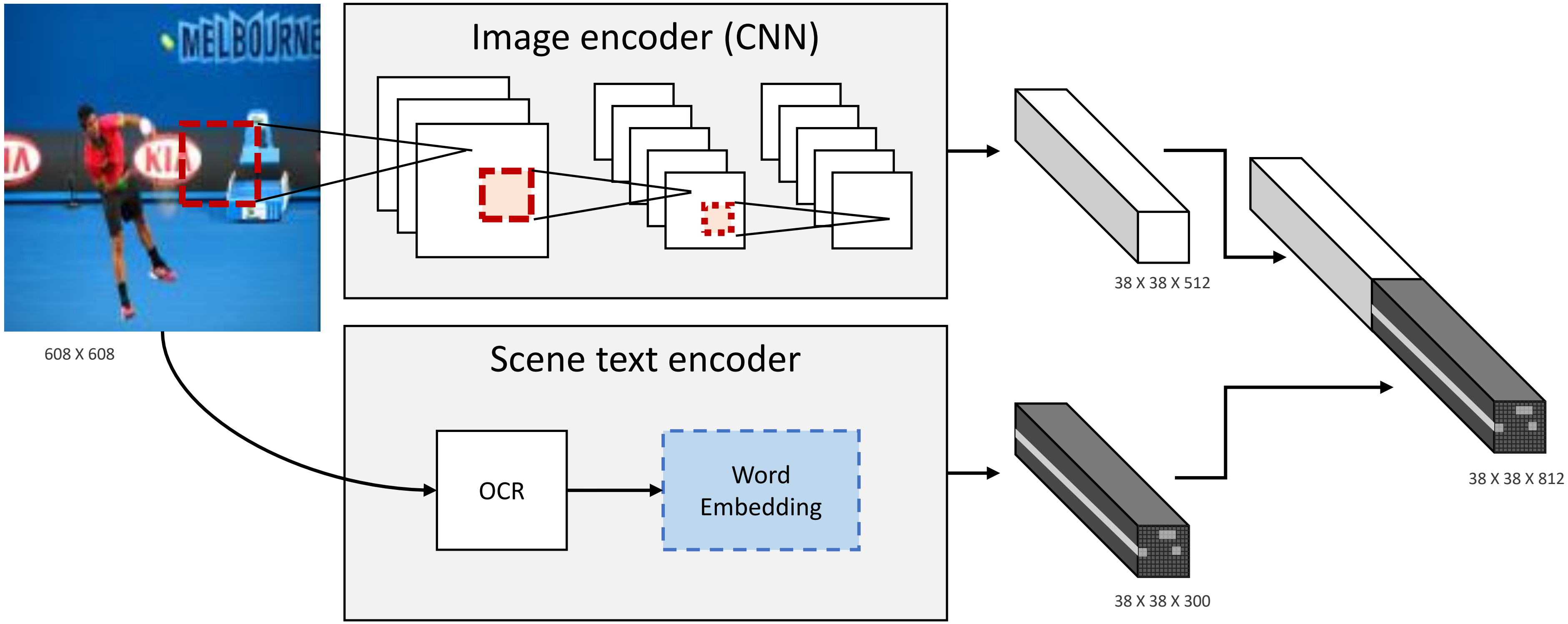
Scene Text VQA Model

Multimodal grid features for STVQA



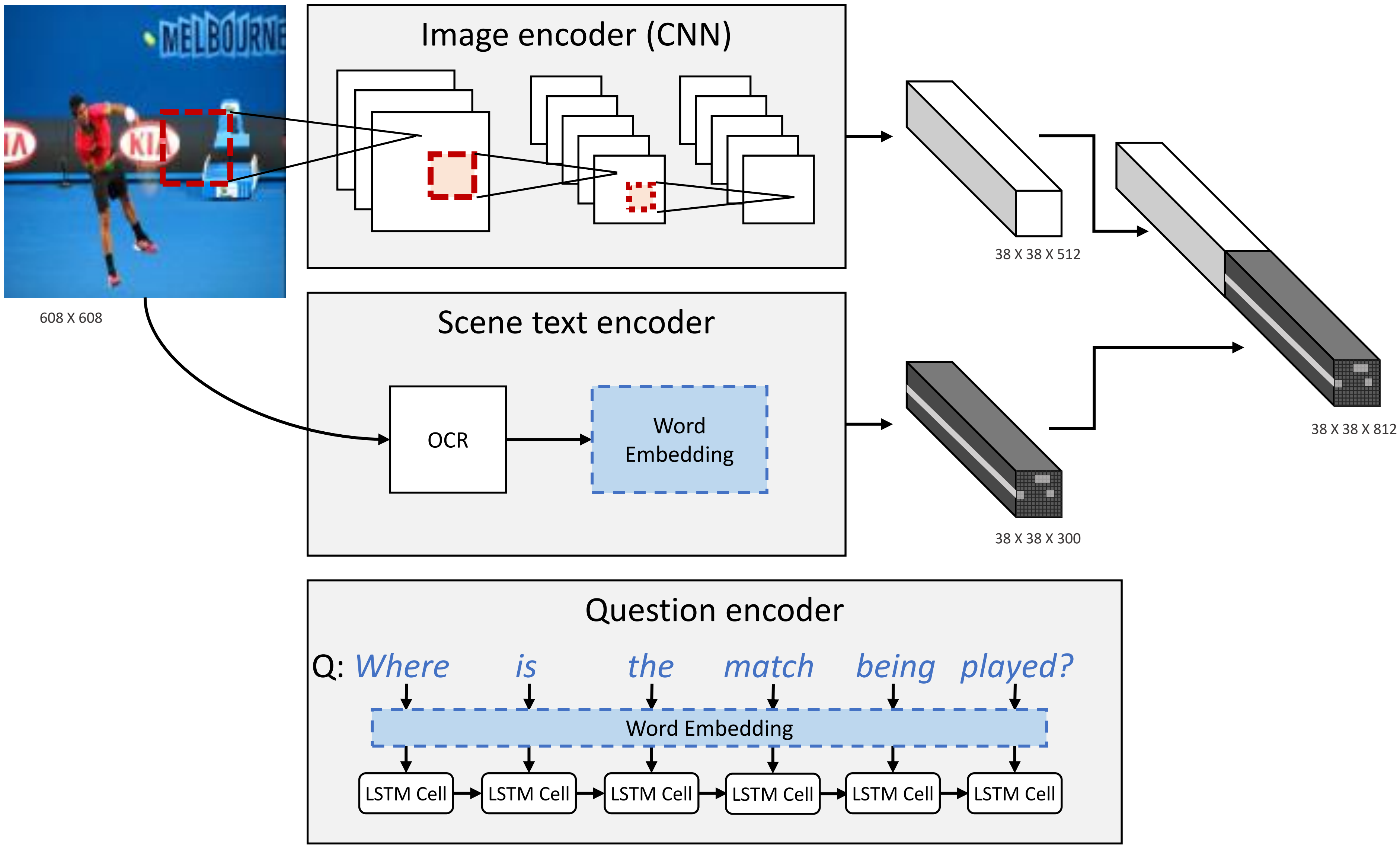
L. Gomez, A. Biten, R. Tito, A. Mafla, D. Karatzas, “Multimodal grid features and cell pointers for Scene Text Visual Question Answering”, PRL, 2021

Multimodal grid features for STVQA

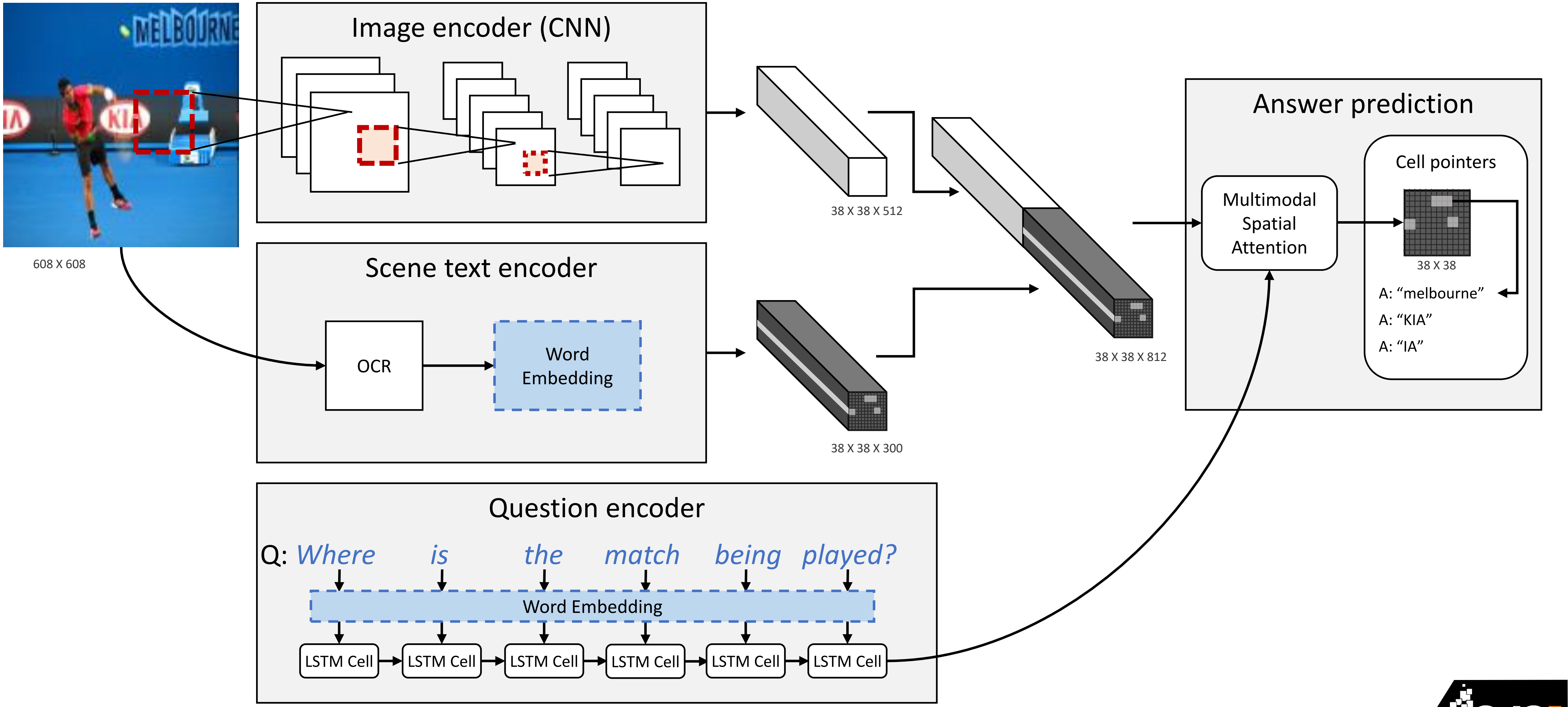


L. Gomez, A. Biten, R. Tito, A. Mafla, D. Karatzas, "Multimodal grid features and cell pointers for Scene Text Visual Question Answering", PRL, 2021

Multimodal grid features for STVQA



Multimodal grid features for STVQA



Experiments

Datasets and Metrics

	ST-VQA	EST-VQA	Custom Test Set
Languages	English	Chinese, English	Spanish, Catalan
Images	23,038	25,239	126
Questions	31,791	28,158	129
Answers	31,791	28,158	129

Metric used: ANSL (Average Normalised Levenstein Similarity)

$$s\left(\alpha_{ij}, o_{q_i}\right) = \begin{cases} \left(1 - NL(\alpha_{ij}, o_{q_i})\right) & if\ NL\left(\alpha_{ij}, o_{q_i}\right) < \tau \\ 0 & if\ NL\left(\alpha_{ij}, o_{q_i}\right) \geq \tau \end{cases}$$

Normalised Levenshtein Distance

$$ANLS = \frac{1}{N} \sum_{i=0}^{N} \left(\max_{j} s\left(\alpha_{ij}, o_{q_i}\right) \right)$$

N: total number of questions in the dataset

M: total number of GT answers per question i

α_{ij}: the ground truth answers

o_{q_i}: the model's answer for the i^{th} question

We trained 35 VQA models, each took 36 hours on a 12GB Titan X Pascal GPU

Baseline: Monolingual VQA

Separate monolingual models for English, Catalan, Spanish, and Chinese, using different pre-trained embeddings.

Rows indicate the test language.

	FastText embeddings								BPEmb ermbeddings			
	Embedding trained on CC				Embedding trained on Wiki				Embedding trained on Wiki			
	en	ca	es	zh	en	ca	es	zh	en	ca	es	zh
en	**0.34**	0.16	0.16	0.18	**0.34**	0.16	0.15	0.16	**0.33**	0.22	0.21	0.21
ca	0.21	**0.32**	0.19	0.16	0.17	**0.32**	0.18	0.16	0.19	**0.33**	0.19	0.19
es	0.16	0.16	**0.33**	0.17	0.18	0.18	**0.33**	0.17	0.20	0.22	**0.34**	0.19
zh	0.13	0.11	0.15	**0.28**	0.18	0.20	0.17	**0.33**	0.19	0.18	0.19	**0.31**

Catalan, Spanish and Chinese performance match English, when each model is tested on the corresponding language: automatic translation tools do not affect performance

Bilingual VQA

The original English model, and 3 x bilingual models using Catalan, Spanish and Chinese FastText embeddings aligned to English by using either Smith et al. or Joulin et al. methods

Test data are embedded using the corresponding bilingual aligned model

	Aligned with Smith et al.			
	en	ca	es	zh
en	**0.34**	**0.29**	**0.30**	0.22
ca	**0.26**	**0.33**	**0.29**	0.19
es	**0.28**	**0.28**	**0.34**	0.20
zh	0.16	0.16	0.17	**0.32**

	Aligned with Joulin et al.			
	en	ca	es	zh
en	**0.34**	**0.27**	**0.28**	**0.25**
ca	**0.27**	**0.32**	0.15	0.12
es	**0.27**	0.14	**0.33**	0.12
zh	**0.23**	0.14	0.15	**0.33**

Aligned models yield improved performance in English, without dropping performance in their original language

Chinese language poses challenges

Multilingual VQA

1 x multilingual model, using BPEmb embeddings on BPE encodings extracted jointly over 275 languages.

	Aligned with Heinzerling et al.			
	en	**ca**	**es**	**zh**
en	**0.35**	**0.30**	**0.28**	**0.24**
ca	**0.30**	**0.35**	**0.30**	0.15
es	**0.28**	**0.32**	**0.34**	0.14
zh	**0.24**	**0.23**	**0.22**	**0.32**

Improvement across all scenarios, including in the monolingual combinations, and across scripts (Chinese)

Training with 1+ languages

3 x bilingual models using Catalan, Spanish and Chinese FastText embeddings aligned to English by using either Smith et al.

Trained simultaneously on data from 2 languages

1 x multilingual model, using BPEmb embeddings on BPE encodings extracted jointly over 275 languages.

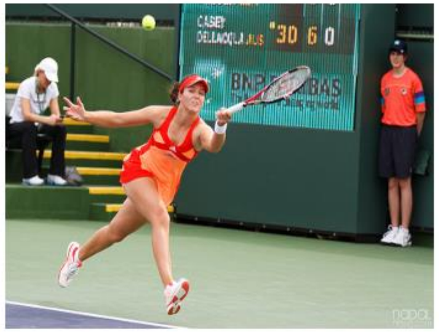
Trained simultaneously on data from different languages

	Aligned with Smith et al.		
	en + ca	**en + es**	**en + zh**
en	**0.34**	**0.34**	**0.34**
ca	**0.33**	0.29	0.26
es	0.32	**0.33**	0.29
zh	0.21	0.21	**0.32**

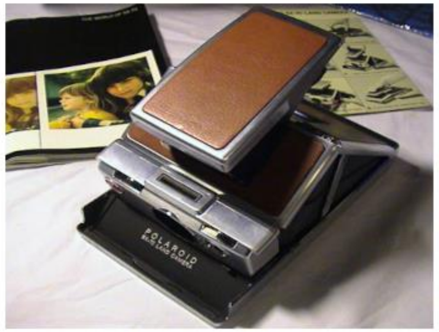
	Aligned with Heinzerling et al.			
	en + ca	**en + es**	**en + zh**	**All**
en	**0.34**	**0.34**	**0.34**	**0.34**
ca	**0.33**	0.31	0.27	**0.34**
es	0.30	**0.34**	0.28	**0.34**
zh	0.15	0.14	**0.31**	**0.31**

Training on multiple languages improves the performance on these languages

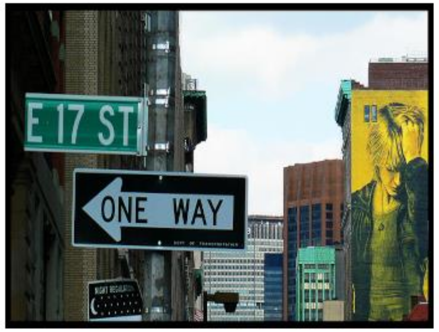
Qualitative Results ST-VQA



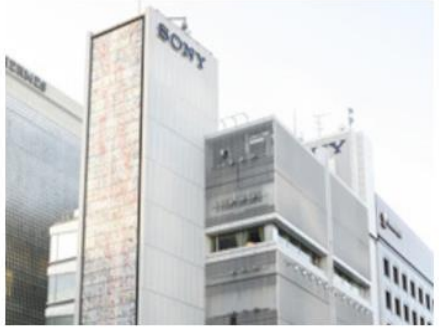
Q: What is the name of the tennis player?
A: Casey



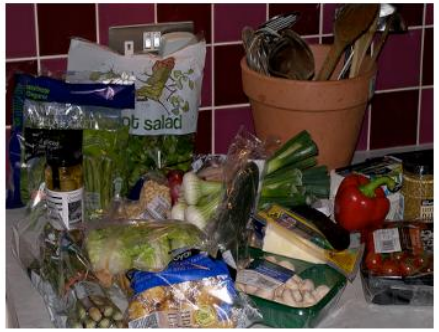
Q: Quina és la marca de la càmera?
A: Polaroid



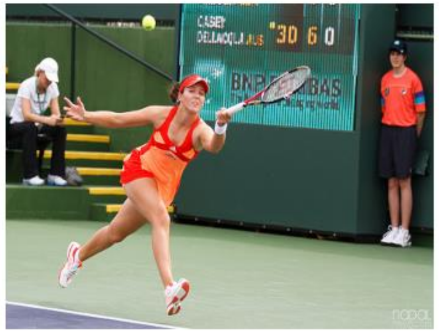
Q: Es esta una calle de un solo sentido o de dos?
A: Un sentido



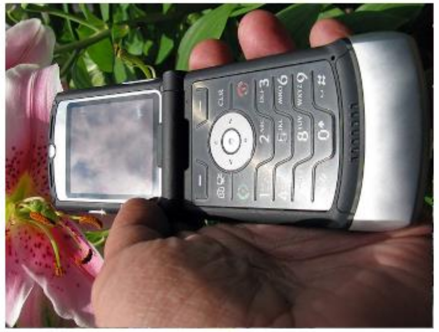
Q: What company name is written on the tallest building (*)
A: SONY



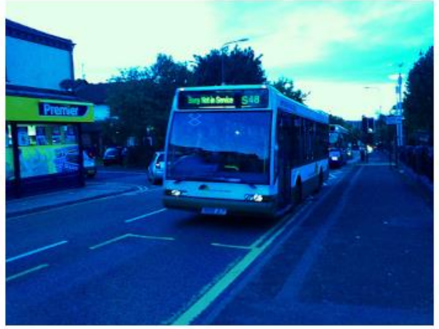
Q: What is the first word in black on the jar?
A: Salad



Q: Com es diu el tennista?
A: Casey



Q: Qué botón se selecciona para eliminar contenido?
A: 8tuv

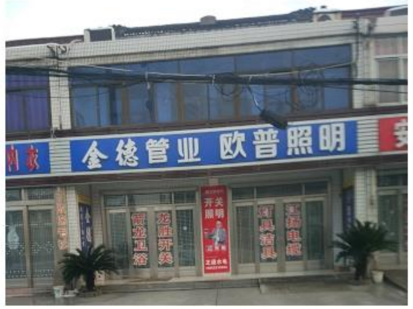


Q: What does the screen in the bus say about service (*)
A: s48

(*) The chinese questions are shown here in English for better readability

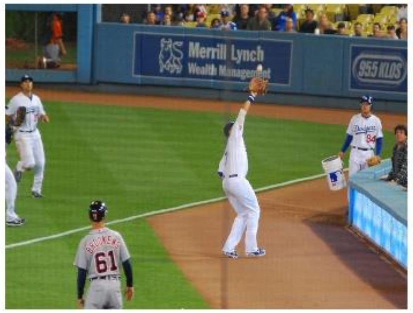
Qualitative results on natively multilingual datasets

EST-VQA (English and Chinese)



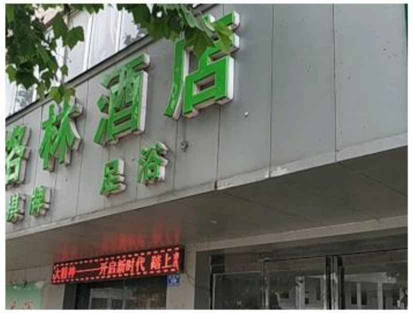
Q: What is the name of the shop in the right?

A: 欧普照明



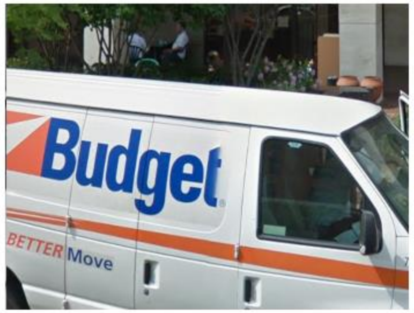
Q: What's the number of player at the bottom left corner of the image?

A: 61



Q: What is the name of this shop?

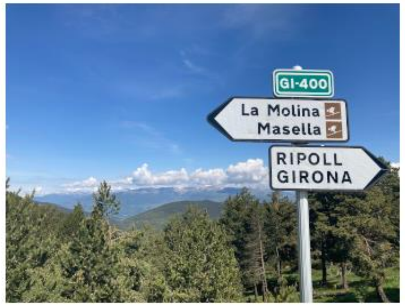
A: 踏上



Q: Which company is this car from?

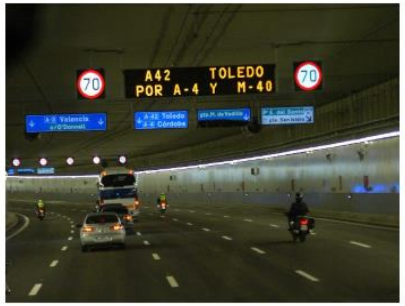
A: Budget budget better

Custom dataset (Spanish and Catalan)



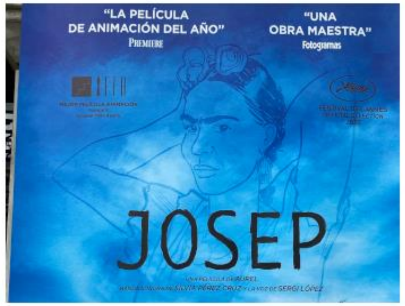
Q: Quina carretera s'indica al cartell verd?

A: GI-400



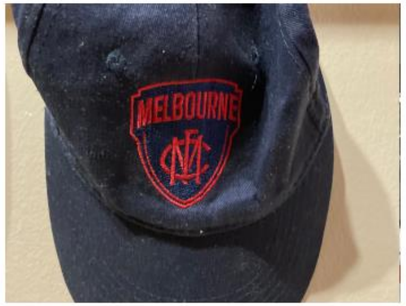
Q: ¿A qué velocidad se puede circular?

A: 70



Q: Quina pel·lícula s'anuncia al cartell?

A: 2020



Q: A qué equipo pertenece la gorra

A: Melbourne

Conclusions

- Explored ways to extend an **existing Scene Text VQA model to a multilingual scenario**, without the need for collecting new data, exploiting multilingual embeddings
- Automatic translation does not affect the monolingual performance of the languages
- Using embeddings aligned to more languages increases performance in said languages
- Training with data in more languages has a positive effect, as expected, even if such data are automatically obtained (helps the model to learn language structure)
- Experiments on native multilingual datasets confirm our results

CVC
Centre de Visió per Computador
Centre de Visió per Computador
Edifici O

Thank you!

Research supported by:

aws *Machine Learning Research Awards*

Google Faculty Research Awards

Fundación BBVA

European Commission

MINISTERIO DE CIENCIA E INNOVACIÓN

Financiado por la Unión Europea NextGenerationEU

Plan de Recuperación, Transformación y Resiliencia

AGENCIA ESTATAL DE INVESTIGACIÓN

DOCTORATS INDUSTRIALS *Aliança de talent!*

Generalitat de Catalunya
Departament d'Economia i Coneixement
Secretaria d'Universitats i Recerca