Cross-Modal Deep Networks for Document Image Classification

Souhail Bakkali[1] Zuheng Ming[1] Mickaël Coustaty[1] Marçal Rusiñol[2]
[1]L3i, University of La Rochelle, France
[2]CVC, Universitat Autònoma de Barcelona, Spain

ICIP 2020

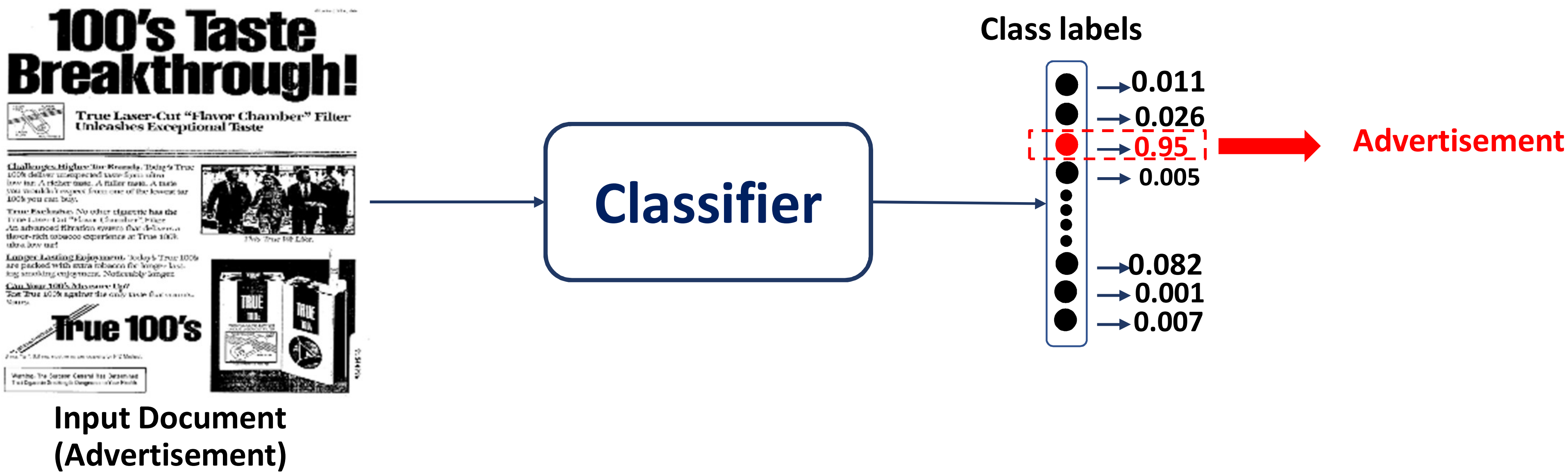
UNION EUROPEENNE
Fonds Européen de développement Régional

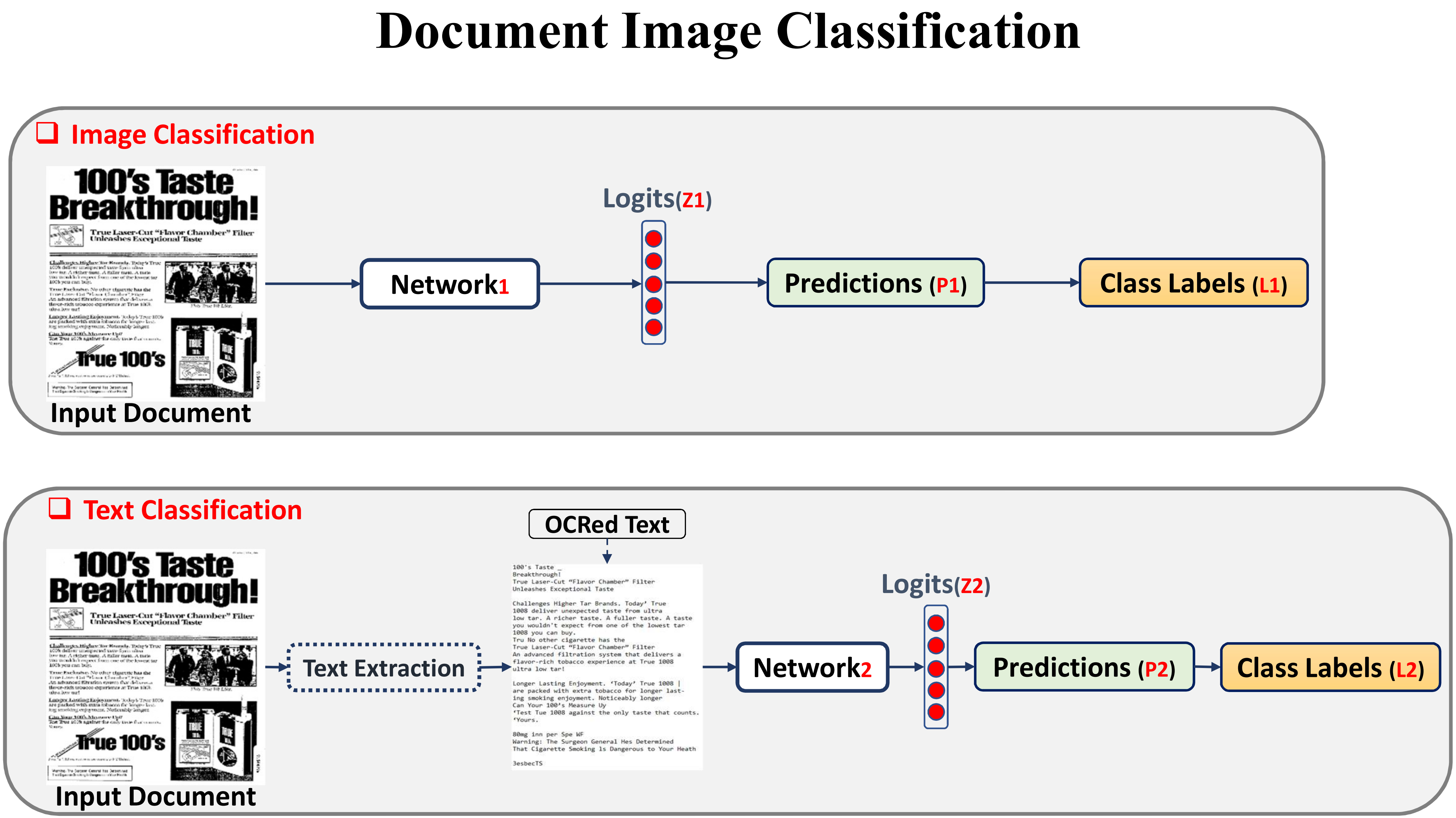
Informatique Image Interaction
Univ. La Rochelle

Computer Vision Center

RÉGION Nouvelle-Aquitaine

Document Image Classification

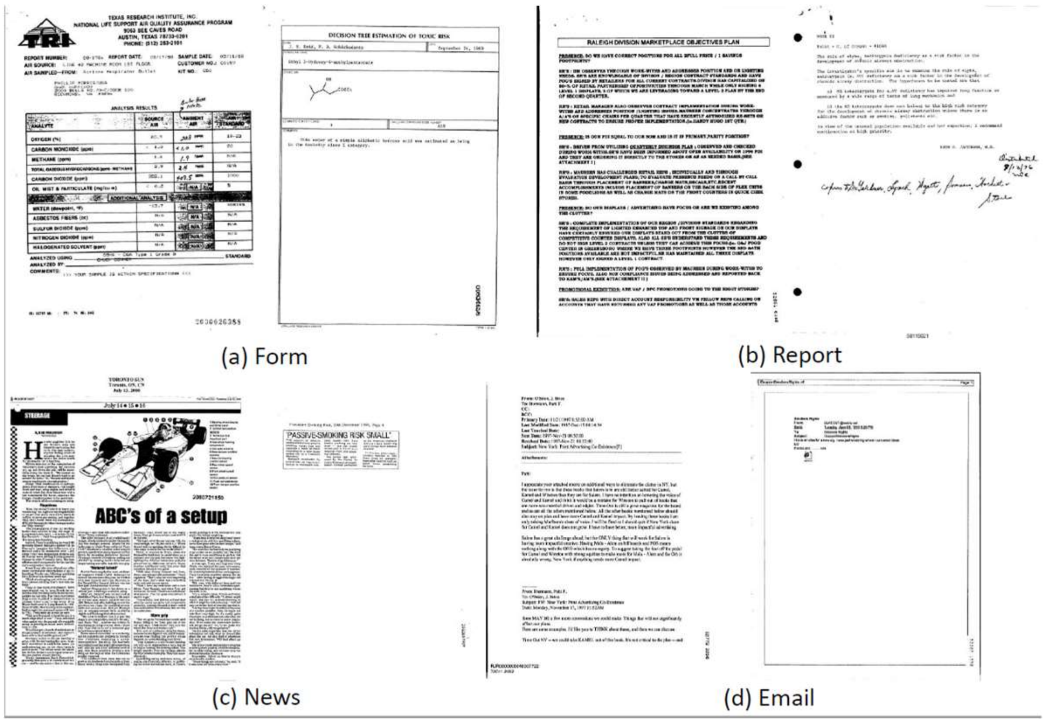



Document Image Classification
Image Classification
100's Taste Breakthrough!
True 100's
Input Document
Network1
Logits(Z1)
Predictions (P1)
Class Labels (L1)
Text Classification
100's Taste Breakthrough!
True 100's
Input Document
Text Extraction
OCRed Text
Network2
Logits(Z2)
Predictions (P2)
Class Labels (L2)

Motivation: **The Need for a Multimodal Analysis**

❑ Challenges:

- Some easy (stable) classes with similar visual content.
- Some hard classes with high variability.
- **10** classes with different complexity levels of the **Tobacco-3482** dataset.



(a) Form (b) Report (c) News (d) Email

❑ Solution:

- Need for a Multimodal Analysis

Architecture Network

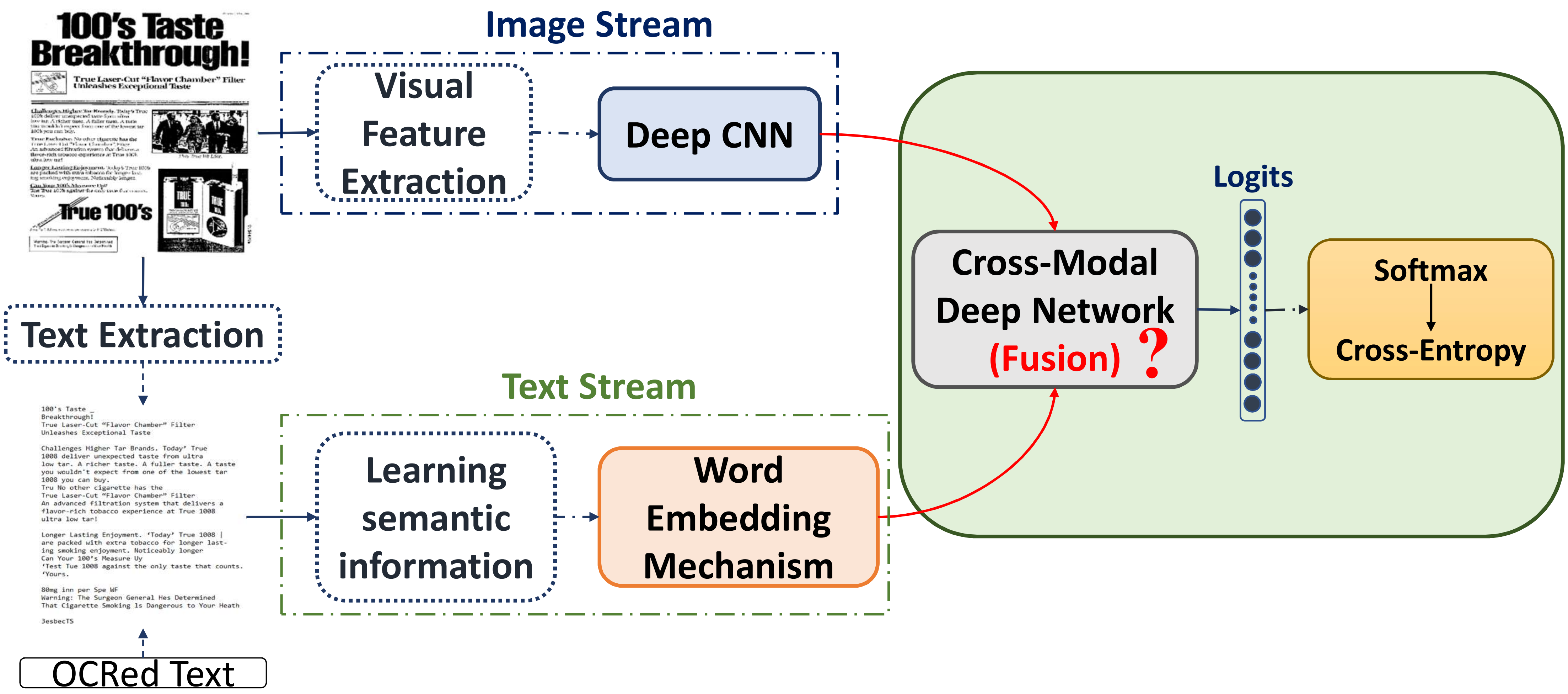
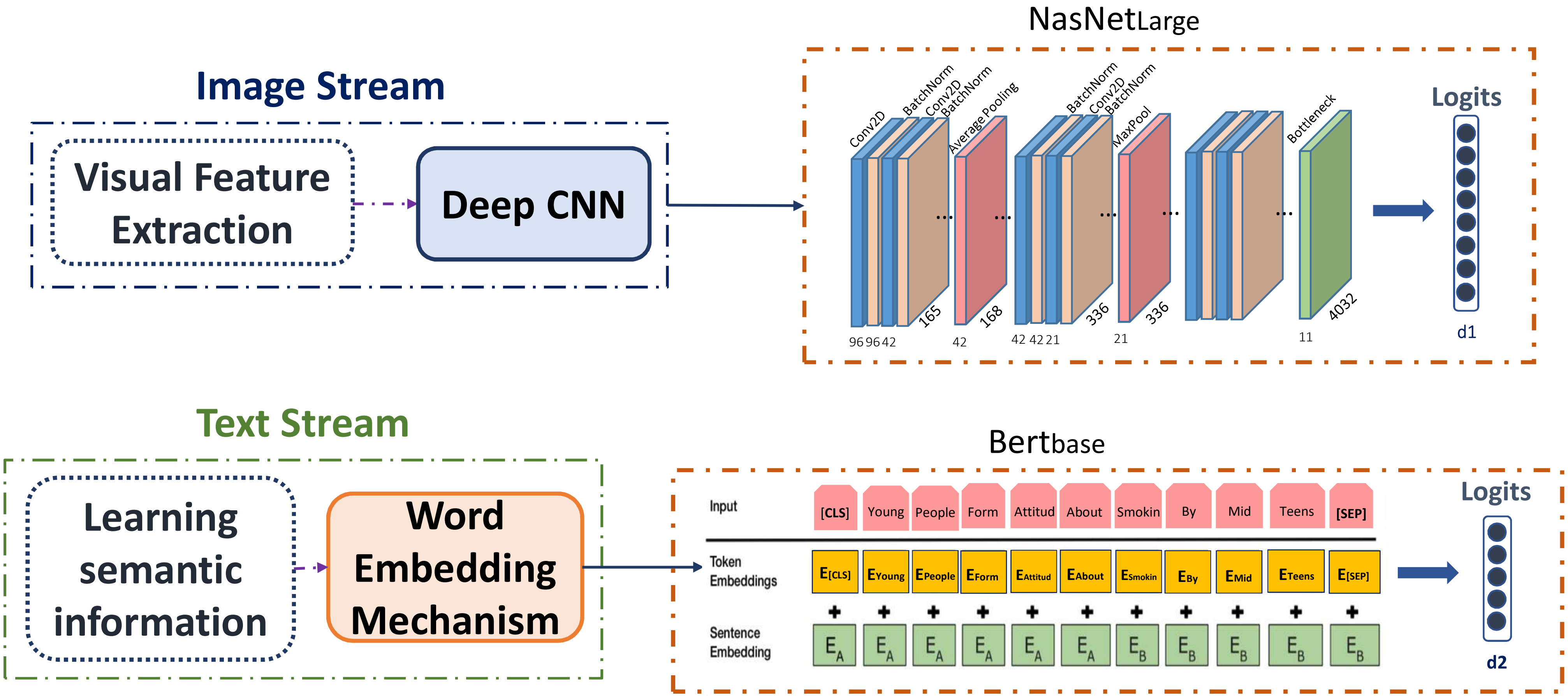
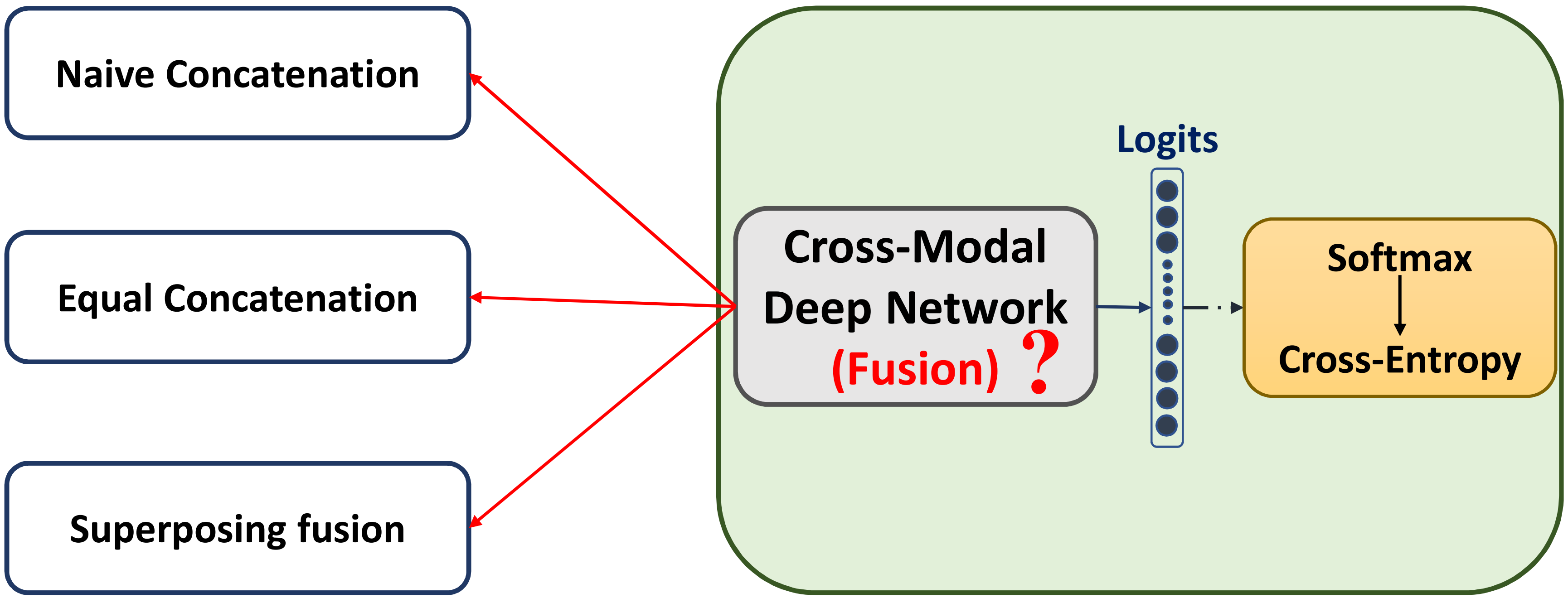


Image and Text Feature Learning

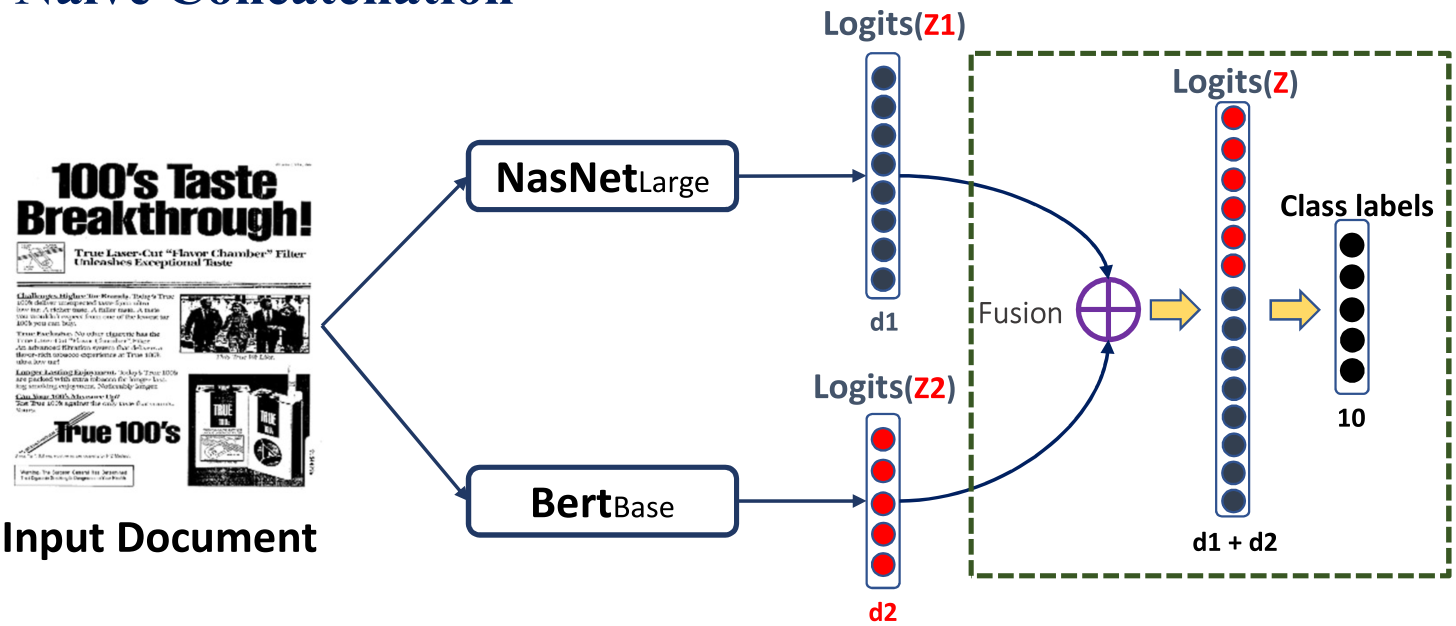


How To Bridge These Two Different Modalities ?



Cross-Modal Feature Learning

❑ Naive Concatenation

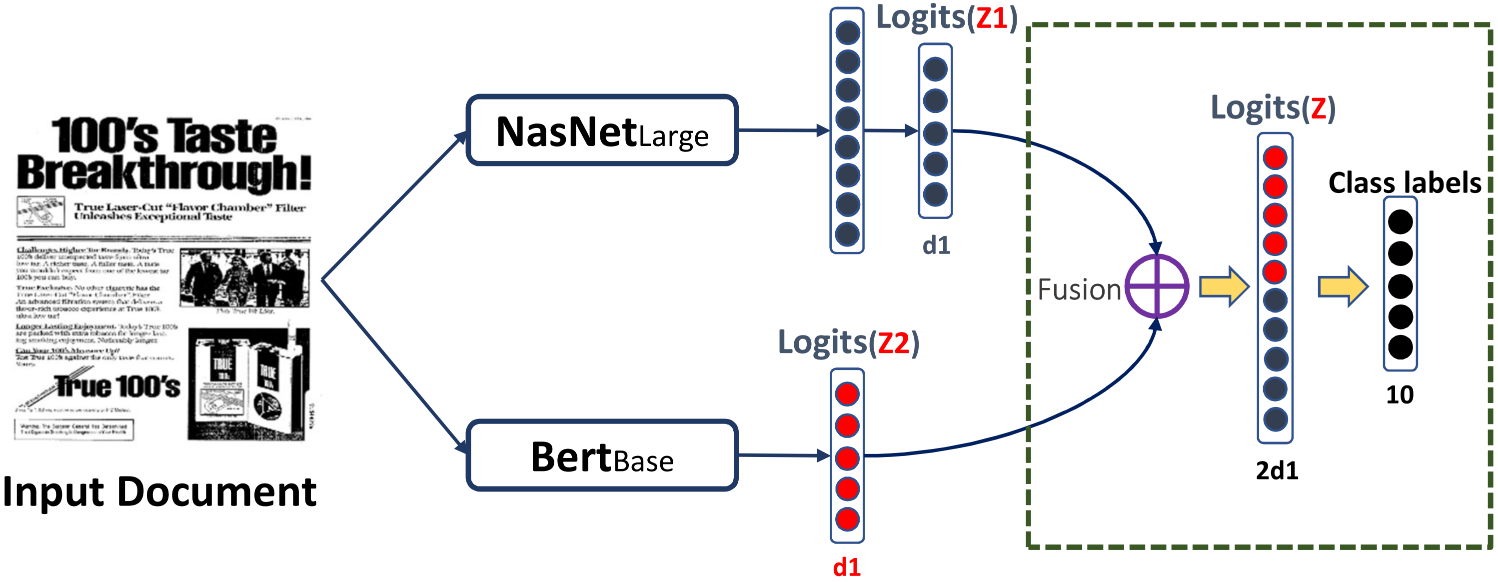


The generated cross-modal features are given by:

$$X_a = [X_1, X_2], \quad X_a \in \mathbb{R}^{d_1+d_2}$$

Cross-Modal Feature Learning

❑ Equal Concatenation

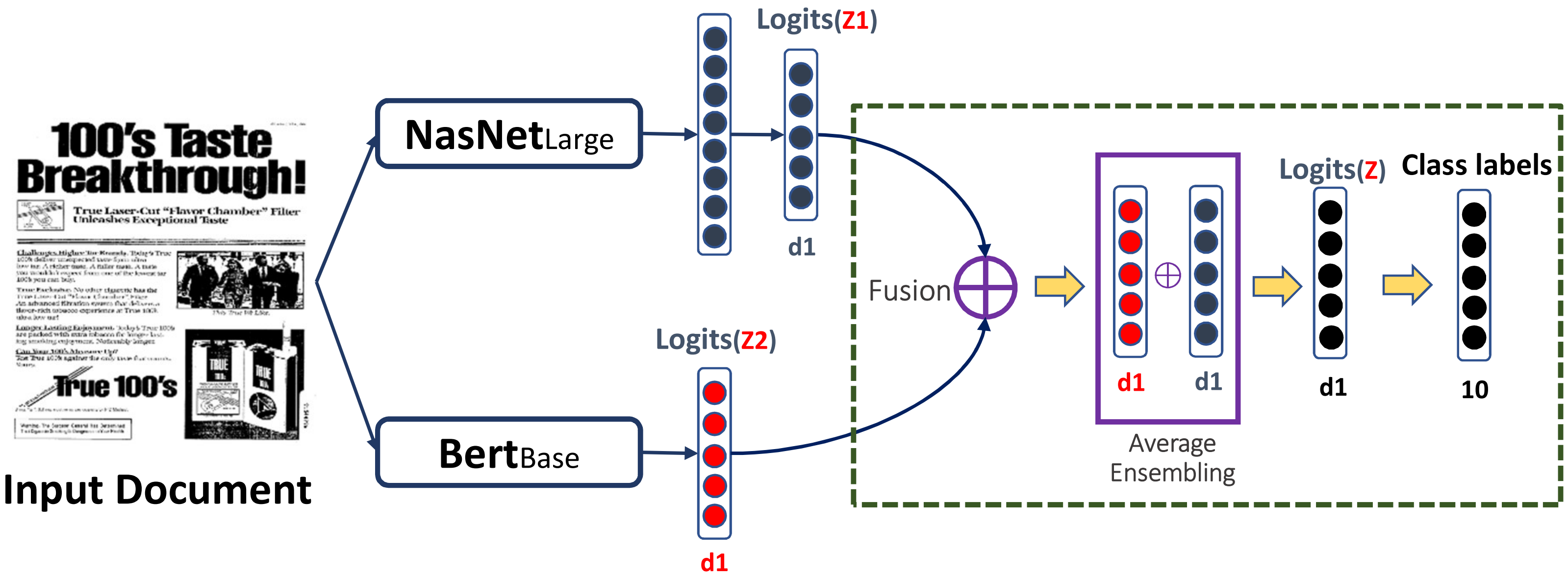


The generated cross-modal features are given by:

$$X_a = [X_1, X_2], \quad X_a \in \mathbb{R}^{2d_1}$$

Cross-Modal Feature Learning

❑ Superposing Concatenation



The generated cross-modal features are given by:

$$X_a = [X_1 + X_2], \quad X_a \in \mathbb{R}^{d_1}$$

Evaluation

Table 1. The Image-stream evaluation of best models from different methods on Tobacco-3482 dataset

Method	Accuracy(%)
AlexNet [12]	90.04
GooGleNet [12]	88.4
VGG-16 [12]	91.01
ResNet-50 [12]	91.13
MobileNetV2 [9]	84.50
InceptionV3 [8]	93.2
NASNet-Large	96.25

Table 2. Accuracy comparison of Text-stream state-of-the-art models on Tobacco-3482 dataset

Method	Accuracy(%)
FastText-CNN [9]	73.8
Feature Ranking (ACC2) [8]	87.1
Glove-CNN1D-LSTM	51
Glove-GRU	61
Bert	97.18

Table 3. Overall accuracy on the Tobacco-3482 dataset

Model	Accuracy(%)	ADVE	Email	Form	Letter	Memo	News	Notes	Report	Resume	Scientific
Single-Modal (Image-NASNet)	96.25	1	1	0.96	0.94	0.98	1	0.90	1	0.78	0.90
Single-Modal (Text-Bert)	97.18	0.97	0.99	0.98	0.93	0.97	0.98	0.89	1	0.96	0.95
Multimodal Model [9]	87.8	0.93	0.98	0.88	0.86	0.90	0.90	0.85	0.71	0.96	0.68
Two Stream Model [8]	95.8	0.94	0.98	0.95	0.98	0.97	0.97	0.88	0.92	1	0.93
Cross-Modal (Naive Concat.)	99.14	1	0.99	0.96	1	1	1	1	0.98	1	0.98
Cross-Modal (Equal Concat.)	98.42	0.98	0.99	0.95	1	0.98	0.97	1	1	0.96	0.98
Cross-Modal (Superposing fusion)	99.71	1	1	0.97	1	1	1	1	1	1	1

Summary

- **Cross-Modal Network:** Learns simultaneously image and text features extracted from document images.
- Three **Feature Fusion** methods to perform Cross-Modal document image classification.
- **State-of-the-art** results compared to single-modal and multi-modal networks.

This work has been co-funded by the Region Nouvelle-Aquitaine under the framework of the CPER-FEDER program