

Cross-Modal Deep Networks for Document Image Classification

Souhail Bakkali¹ Zuheng Ming¹ Mickaël Coustaty¹ Marçal Rusiñol²

¹L3i, University of La Rochelle, France

²CVC, Universitat Autònoma de Barcelona, Spain

ICIP 2020



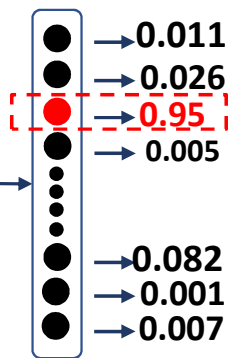
Document Image Classification



Input Document
(Advertisement)



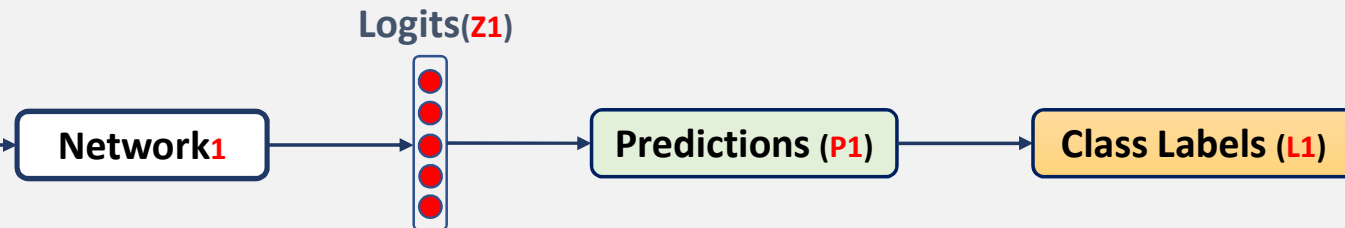
Class labels



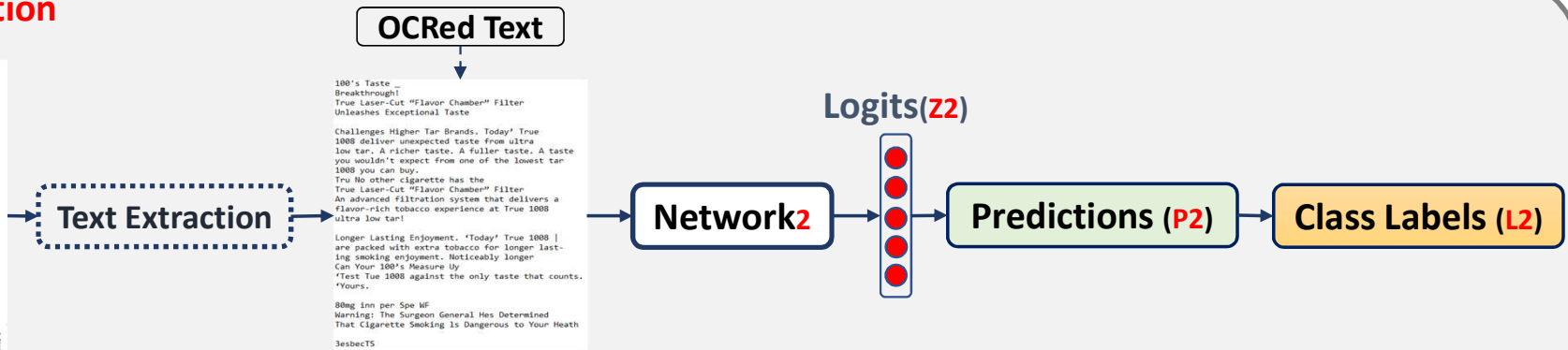
Advertisement

Document Image Classification

Image Classification



Text Classification



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.

Solution:

- Need for a Multimodal Analysis

(a) Form: A document titled 'DESIGN FREE ESTIMATION OF SOILS WITH' containing a table of analysis results. The table has columns for 'ANALYSIS RESULTS', 'SAMPLE', and 'UNIT'. It lists various soil parameters like 'LIQUID LIMIT (%)', 'PLASTICITY INDEX (%)', and 'SHRINKAGE (%)' with corresponding values and units.

(b) Report: A document titled 'BALFOUR DIVISION MARKETPLACE OBJECTIVES PLAN'. It contains several sections of text, including 'PURPOSE', 'SCOPE', and 'OBJECTIVES'. The text is dense and appears to be a formal business or organizational plan.

(c) News: A newspaper clipping titled 'ABC's of a setup' with a sub-headline 'PASSIVE-SMOKING RISK SMALL'. The article includes a photograph of a person working on a vehicle and discusses the risks of passive smoking.

(d) Email: An email interface showing a list of messages on the left and the content of a selected message on the right. The message content includes a header with 'From: [Name]', 'To: [Name]', and 'Subject: [Subject]', followed by the main body of the email text.

Architecture Network

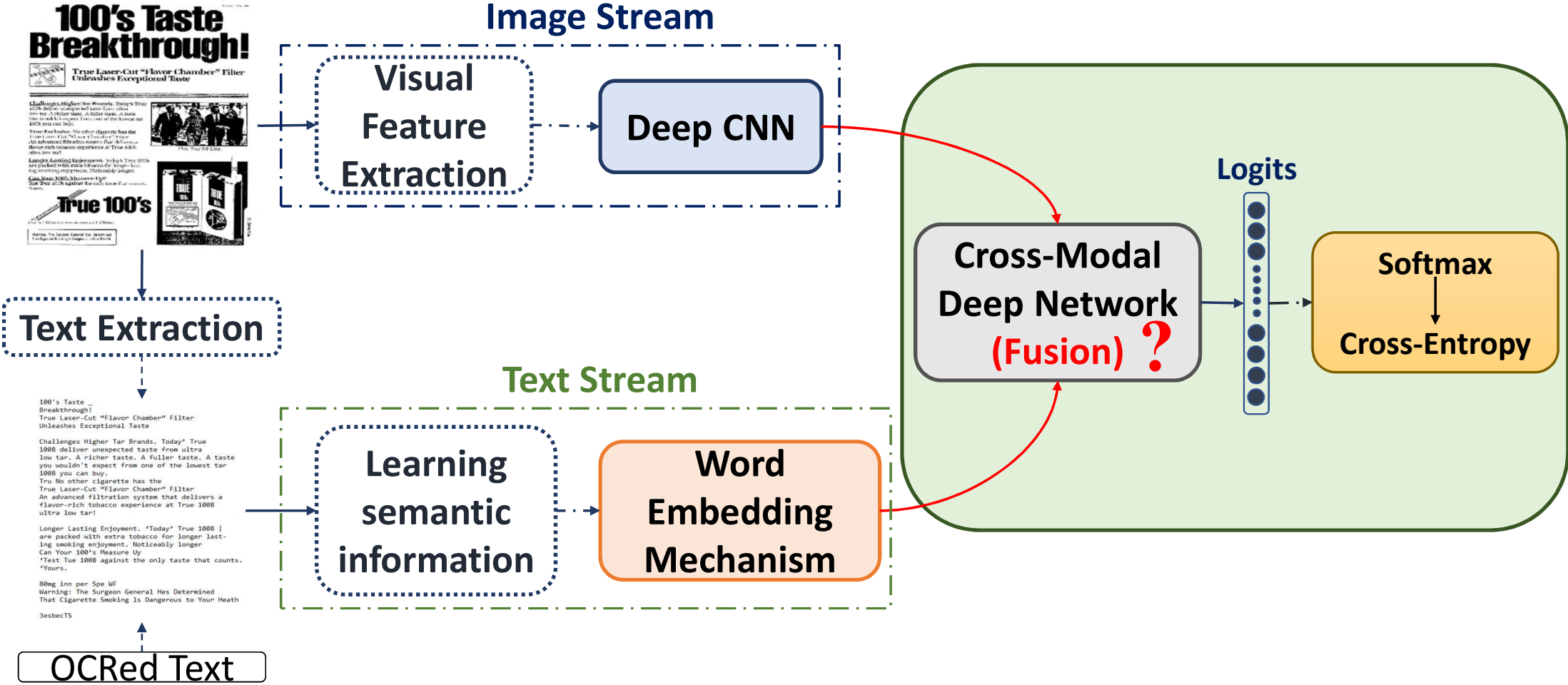
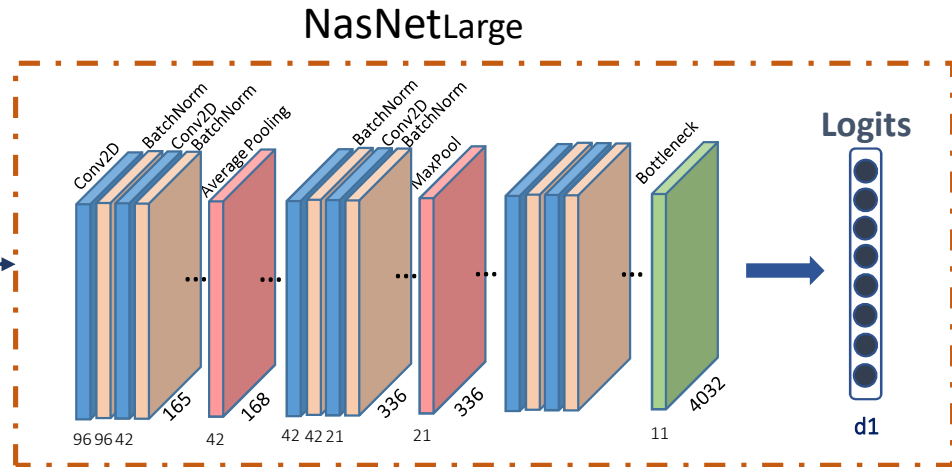
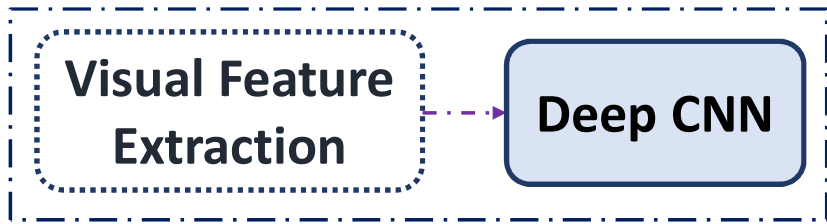
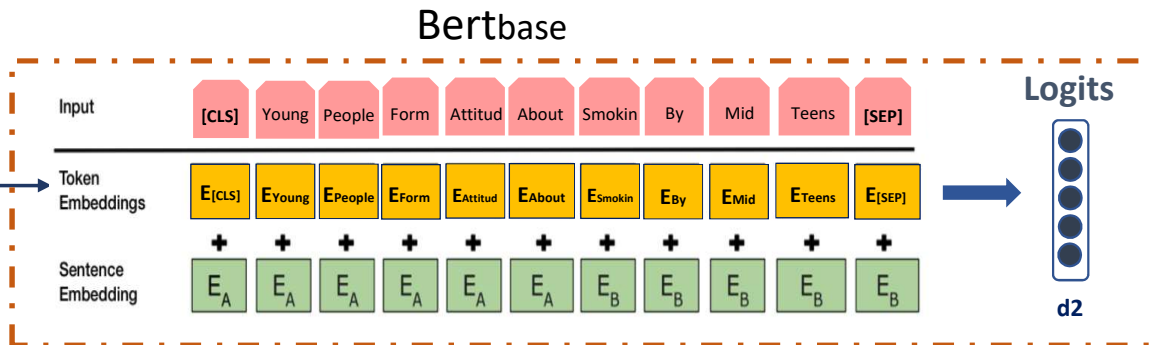
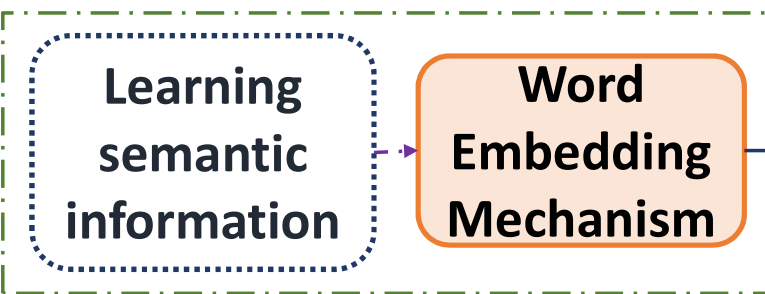


Image and Text Feature Learning

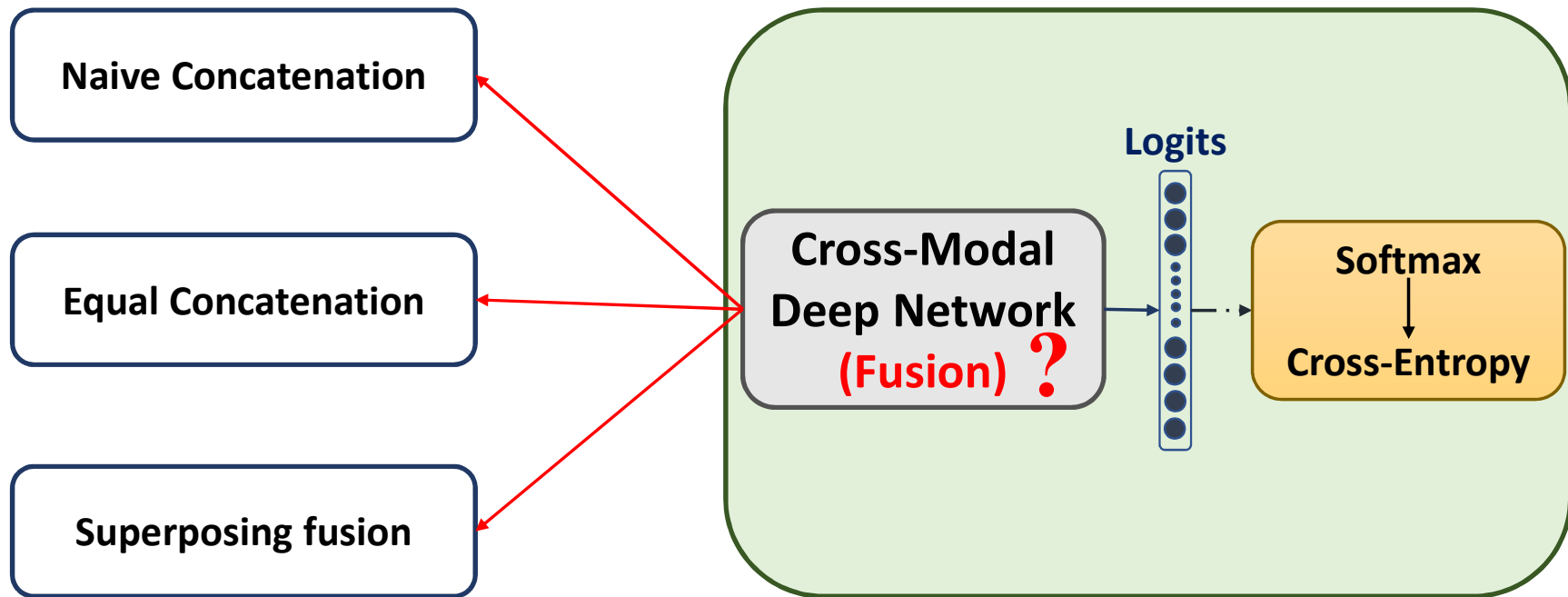
Image Stream



Text Stream

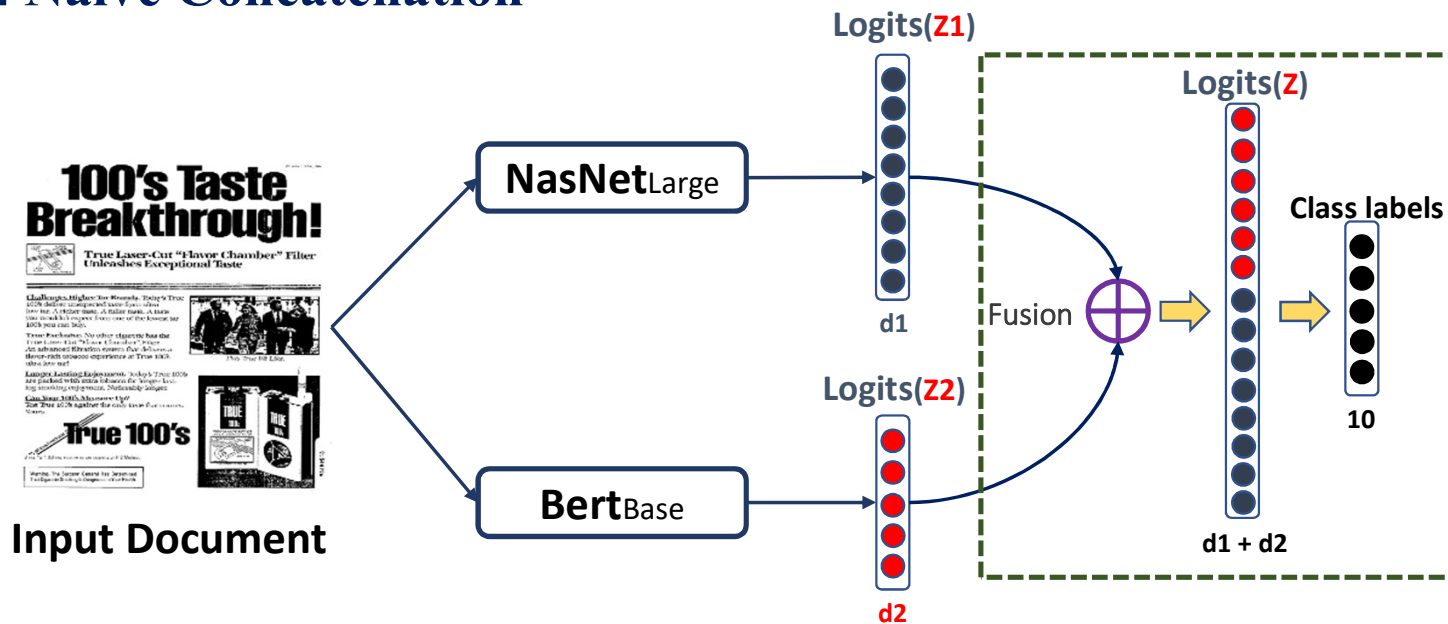


How To Bridge These Two Different Modalities ?



Cross-Modal Feature Learning

□ Naive Concatenation

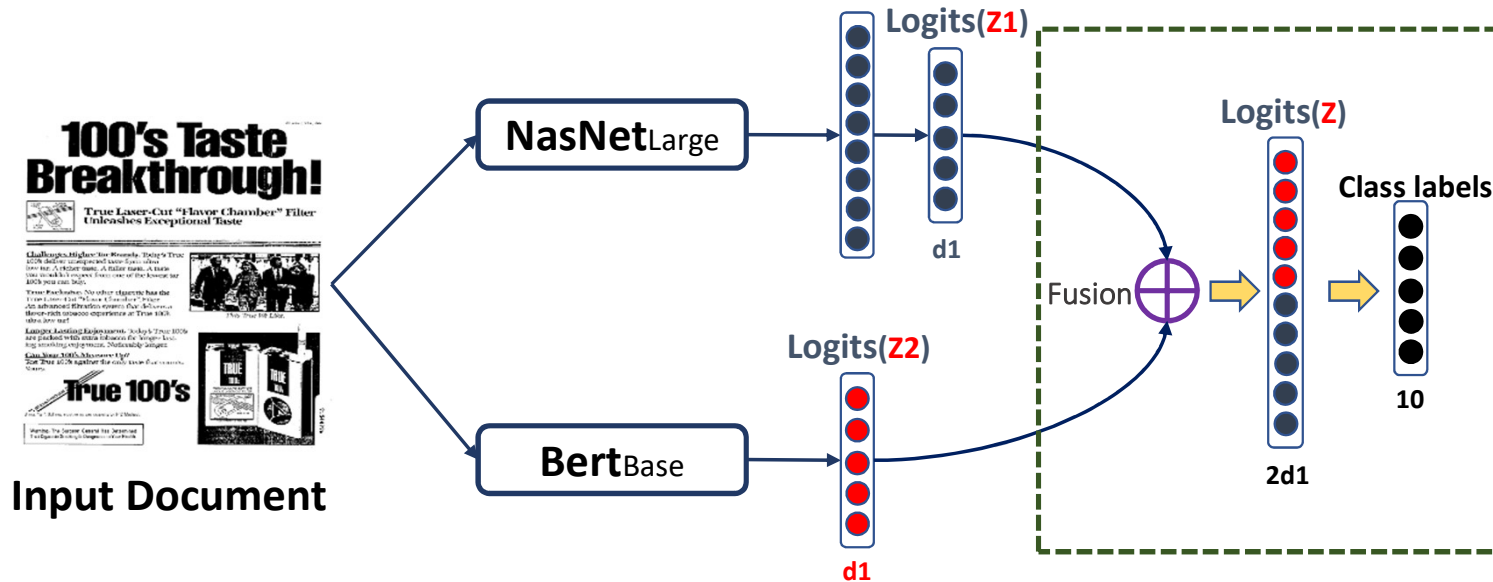


The generated cross-modal features are given by:

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

Cross-Modal Feature Learning

□ Equal Concatenation

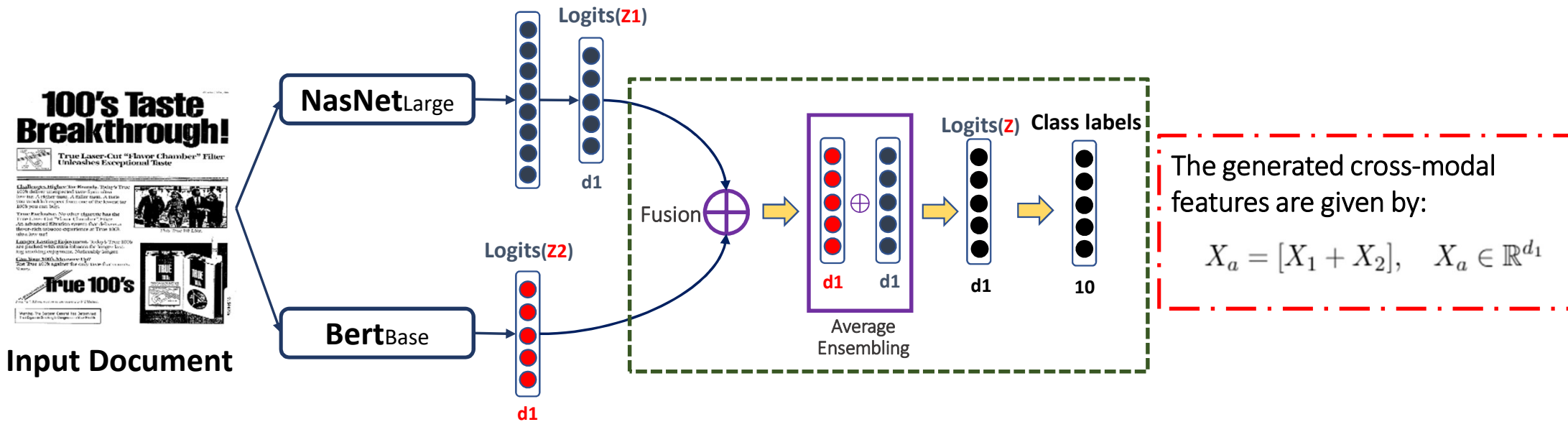


The generated cross-modal features are given by:

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

Cross-Modal Feature Learning

□ Superposing Concatenation



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