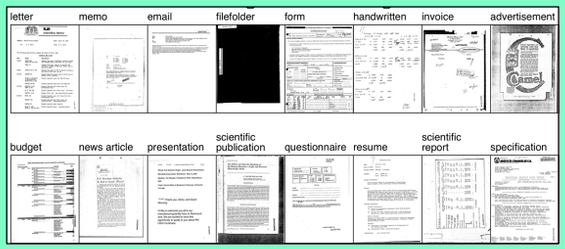
Stefano P. Zingaro | Giuseppe Lisanti | Maurizio Gabbrielli

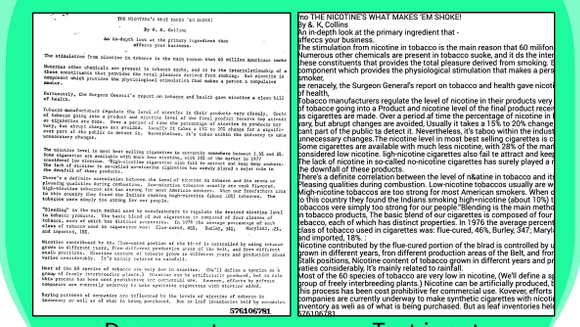
25th International Conference on Pattern Recognition (ICPR 2020)

Department of Computer Science and Engineering - University of Bologna

RVL-CDIP subset of *Tobacco Legacy Document Library*



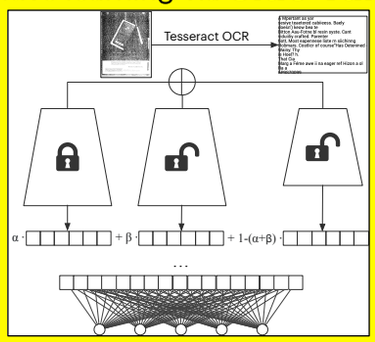
Tobacco3482 subset of RVL-CDIP



Document Image

Text input source

Additive learning via **Side-Tuning**



We select one off-the-shelf architecture pre-trained on ImageNet for image classification, and one CNN for text classification. We choose the side network (the one with locked weights) to be the same as the one for the image classification task to favor visual features over textual. We then merge the outputs by linearly combining the features resulting from each network and finally, we add a fully connected layer before the actual classifier.

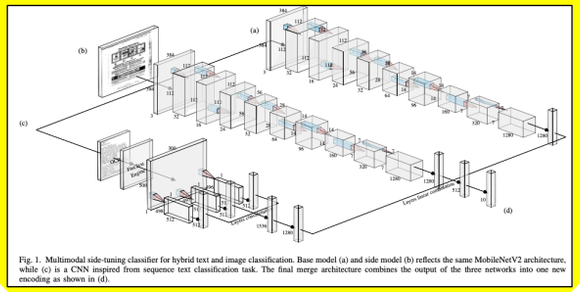


Fig. 1. Multimodal side-tuning classifier for hybrid text and image classification. Base model (a) and side model (b) reflects the same MobileNetV2 architecture, while (c) is a CNN inspired from sequence text classification task. The final merge architecture combines the output of the three networks into one new encoding as shown in (d).

The proposed network architecture

Multimodal Side-Tuning for Document Classification

TABLE IV
OVERALL AND PER-CLASS ACCURACY ON THE TOBACCO3482 DATASET COMPARED WITH THE RESULTS FROM [8]. THE SELECTED ALPHA CONFIGURATION FOR THE MULTIMODAL SIDE-TUNING IS $a_0 = 0.3$, $a_1 = 0.2$, AND $a_2 = 0.5$ FOR MOBILENETV2 AND $a_0 = 0.3$, $a_1 = 0.3$, AND $a_2 = 0.4$ FOR RESNET50. BEST RESULTS IN BOLD.

Model	OA	Adve	Email	Form	Letter	Memo	News	Note	Report	Resume	Scientific
Audebert	87.8%	93.0%	98.0%	88.0%	86.0%	90.0%	90.0%	85.0%	71.0%	86.0%	68.0%
Text	67.8%	93.3%	29.5%	77.0%	58.8%	49.7%	63.6%	68.7%	52.0%	60.7%	**79.9%**
Multimodal (ResNet50)	90.3%	**96.1%**	98.3%	**90.8%**	91.7%	**93.5%**	**95.5%**	87.6%	**76.7%**	89.4%	68.0%
Multimodal (MobileNetV2)	**90.5%**	94.8%	**99.1%**	88.7%	**93.2%**	93.0%	**95.5%**	**89.7%**	76.2%	**95.3%**	67.4%

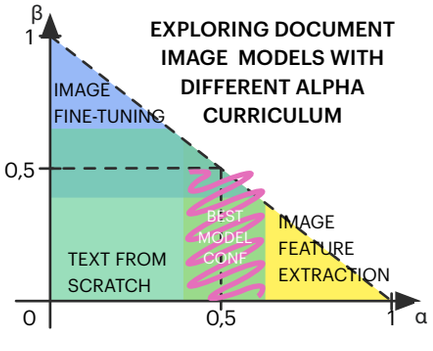
TABLE II
OVERALL ACCURACY ON THE TOBACCO3482 IMAGE DATASET USING TWO DIFFERENT OFF-THE-SHELF ARCHITECTURES FOR BOTH BASE AND SIDE MODEL IN THE SIDE TUNING FRAMEWORK. BEST RESULT IN BOLD.

Model (base architecture)	#Params	OA
Image (ResNet50)	≈ 51M	87.2%
Image (MobileNetV2)	≈ 7M	88.0%
Multimodal (ResNet50)	≈ 57M	90.3%
Multimodal (MobileNetV2)	≈ 12M	**90.5%**

EXPERIMENTAL RESULTS

TABLE I
BASELINE MODELS AND MULTIMODAL OVERALL ACCURACY FOR TOBACCO3482 USING MOBILENETV2 (VISUAL FEATURES) AND 1D CNN (TEXTUAL FEATURES) ARCHITECTURES. BEST RESULT IN BOLD.

Model	#Params	OA
Text	≈ 1.8M	67.8%
Image (fine-tuning)	≈ 3.5M	84.0%
Image (side-tuning)	≈ 7M	88.0%
Multimodal (side-tuning)	≈ 12M	**90.5%**

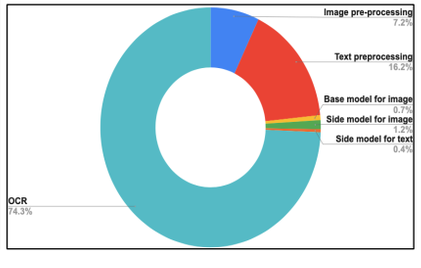


EXPLORING DOCUMENT IMAGE MODELS WITH DIFFERENT ALPHA CURRICULUM

TABLE III
OVERALL ACCURACY ON THE RVL-CDIP DATASET COMPARED WITH THE RESULTS FROM PREVIOUS WORKS. MODALITIES OF THE DATA SOURCE ARE IMAGE (I), TEXT (T), OR BOTH (I+T). THE SELECTED ALPHA CONFIGURATION FOR THE MULTIMODAL SIDE-TUNING IS $a_0 = 0.3$, $a_1 = 0.2$, AND $a_2 = 0.5$ FOR MOBILENETV2 AND $a_0 = 0.3$, $a_1 = 0.3$, AND $a_2 = 0.4$ FOR RESNET50. BEST RESULT IN BOLD.

Model	#Params	Modality	OA
CNNs [19]	≈ 62M	I	89.8%
Audebert [8]	≈ 8M	I+T	90.6%
AlexNet + SPP [22]	≈ 62M	I	90.94%
VGG16 [23]	≈ 138M	I	90.97%
VGG16 + ULMFit [6]	≈ 162M	I+T	**93.6%**
Text	≈ 1.8M	T	80.5%
Multimodal (MobileNetV2)	≈ 12M	I+T	92.2%
Multimodal (ResNet50)	≈ 57M	I+T	92.7%

TIMING OF DOCUMENT INFERENCE
(average of 1124ms on GPU)



szingaro.github.io/multimodal-side-tuning