

# Multimodal Document Image Classification

Rajiv Jain and Curtis Wigington

Adobe Research

{rajivjain, wigington}@adobe.com

**Abstract** — State-of-the-art methods for document image classification rely on visual features extracted by deep convolutional neural networks (CNNs). These methods do not utilize rich semantic information present in the text of the document, which can be extracted using Optical Character Recognition (OCR). We first study the performance of state-of-the-art text classification approaches when applied to noisy text obtained from OCR. We then show that fusing this textual information with visual CNN methods produces state-of-the-art results on the RVL-CDIP classification dataset.

**Keywords:** Classification; Document Image; Multimodal

## I. INTRODUCTION

The task of document image classification can be viewed as categorizing a given document into a set of defined classes. This capability is important in the context of document management systems where metadata for the class type can be useful in the organization or retrieval of documents in large collections. It can also help automate document image workflows by routing a document when classes of interest are detected.

The success of Convolutional Neural Networks (CNNs) for object classification in computer vision ([1], [2]) has led to several studies exploring its use for document image classification ([3], [4], [5]). With large datasets now available for document image classification [3], CNN architectures used in computer vision have become the state-of-the-art approach for this task [6].

While documents contain rich visual information such as its layout, fonts, and figures, one major difference between natural imagery and document images is that the underlying content often contains a much larger amount of text that can be extracted through Optical Character Recognition (OCR). However, there has been relatively little research into how to integrate textual semantic information into visual CNN approaches for document image classification.

This paper first explores text classification on text extracted from OCR and then explores whether combining both text and visual modalities of the document can improve the accuracy of CNN architectures for document classification. This work makes the following contributions:

1. State of the art performance on the RVL-CDIP Tobacco Corpus using both spatial and feature fusion of text and visual modalities.
2. Results demonstrating text classification models using pre-trained language models can provide substantial performance gains even on noisy text from OCR.

The remainder of the paper is laid out as follows. Section II provides an overview of related work, Section III presents our text classification approach, Section IV presents approaches for multimodal document classification, Section V contains the experimental setup, Section VI provides our results, and Section VII provides conclusions and future work.

## II. RELATED WORK

### A. Document Image Classification

Recent studies into document image classification can be split into classical approaches that use hand coded features and more recent work that has explored deep convolutional neural networks with greater success. A review of many of the classical approaches can be found in the survey paper from [7].

Currently, state-of-the-art methods for document image classification utilize CNNs. Kang et al. [4] introduced CNNs to the task of document classification using a custom network and significantly decreasing the error rate in comparison to classical feature based approaches. Harley et al. [3] extended this further by first introducing the RVL-CDIP collection, a large dataset of 400,000 documents and 16 classes. They showed that the AlexNet CNN architecture used for object classification could achieve 90% accuracy for document classification. Others have followed this work by exploring more complex neural network architectures [5], [8], [9], or by exploring augmentation, architecture settings, and other hyperparameters to improve performance [10]. The current state of the art for this task belongs to Das et al. [6]. They followed the experimental procedure of Harley et. al, but instead used a VGG-16 [1] architecture and obtained an accuracy of 92.2% using an ensemble combining classification scores from sub-regions as well as the full image.

### B. Text Classification

Text classification has been widely studied due to applications such as spam filtering, sentiment classification, and topic categorization. Traditional methods have used a bag of words or character n-grams followed by a classifier [11]. While these approaches are often still competitive for many tasks when distinctive words are sufficient to separate the classes, recent research has explored word embeddings and deeper neural network models to capture a stronger semantic representation of the text with greater success.

Current baselines include an average of pretrained word embeddings such as Word2Vec [12] or GLOVE [13]. These approaches provide a better semantic classification rather than exact word distributions. Kim et al. [14] extended this further by building a multi-layer CNN with Word2Vec features for

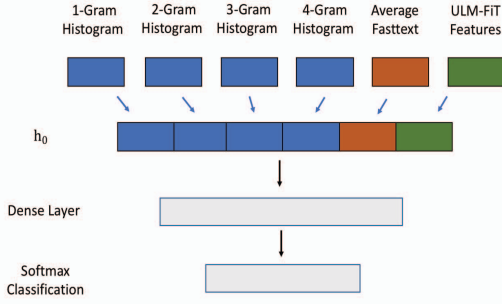


Figure 1. An illustration of the ensemble text classification approach.

each word as input for sentence classification. This was outperformed by Zhang et al. [15] by utilizing a character CNN to handle out of vocabulary words. Joulin et al. [16] introduced FastText, a shallow linear model that incorporates sub-word embeddings, with results that were competitive with deeper neural networks. Linguistic approaches such as ULMFiT [17] and ELMO [18] utilize language models pre-trained on a large corpus of text and achieve large performance gains over previous approaches.

### C. Multimodal Document Image Classification

A few studies have explored multimodal approaches for tasks related to document image classification, generally finding that the combination of both modalities is beneficial to performance. Augereau et al. [19] combine a bag of visual words and bag of text words features leading to increased performance on classification of industrial documents. Noce et al. [20] explored structural embedding of textual information by augmenting the image with colors representing textual topics for use with a CNN for classification. Rusinol et al. [21] showed an improvement from late fusion using pixel intensity features combined with text features from a topic model for page stream segmentation. This was extended by Wiedemann et al. in [22] using a concatenation of convolutional word and image features for improved page stream segmentation. To our knowledge, this is the first detailed study of multimodal document classification on the public RVL-CDIP dataset.

## III. OCR TEXT CLASSIFICATION

Unlike many text classification datasets, a unique challenge for text classification in the context of document images is that text extracted using OCR can have a substantial number of character or word errors or even missing text. The current state of the art for text classification involves classifying contextual features from language models trained on large quantities of text and may be sensitive to OCR errors. Recent studies [23], [24] have shown that a combination of n-gram or word embeddings can improve classification performance. Hence, we utilize 3 text features that represent text at the sequence, word, and character level and then use an ensemble method for classification that combines these underlying features.

For sequence level features, we use ULMFiT [17], a state-of-the-art approach for text classification that utilizes transfer

learning to adapt language model features for classification. It is built around the AWD-LSTM [25] language model that is first trained in a self-supervised manner on a large external corpus and then fine-tuned on the training dataset used for classification. Classification is performed using the hidden states of the LSTM model. Similar to the original ULM-FiT paper, we represent the text in a document using  $V_U$ . A sequence of text with  $T$  words that have hidden states  $H = \{h_1, h_2, \dots, h_T\}$  is represented by the concatenation of the following 3 features:

$$V_U = [h_T, \text{maxpool}(H), \text{meanpool}(H)] \quad (1)$$

We use the default parameters from the ULMFiT paper with  $h$  having a dimensionality of 400.

FastText embeddings [16] are used to represent word level features for text classification. We use the 300-dimensional word vectors pretrained on the common crawl corpus provided by the authors with sub-word embeddings to add robustness for out of vocabulary words [26]. We use  $V_F$  to denote the average FastText feature vector for all of the words in the document.

Finally, character N-grams are utilized to capture a character-level signal for each document. They are a classical text feature that have been shown to be effective for classification and retrieval of noisy text [11],[27]. These simple features are still competitive with more recent approaches for some datasets [16]. We utilized a normalized histogram of character N-gram occurrences in the document ( $V_{C_N}$ ), with values of  $N$  including  $\{1, 2, 3, 4\}$ .

In order to combine these character, word, and sequence features for classification, we created an ensemble model. First, we added a dense layer for each of the individual features and concatenated them as shown in Equation (2). Here,  $\varphi$  is the tanh activation function and  $W_i$  is a weight matrix that projects the individual features into  $\mathbb{R}^{50}$ .

$$h_0 = [\varphi(W_i V_i + b_0) \mid \forall_{i \in \{V_U, V_F, V_{C_1}, V_{C_2}, V_{C_3}, V_{C_4}\}}] \quad (2)$$

Next,  $h_0$  is followed by a single dense layer with 256 units and softmax layer to perform the classification. An illustration of our text classification architecture can be seen in Figure 1.

## IV. MULTIMODAL FUSION

The taxonomy of multimodal fusion is generally split into early fusion, which combines feature representations of the modalities, and late fusion, which combines the predictions from separate classifiers for each modality [28]. This work focuses on early fusion to combine a textual representation with a visual CNN classifier. Section A presents an early spatial fusion method that projects textual features into the spatial domain for input into a CNN classifier. Section B presents four early feature fusion methods that fuse textual and visual features obtained from the second to last layer from their respective classifiers. We compare these approaches to a baseline late fusion approach, which averages the predictions from classifiers built for each modality.

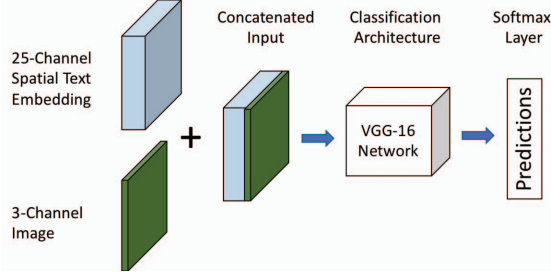


Figure 2. Illustration of the multimodal spatial fusion combining the visual and text modalities into a joint 2-dimensional input.

Recent studies [5],[6] have compared several variants of CNNs that have been applied in computer vision and found that the VGG-16 [1] network produces state-of-the-art results for document image classification. Hence the VGG-16 network is used as our image classification network architecture for both fusion approaches.

#### A. Spatial Fusion

The goal of spatial fusion is to create a combined text and image input representation that can be stacked together as input into a CNN. However, images are two dimensional pixel values while text is typically treated as a one dimensional sequence, making it difficult to directly perform early fusion with their raw representations. In the case of document image classification, the text produced through OCR comes directly from the document. Hence the spatial location of the text can be used to embed a feature representation corresponding to the 2-dimensional bounding box of where the text occurs in the document. This can be seen as providing richer semantic information correlated with the coarse visual representation used by feeding a resized image into a CNN network.

In order to create this 2-dimensional spatial text input, we first represent each word in the document with an  $N$  dimensional embedding vector. Next, a zero initialized tensor with dimensionality  $H \times W \times N$ , where  $H$  and  $W$  are the height and width of the resized image are used for document image classification. For each of the words in the document, we set the tensor value corresponding to the bounding box coordinates of the word to the word's  $N$ -dimensional feature representation. This is similar to the approach taken by Noce et al. [20], but rather than trying to embed the text representation into three color channels, which can suffer from quantization, we instead use the full  $N$ -dimensional feature representation of the text. The 3-channel ( $H \times W \times 3$ ) input image is concatenated with the  $H \times W \times N$  text representation to create an  $H \times W \times (N+3)$  multimodal spatial representation of the input. This multi-channel input can then be used for classification with a CNN as shown in Figure 2.

#### B. Multimodal Feature Fusion

The goal of multimodal feature fusion is to combine a feature vector representation of both the text ( $V_T$ ) and visual ( $V_I$ ) modalities for joint classification. The vector representation for each modality is obtained using the activations from the second to last layers of both the text and image classifiers. Note that  $V_I$ , could come from the VGG-16 network using either the actual document image or spatial

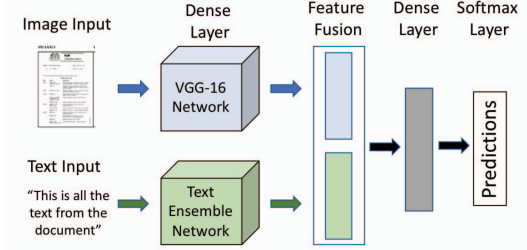


Figure 3. Illustration of the multimodal feature fusion combining a feature representation of the visual and text modalities.

fusion input presented in the previous section. We explore 4 early feature fusion approaches described as follows:

**Concatenation:** The image and text features are stacked together.

$$F_{Concat} = [V_T, V_I] \quad (3)$$

**Addition:** Each of the elements of both features are added together. This operation is preceded by a dense layer projecting the features into the same dimensionality.

$$h_T = \tanh(W_T V_T + b_T) \quad (4)$$

$$h_I = \tanh(W_I V_I + b_I) \quad (5)$$

$$F_{Add} = h_T + h_I \quad (6)$$

**Multimodal Compact Bilinear Pooling:** Multimodal bilinear pooling captures interactions between two modalities by calculating the outer product between the two vectors as shown in Equation (7). This was first applied in [29], but due to the size of the full outer product we choose to use the compact bilinear version of this algorithm presented in [30], [31], which uses the count sketch operation to reduce the output dimensionality.

$$F_{Bilinear} = V_T \otimes V_I \quad (7)$$

**Multimodal Gated Units:** The multimodal gated unit is first described by Arevalo et al. in [32]. It uses a self-attention mechanism to learn a gating function to combine multimodal features prior to their use for classification. The gating function and its output  $F_{Gated}$  can be described with the following equations, where  $h_T$  and  $h_I$  from Equations (4) and (5) project the image and text feature representations into the same dimensionality. The gating mechanism  $z$  controls how much each modality contributes to the classification.

$$z = \sigma(W_z * [V_I, V_T]) \quad (8)$$

$$F_{Gated} = z \cdot h_T + (1 - z) \cdot h_I \quad (9)$$

To perform classification, each of the feature fusion approaches is followed by a dense layer and a softmax layer for prediction as illustrated in Figure 3. As a baseline, we also compare to an average of the independent classifier predictions for each modality, which we refer to as  $F_{AVG}$ . This can be viewed as a baseline late fusion technique.

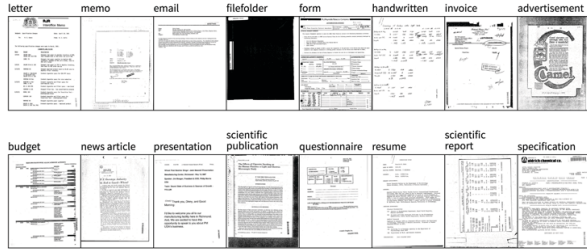


Figure 4. Examples of the 16 classes in the RVL-CDIP dataset.

## V. EXPERIMENTAL SETUP

### A. Dataset and Evaluation Metric

The RVL-CDIP [3] collection is used to evaluate our approach for multimodal document image classification. The dataset consists of 400,000 document images from 16 different classes and provides a training set of 320,000 images as well as a validation and test set each with 40,000 images. Examples of the 16 classes can be seen in Figure 4. In order to train our fusion approaches, we utilize the Tesseract OCR [33] engine to extract text from each of the pages, keeping the bounding box for the location of each word. In order to measure performance on the dataset and compare our results, we report the classification accuracy for each of the classes.

### B. Implementation Details

For our experiments, the neural networks were trained using stochastic gradient descent with learning rates ranging from .001 to .00001. We also use a decaying step function similar to [5]. We train the network on a single 1080TI card for 25 Epochs, keeping the model with the highest validation accuracy. More detailed implementation details for each of the components of the study is given below.

**OCR Text Classification:** We first preprocess the text by converting it to lowercase and removing extra spaces and line breaks. For character N-grams we utilize this raw text directly. Similar to [23] we use a hashing function to allow a histogram to represent all words within a fixed size. We choose [50,2500,20000,20000] for the 1,2,3,4-gram histogram sizes respectively. For FastText and ULMFiT, we clean the text with a script provided by the authors of FastText, which preserves common punctuation, converts the digits to words and removes all other non-alphabetic characters.

**Image Classification:** We resize all the images to be 224x224 and follow the training procedure in [5] to train the VGG-16 architecture with pretrained ImageNet weights. Since RVL-CDIP consists of single-channel images, we copy the greyscale channel to produce a 3-channel input compatible with the ImageNet pre-trained weights.

**Multimodal Spatial Fusion:** Due to computational resource limitations of the GPU used for training, a 25-dimensional embedding is used to limit the input dimensionality. We chose FastText as the word embedding representation due to its ability to represent out of vocabulary words in a low

Approach	Accuracy
Word CNN [14]	77.5
FastText – 25D [16]	78.9
FastText – 300D bigram [16]	79.7
Bag of Words [15]	80.3
Char CNN [15]	81.6
Char-Ngram [15]	82.4
ULMFiT [17]	85.8
<b>Ensemble Approach (Ours)</b>	<b>86.4</b>

Table 1. OCR text classification accuracy on the RVL-CDIP dataset showing ULMFiT and our ensemble perform best.

dimensional space. Given the 28-channel concatenated image and text multimodal input, it is difficult to directly reuse pretrained VGG-16 weights that were built for 3-channel inputs due to the discrepancy in the first layer of the architecture. In order to use the pre-trained weights, we set the convolutional filter weights corresponding to the 25-channel text representation in the first layer to zero, while keeping the 3-channel image weights. We found this to train better than using an entirely random initialization.

**Multimodal Feature Fusion:** While the four feature fusion approaches are trainable end-to-end within a neural network, we found it was more stable during the training if we trained the text and image networks classifier independently and then fine-tuned the fusion classifier with both. We believe this is due to the different learning rates needed by each modality. For the feature fusion classification, we use a dense layer size of 256 units with ReLU activations.

## VI. EXPERIMENTS AND RESULTS

We performed three experiments on the RVL-CDIP dataset with our text classification and multimodal fusion approaches. First, we compared the accuracy of our ensemble text classification from Section III model to other state of the art text classification approaches to understand the text classification performance when applied to OCR. Next, for the second experiment we measured the accuracy of the spatial fusion on the RVL-CDIP dataset. Finally, we measured the accuracy of the feature fusion in the third experiment. Examples of our results can be seen in Table 6.

### A. OCR Text Classification

We compare our proposed ensemble text classification approach in Section III to a number of other recently proposed text classification approaches in Table 1. Since results using previous techniques were not reported on this dataset, we use public implementations with default settings where available. For the Bag-of-words and Bag-of-N-grams we use TF-IDF histogram followed by a fully connected layer with tanh activations and a softmax layer as implemented in [15].

The results show that a language model based approach such as ULMFiT can provide large increases in accuracy, even when applied to noisy text produced by OCR. The ensemble method from Section III, provides a small improvement over ULMFiT showing that much of the discriminative signal provided by the additional FastText and



VGG-16 Input	Accuracy
Spatial Word Embedding	84.7
Image + Word Presence Channel	91.0
Image (baseline)	91.1
<b>Image + Spatial Word Embedding</b>	<b>92.3</b>

Table 2. Spatial fusion accuracy compared to using the image, spatial text input or word presence input.

character-Ngram features is captured through ULMFiT. We chose the classification model with the lowest validation error rate for the remainder of the fusion experiments, which was the ensemble approach.

### B. Multimodal Spatial Fusion

The results for the multimodal spatial fusion can be found in Table 2. In our experiments, we compare the performance of our 28-channel spatial fusion input with the image-only input (3-channel) and spatial text tensor input (25-channel). To provide a better intuition of whether it is the semantic information from the word embeddings or knowledge of the word’s spatial positions that impacts classification most, we also report results from a third comparison by substituting the  $N$  channel spatial text representation with a single Boolean channel that captures whether or not a word is present at a given pixel location.

The classification performance for the spatial fusion of text features outperforms the 25-D FastText classification shown in Table 1, showing the importance of word placement for classification. The spatial fusion of text and image input reduces the error rate by 12.4% over the baseline method, which only used the image as input. This is in contrast to using the “Image + Word Presence” results that show only using the spatial positions of the word without the word embedding did not increase the classification accuracy. This provides evidence that incorporating finer grained semantic information from word embeddings with the raw pixel input in the spatial domain is responsible for the increase in classification performance.

### C. Multimodal Feature Fusion

Table 3 shows the results for each of the four fusion approaches described in Section IV B. The fusion is performed with the second to last layer from both the VGG-16 and text classification ensemble network from Section III. We experimented and reported results using either a 3-channel

Fusion Approach	Accuracy	
	Image Input	Spatial Fusion Input
Image Input (baseline)	91.1	92.3
$F_{AVG}$	92.8	93.3
$F_{Gated}$	92.9	93.0
$F_{Bilinear}$	93.4	93.4
$F_{Concat}$	93.5	93.5
$F_{Add}$	<b>93.6</b>	<b>93.5</b>

Table 3. Feature fusion accuracy with image-only and spatial fusion input.  $F_{Add}$  has a 28.1% relative improvement over the baseline.

Class Label	F1-Score			Improvement (Absolute / Relative)
	Text Only	Image Only	Fusion $F_{Add}$	
letter	89.7	89.7	92.2	2.5 / 24.3%
form	82.3	83.5	86.9	3.4 / 20.6%
budget	86.6	90.3	93.1	2.8 / 28.9%
invoice	91.2	92.1	94.2	2.1 / 26.6%
presentation	81.7	84.6	88.5	3.9 / 25.3%
questionnaire	89.7	87.7	92.1	4.4 / 35.8%
resume	98.1	96.1	98.5	2.4 / 61.5%
memo	91.8	93.7	95.4	1.7 / 27.0%
email	97.5	98.6	98.9	0.3 / 21.4%
handwritten	75.5	95.1	95.3	0.2 / 4.1%
advertisement	69.6	93.0	93.8	0.8 / 11.4%
sci. report	82.6	81.3	88.2	6.9 / 36.9%
sci. pub.	92.1	94.1	95.1	1.0 / 16.9%
specification	93.6	94.0	96.1	2.1 / 35.0%
file folder	70.5	95.0	94.9	-0.1 / -2.0%
news article	86.7	92.7	93.6	0.9 / 12.3%

Table 4. Class specific F1-Score for the image, text, and  $F_{Add}$  feature fusion. Several classes show large relative improvements.

image or spatial fusion input into the VGG-16 network. The best performing method is the addition of the text and visual features using the image-only input, with an accuracy of 93.6. This represents a 28.1% reduction in error over the baseline approach. The accuracy of the overall fusion approach does not increase when combining the spatial fusion and feature fusion. This could be due to the richer textual features used in the feature fusion as well as overlap in the feature representation. The bilinear, concatenation, and addition fusion approaches perform similarly on the dataset. This indicates that richer interaction between the multimodal features provided by compact bilinear pooling do not add additional discriminative capability for this experiment. Further inspection of the gated fusion showed that its attention mechanism was overweighting the image features leading to the decreased performance.

Class specific F1-Scores for the image-only, text-only, and feature fusion ( $F_{Add}$ ) classifiers are shown in Table 4. Classes

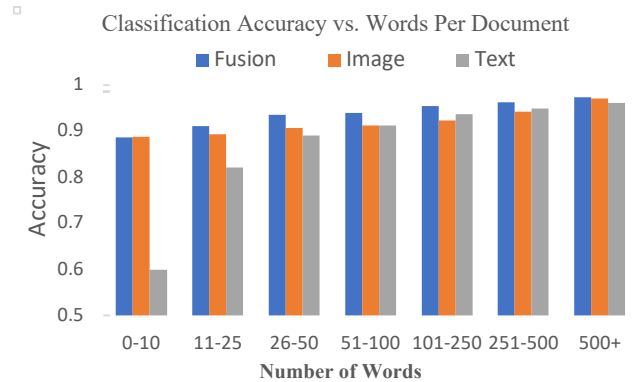


Figure 5. This chart comparing accuracy and # of words shows the benefit of multimodal fusion increases with the number words in the document.

Approach	Description	Accuracy
Hartley et. al [3]	AlexNet Architecture	89.8
Tensmeyer et. al [10]	AlexNet with larger images and spatial pyramid pooling	90.9
Csurka et al. [8]	Google LeNet Architecture	90.7
Afzal et al. [5]	VGG-16 Architecture	91.0
Das et al. [6]	Ensemble of VGG-16 networks using sub regions	92.2
Our Baseline	VGG-16 Architecture	91.1
<b>Spatial Fusion (ours)</b>	<b>Image &amp; spatial text embedding with VGG-16</b>	<b>92.3</b>
<b>Feature Fusion (ours)</b>	<b>Image + text feature fusion using addition</b>	<b>93.6</b>

Table 5. Comparison of our fusion approach with the existing state of

that benefit the most from the fusion (resume, questionnaire, scientific report) are cases where the image classification actually outperforms the text classification and have large amounts of text. Similarly, cases where fusion improves the least (handwritten, advertisement, file folder) have the lowest text classification accuracy and typically involve documents with very little or even no legible text.

In order to understand the effect of the number of words found in the document with the classification accuracy we create the bar chart in Figure 5 for the text-only, image-only, and feature fusion ( $F_{Add}$ ) classifiers. In the context of OCR errors, we define a word here to be a contiguous set of alpha numeric characters separated by punctuation or spaces. When there are less than 10 words, the text classification performance is much lower than the image classification, and hence there is little improvement from the fusion. When there are between 100-500 words, text-only classification actually outperforms image-only classification, leading to larger relative increases in the fusion performance.

#### D. Comparison to the State of the Art

Table 5 compares our proposed multimodal document classification with other approaches that have reported results on the RVL-CDIP dataset. Both our spatial fusion and feature fusion approaches outperform the existing state of the art presented by Das et al. [6]. Multimodal feature fusion leads to a 1.4% absolute improvement in accuracy or 17.9% reduction in error over their approach. Their work used an ensemble of image classifiers while our study only used a single VGG-16 image classifier. It is possible that the use of a stronger visual classifier such as their ensemble method could further improve the overall fusion in the future.

## VII. CONCLUSION

This study demonstrates that fusion of semantic information from OCR with visual information from the document image can increase the performance of visual-only

document image classification algorithms. Both early spatial and feature fusion techniques improve over a baseline visual CNN classifier and outperform the state of the art reported on this dataset. We also showed that newer text classification techniques that adapt pre-trained language models can lead to large performance gains for text classification. We believe that future work using actual probabilities from the OCR model rather than using it as a black box may lead to additional accuracy gains.

## REFERENCES

- [1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv Prepr. arXiv1409.1556*, 2014.
- [2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NuerIPS*, 2012, pp. 1097–1105.
- [3] A. W. Harley, A. Ufkes, and K. G. Derpanis, "Evaluation of deep convolutional nets for document image classification and retrieval," in *ICDAR*, 2015.
- [4] Y. Li, D. Doermann, P. Ye, L. Kang, and J. Kumar, "Convolutional Neural Networks for Document Image Classification," in *ICPR*, 2014, pp. 3168–3172.
- [5] M. Z. Afzal, A. Kölsch, S. Ahmed, and M. Liwicki, "Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image classification," in *ICDAR*, 2017, vol. 1, pp. 883–888.
- [6] A. Das, S. Roy, U. Bhattacharya, and S. K. Parui, "Document Image Classification with Intra-Domain Transfer Learning and Stacked Generalization of Deep Convolutional Neural Networks," in *ICPR*, 2018, pp. 3180–3185.
- [7] N. Chen and D. Blostein, "A survey of document image classification: Problem statement, classifier architecture and performance evaluation," *IJDAR*, 2007.
- [8] G. Csurka, D. Larlus, A. Gordo, and J. Almazan, "What is the right way to represent document images?," *arXiv Prepr. arXiv1603.01076*, 2016.
- [9] A. Kolsch, M. Z. Afzal, M. Ebbecke, and M. Liwicki, "Real-Time Document Image Classification Using Deep CNN and Extreme Learning Machines," in *ICDAR*, 2018, pp. 1318–1323.
- [10] C. Tensmeyer and T. Martinez, "Analysis of convolutional neural networks for document image classification," in *ICDAR*, 2017, vol. 1, pp. 388–393.
- [11] W. B. Cavnar and J. M. Trenkle, "N-gram-based text categorization," in *Symposium on document analysis and information retrieval*, 1994, vol. 161175.
- [12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NuerIPS*, 2013, pp. 3111–3119.
- [13] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.
- [14] Y. Kim, "Convolutional neural networks for sentence classification," *EMNLP*, pp. 1746–1751, 2014.
- [15] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *NuerIPS*, 2015, pp. 649–657.
- [16] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *Eur. Chapter Assoc. Comput. Linguist.*, pp. 427–431, 2017.
- [17] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *ACL*, vol. 1, pp. 328–339, 2018.
- [18] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *ACL*, vol. 1, pp. 2227–2237, 2018.
- [19] O. Augereau, N. Journet, A. Vialard, and J.-P. Domenger, "Improving classification of an industrial document image database by combining visual and textual features," in *Document Analysis Systems*, 2014, pp. 314–318.

- [20] L. Noce, I. Gallo, A. Zamberletti, and A. Calefati, “Embedded Textual Content for Document Image Classification with Convolutional Neural Networks,” in *ACM Symposium on Document Engineering*, 2016, pp. 165–173.
- [21] M. Rusiñol, V. Frinken, D. Karatzas, A. D. Bagdanov, and J. Lladós, “Multimodal page classification in administrative document image streams,” *IJDAR*, 2014.
- [22] G. Wiedemann and G. Heyer, “Page Stream Segmentation with Convolutional Neural Nets Combining Textual and Visual Features,” *arXiv Prepr. arXiv1710.03006*, 2017.
- [23] J. A. Botha, E. Pitler, J. Ma, A. Bakalov, A. Salcianu, D. Weiss, R. McDonald, and S. Petrov, “Natural language processing with small feed-forward networks,” *EMNLP*, 2017.
- [24] D. Kiela, C. Wang, and K. Cho, “Dynamic meta-embeddings for improved sentence representations,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 1466–1477.
- [25] S. Merity, N. S. Keskar, and R. Socher, “Regularizing and optimizing LSTM language models,” *ICLR*, 2017.
- [26] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, “Advances in Pre-Training Distributed Word Representations,” in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018.
- [27] R. Jain, D. W. Oard, and D. Doermann, “Scalable ranked retrieval using document images,” in *Document Recognition and Retrieval XXI*, 2014, pp. 9021–9030.
- [28] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019.
- [29] T.-Y. Lin, A. RoyChowdhury, and S. Maji, “Bilinear cnn models for fine-grained visual recognition,” in *CVPR*, 2015, pp. 1449–1457.
- [30] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, “Compact bilinear pooling,” in *CVPR*, 2016, pp. 317–326.
- [31] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, “Multimodal compact bilinear pooling for visual question answering and visual grounding,” *EMNLP*, 2016.
- [32] J. Arevalo, T. Solorio, M. Montes-y-Gómez, and F. A. González, “Gated multimodal units for information fusion,” *arXiv Prepr. arXiv1702.01992*, 2017.
- [33] R. Smith, “An overview of the Tesseract OCR engine,” in *ICDAR*, 2007, vol. 2, pp. 629–633.

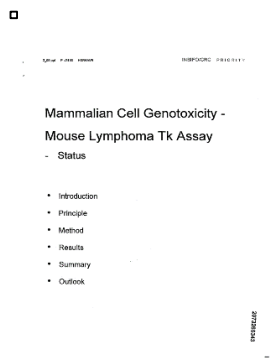
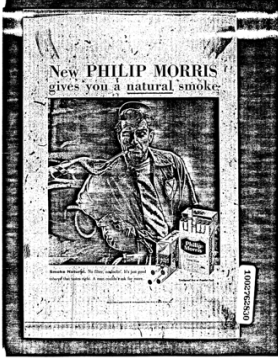
Document Image	OCR Text	Ground Truth	Predicted Classes
	2soppt P~et00 SRW INBIFOCRC PRIORITY Mammalian Cell Genotoxicity - Mouse Lymphoma Tk Assay - Status * Introduction ° Principle * Method ° Results * Summary * Outlook Evesszeoz	Scientific Report	Image: Presentation Text: Scientific Report Fusion: Scientific Report
	» ° So 5 R @ a o	Advertisement	Image: Advertisement Text: File Folder Fusion: Advertisement

Table 6. Examples from the RVL-CDIP dataset where the image-only or text-only classification predicted incorrectly, but the fusion technique predicted correctly. These examples show the challenges with noisy OCR and binarized images.