A. Implementation Details

In this section, we provide full implementation specifics of UniTNT and divide it into three parts – (1) architecture; (2) training procedure; (3) Scene-text information.

A.1. Architecture

We harness the model agnosticism of UniTNT and apply it to two top-performing VL models. Specifically, we utilize the publicly-available code bases of ALBEF [12] and BLIP [11] and apply our method to them. We design our approach in a modular way enabling simple integration into existing models. Below we list the architectural specifics for both UniTNTALBEF and UniTNTBLIP.

OCR Encoder  We use a pretrained BERT-base [6] as our encoder and introduce it with 2-dimensional information, as can be seen in Equation 1. Specifically, we use three separate embedding layers (i.e., torch.nn.Embedding) for the word token and its x and y axis positions for both the OCR and the question. In particular, we define the minimal and the maximal spatial position as 0 and 1000, respectively, and set these values for the question tokens (referred to as “pseudo-2D information” in the main paper). We restrict the number of OCR and question token lengths to 128 and 35, respectively. Next, we sum the 2D-related embeddings and pass them in a 2-layer MLP with a hidden dimension of 768 for additional processing. Finally, we multiply it by $\alpha$ (set to 0.1) and sum it with the token representation to obtain the final one fed into the encoder.

\[ \text{OCR Encoder} \]

Figure 1: OCR prevalence in VQAv2. Histogram of the number of OCR instances per-image in VQAv2 dataset.

VL-OCR Decoder  In order to introduce the pretrained decoder with scene-text information, we create new OCR Cross Attention (OCR-CA) blocks and place them in parallel to the existing VL ones. Such newly added components are identical to the existing ones and initialized with the pretrained weights of the latters'. To fuse the outputs of the OCR CA and the VL CA, $F_{\text{OCR}}$ and $F_{\text{VL}}$, we concatenate them along the channel dimension and pass them via attention based 2-layers MLP with a hidden size of 768 to obtain $F_{\text{attn}}$, an attention map that multiplies $F_{\text{OCR}} \odot F_{\text{attn}}$. Namely, this mechanism highlights the important and meaningful features in $F_{\text{OCR}}$ and masks the less relevant ones. Then, we pass the multiplication output via a learnable gating module (by multiplying it by $\tanh(\beta)$, where $\beta$ is learnable and initialized to 0), aimed to gradually blend the OCR features into the existing VL one.

\[ \text{VL-OCR Decoder} \]

---

*Work done during an Amazon internship.

1 https://github.com/salesforce/ALBEF
2 https://github.com/salesforce/BLIP
3 https://huggingface.co/docs/transformers/model_doc/bert
A.2. Training Procedure
We train all of our models to minimize $L_{\text{UniTNT}} = L_{\text{base}} + \alpha_1 L_{\text{OCR-LM}} + \alpha_2 L_{\text{OCR-BC}}$ using 8 A100 GPUs, where $\alpha_1$ and $\alpha_2$ are hyperparameters.

Visual Question Answering We train both UniTNT_{ALBEF} and UniTNT_{BLIP} on a unified Text-Non-Text VQA dataset, containing VQAv2 [1], TextVQA [16] and ST-VQA [3] for 10 epochs using a batch size of 8 and 16 for ALBEF and BLIP, respectively. Moreover, we set $\alpha_1 = \alpha_2 = 1$ and keep the other training-related hyperparameters as in the original papers.

Image Captioning We train UniTNT_{BLIP} on a unified Text-Non-Text CAP dataset, comprised of COCO Captions [17], TextCaps [15], for 5 epochs with batch size of 32. We set $\alpha_1 = 0.05$ and $\alpha_2 = 0$ since contrary to VQA, CAP does not contain textual information available both in training and inference time, making it infeasible to implement OCR-BC. Moreover, we keep the rest of the hyperparameters as in BLIP.

A.3. Scene-text information
As specified in the paper, we extract the scene-text information (word tokens and 2-dimensional position) for all the VQA and CAP datasets (both the general and scene-text counterparts) using Amazon Text-in-Image. To better understand the prevalence of OCR in the non-scene-text datasets, we plot the statistics of OCR in VQAv2 in Fig. 1 (same images are in COCO Captions as well). While a large portion of the images does not contain text in them, there is a large amount of such with OCR (38.36% and 38.03% of train and test images contain OCR). Since OCR conveys meaningful information, it sheds light on the significant improvement of UniTNT up his baselines (ALBEF and BLIP).

B. Datasets

B.1. Visual Question Answering
VQAv2 contains 204,721 images (82,783, 40,504, and 81,434) from COCO [13], 1,105,904 questions (443,757, 214,354, and 447,793), and 6,581,110 answers (4,437,570, 2,143,540, and the test answers are held-out). Answering the questions requires vision-language understanding and commonsense knowledge. Each question has ten ground-truth answers.

TextVQA contains 28,408 images from OpenImages [10], 45,336 questions and 453,360 ground-truth answers. The annotators were instructed to formulate questions that require reasoning from the text in the image. As in VQAv2, each question has 10 ground-truth answers.

ST-VQA is a fusion of computer-vision datasets – ImageNet [5], VizWiz [2], Visual Genome [9], IIIT Scene Text Retrieval [14], ICDAR 2013 [8], ICDAR 2015 [7] and COCO Text [17]. It contains 31K questions, split into training (26K) and testing (5K), requiring scene-text understanding.

B.2. Image Captioning
COCO Captions contains over one and a half million captions describing over 330,000 images from the COCO dataset. Each image has five human-generated captions.

TextCaps is composed of 28,408 images and 142,040 captions (5 captions per image). The images are from the TextVQA dataset, and the captions are based on the text in the image. Specifically, models have to reason over the scene-text information to generate correct captions.

C. The Impact of Training Data
In this section, we study the effect of the different combinations of training datasets and report our findings in Tab. 1. In particular, we experiment with UniTNT and BLIP in Visual Question Answering and Image Captioning using separate training on vision-oriented and OCR-oriented datasets and combined training. In VQA, using both dataset types leads to the best standalone and average performance in the tested benchmarks. This attests to the symbiosis between general and scene-text-oriented VQA, encouraging avoidance of the common practice of separate finetuning.

However, using a unified training set in CAP leads to the best COCO Captions and average results, but not in TextCaps. Specifically, separate finetuning on TextCaps achieves a CIDEr score of 130.5, compared to 119.1 in the combined training. It corresponds with [15], which shows that combining COCO Captions with an upsampled version of TextCaps reduces the model’s performance on the former. It is because while training on TextCaps encourages the model to insert OCR into the caption, training on COCO Captions which barely contains OCR in its captions, penalizes such behavior, leading to an intrinsic tradeoff. To better understand the effects of training models solely on TextCaps, we qualitatively test them on COCO Captions. Notably, we finetune both BLIP and UniTNT of TextCaps and demonstrate their performance on COCO Captions in Fig. 2. Our analysis shows that as TextCaps contains OCR in all its captions, separate finetuning causes models to fixate on OCR, regardless of their importance. Moreover, in images without an OCR signal, the models sometimes hallucinate text in the image. While both models showcase similar behavior, since UniTNT has better scene-text understanding, it is more prone to such phenomena. It is also expressed in Tab. 1, where BLIP and UniTNT trained
Table 1: The impact of training data. We show the effect of each dataset configuration for training UniTNT and BLIP.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BLIP</td>
<td>✓</td>
<td>✓</td>
<td>40.16</td>
<td>30.12</td>
<td>35.14</td>
<td>84.8</td>
<td>112.7</td>
<td>98.8</td>
</tr>
<tr>
<td>UniTNT&lt;sup&gt;+&lt;/sup&gt;BLIP</td>
<td>✓</td>
<td>✓</td>
<td>37.01</td>
<td>50.19</td>
<td>43.60</td>
<td>70.4</td>
<td>130.5</td>
<td>100.5</td>
</tr>
<tr>
<td>BLIP&lt;sup&gt;*&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td>76.39</td>
<td>20.36</td>
<td>38.45</td>
<td>83.3</td>
<td>59.4</td>
<td>96.4</td>
</tr>
<tr>
<td>UniTNT&lt;sup&gt;+&lt;/sup&gt;BLIP</td>
<td>✓</td>
<td>✓</td>
<td>79.68</td>
<td>36.33</td>
<td>58.01</td>
<td>133.7</td>
<td>59.6</td>
<td>96.7</td>
</tr>
<tr>
<td>BLIP&lt;sup&gt;*&lt;/sup&gt;</td>
<td>✓</td>
<td>✓</td>
<td>77.40</td>
<td>32.43</td>
<td>54.92</td>
<td>83.4</td>
<td>101.4</td>
<td>117.4</td>
</tr>
<tr>
<td>UniTNT&lt;sup&gt;+&lt;/sup&gt;BLIP</td>
<td>✓</td>
<td>✓</td>
<td>79.90</td>
<td>55.21</td>
<td>67.56</td>
<td>134.0</td>
<td>119.1</td>
<td>126.6</td>
</tr>
</tbody>
</table>

on TextCaps obtain 84.8 and 70.4 on COCO Captions, respectively. Despite the improved performance on TextCaps when performing separate finetuning on it, our findings highlight its drawbacks. Thus, we claim that also in CAP, combined training should be applied.

From a general view, we hypothesize that since numerous valid captions exist for a given image, both with and without OCR, the model struggles to decide whether to use the OCR in its caption. Due to the datasets’ sizes in combined training that favors the vision-oriented ones, the model opts to reduce its use of OCR, not fully maximizing its performance on TextCaps. It is contrary to VQA, where the conditioning over the question makes it easier for the model to decide whether to use OCR or not (e.g., "What is written in the sign?" versus "What color is this shirt?").

D. Qualitative Analysis

Visual Question Answering We provide an additional qualitative demonstration of UniTNT and compare it to BLIP and M4C on both TextVQA validation set (Fig. 3) and VQAv2 test set (Fig. 4). We depict in the four left-most columns success-cases and the rightmost, fail cases, and color in green the correct answers and red, incorrect ones. Moreover, we divide the figures such that the upper part corresponds with the benchmark’s goal (VQAv2 – see, TextVQA – read) and the lower one with the counterpart goal (VQAv2 – read, TextVQA – see). These results further demonstrate that UniTNT is capable of reasoning over both visual and scene-text information, while other competing methods perform unsatisfactorily on at least one of the benchmarks. Moreover, as the visualizations in Fig. 4 testify, granting scene-text understanding also benefit VQAv2, corresponding with the quantitative evidence in the main paper. It is demonstrated in the bottom part of the figure, where the OCR is crucial for answering the questions or providing meaningful information that facilitates answering them.

Image Captioning Similar to the VQA demonstration, we present a visualization of UniTNT performance on TextCaps (Fig. 5 and COCO Captions (Fig. 6) and compare the performance to M4C and BLIP. On the left columns, we show images where our method outperforms the other methods, and on the right, its failure cases. Moreover, we list the CIDEr scores of each prediction and color in green the highest one. These findings attest that BLIP is incapable of incorporating scene-text information, which results in relatively low CIDEr results. Interestingly, M4C is too overfitted for TextCaps, causing it to fail completely on COCO Captions where OCR is scarce. Specifically, it focuses on the OCR regardless of their importance (e.g., the third example in the last row of Fig. 6) and thus provides an irrelevant caption. Despite the intrinsic tradeoff described in the paper between TextCaps and COCO Captions, UniTNT is capable of providing adequate captions for both benchmarks. Specifically, our method is the only one to cope satisfactorily on both benchmarks altogether and is capable of harnessing both scene-text and visual information.
Figure 2: Qualitative demonstration of the effects of finetuning on TextCaps. BLIP and UniTNT results of COCO Captions when finetuned solely on TextCaps. In some cases, scene-text understanding helps the models, but it also leads to over-reliance on the OCR signal and even to the hallucination of OCR. While such phenomena occur in both models, it is more prevalent in UniTNT due to its better scene-text understanding.
Figure 3: Qualitative demonstration on TextVQA validation. UniTNT, M4C, and BLIP answers, containing both success (left) and fail (right) cases of our method on image-question pairs that require mainly reading (top) and ones that require also visual reasoning (bottom).
Figure 4: **Qualitative demonstration on VQAv2 test.** UniTNT, M4C, and BLIP answers, containing both success (left) and fail (right) cases of our method on image-question pairs that require mainly vision (top) and ones that require also scene-text understanding (bottom).
Figure 5: **Qualitative demonstration on TextCaps.** UniTNT, M4C-Captioner, and BLIP answers, containing both success (left) and fail (right) cases of our method alongside the per-caption CIDEr score.
Figure 6: Qualitative demonstration on COCO Captions. UniTNT, M4C-Captioner, and BLIP answers, containing both success (left) and fail (right) cases of our method alongside the per-caption CIDEr score.
References


