Generating Text from Images in a Smooth Representation Space

Graham Spinks and Marie-Francine Moens

Department of Computer Science, KU Leuven, Leuven, Belgium
graham.spinks@cs.kuleuven.be
sien.moens@cs.kuleuven.be

Abstract. A methodology is described for the generation of relevant captions for images of an extensive medical dataset in the ImageCLEF 2018 Caption Prediction competition. Automatic and accurate textual descriptions of images could help relieve workload pressure for specialists and assist clinical professionals in multiple areas. Instead of generating textual sequences directly from images, we first learn a smooth, continuous representation space for the captions. Subsequently the task is reduced to the minimization of the mapping loss from image to continuous representation through the use of a deep convolutional neural network. We illustrate how our system learns to generate captions by aligning relevant embeddings. The submitted run achieves a score of roughly 13.76% and ranks 4th out of the 5 participating teams. The top submission in the competition achieved a score of 25.01%.

1 Introduction

In this paper, our participation is described in the 2018 ImageCLEF Caption Prediction task [5] which is a part of the 2018 ImageCLEF competition [6]. The goal is to regenerate the original caption for a set of images where the caption is essentially a concise textual interpretation of the content of the image. The dataset consists of 4 million diverse images that cover a range of radiology/clinical data and was collected from open access biomedical journal articles (Pubmed Central). No additional external data was used for this submission.

A large amount of effort is dedicated in medical fields to correctly interpret and describe various images. Automation of this process might help reduce the bottleneck in certain diagnosis pipelines and help medical professionals focus on more important tasks.

Generating captions from images is also a task that requires an understanding of data representations in neural networks. The cross-modal nature of the task implies a successful alignment of visual and textual data. The nature of these modes are quite distinct as continuous images usually demand different processing techniques than discrete texts.

While the ImageCLEF competition also contains a concept detection subtask on the same dataset, this submission focuses on directly generating captions without any additional intermediate steps. This approach has the advantage

that the text generation doesn't depend on any pre-fabricated conceptual labels and is directly inferable from images.

Current text-to-image systems that employ neural networks typically combine a Convolutional Neural Network (CNN) with a discrete decoder in the form of a Recurrent Neural Network (RNN), which in practice is often a Long Short-Term Memory (LSTM) network [3][9][10]. The difficulty of such approaches often lies in the discrete nature of natural language sentences. Back-propagation is challenging for such data as the gradient of the error becomes infinite on the boundary of discrete symbols.

In order to alleviate this problem, our approach starts by creating a smooth continuous code space for text, which is characterized by a coherent local structure where similar inputs are mapped to nearby codes. This contrasts with autoencoders that simply learn an identity mapping with unstructured latent representations. The advantage is that complex modifications can be made to the text while traversing the data manifold for slightly modified sentences. In order to obtain such a representation we use an Adversarially Regularized Autoencoder (ARAE) [8] which trains a discrete autoencoder in an adversarial setting.

In a subsequent step, we align the images to the continuous data manifold of the captions rather than to the discrete natural language. This has the benefit that in this stage we avoid the complex and costly discrete decoder step which is present in traditional image-to-text systems. Once image and text representation are aligned, we can decode the aligned vector with the decoder we obtained in the previous step, thus obtaining natural language text for each image.

We will show that our method creates a textual representations space from which the input can easily be reconstructed. By aligning the visual input to this space, we create varied captions for the images and obtain a score of 13.76% on the test set.

2 Methodology

We will briefly mention how the data was prepared before discussing the creation of the text representation as well as the caption generation. An overview of the entire methodology is presented in figure 1.

2.1 Data Preprocessing

In order to simplify the caption generation task, all words are converted to lowercase while any words that appear less than 100 times in the entire dataset are replaced by out-of-vocabulary markers. The remaining vocabulary contains 4303 different words. Any captions that exceed the length of 15 words are capped while any captions that are shorter are padded.

All images are randomly cropped to achieve data augmentation and transformed to 256x256 resolution. The images are normalized with a mean and standard deviation of 0.5.

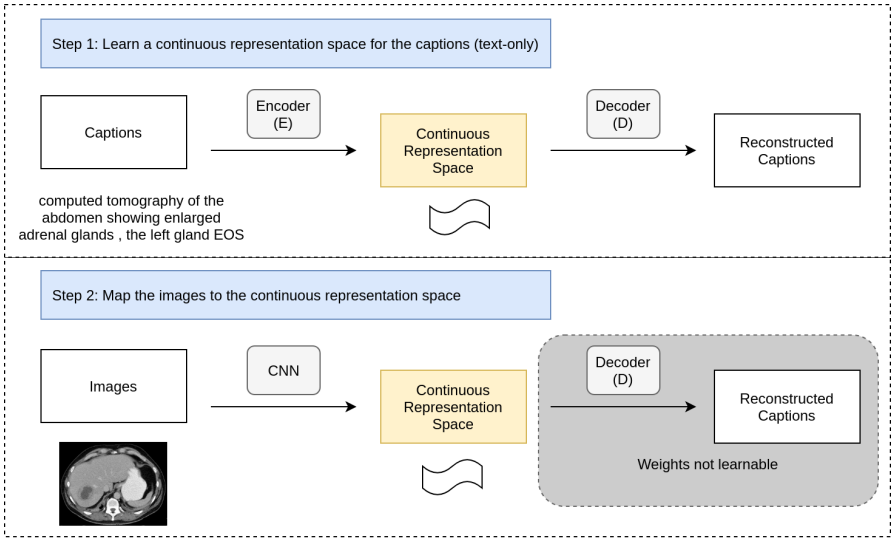

Fig. 1. Overview of the methodology. The encoder (E) of the ARAE model creates a textual representation for all captions which can be decoded to the original input with the decoder(D). In a second step each image is mapped to the continuous representation space with a CNN. D, for which the weights are frozen in this step, decodes the mapping to a caption for each image. CT image reference [4].

2.2 Text Representation

While there are several methodologies to create dense continuous representations for discrete structures, each comes with both advantages and disadvantages. One might consider for example vector-based word or sentence embeddings that are trained by predicting the context of a word or one might simply use an autoencoder that reconstructs the original text from a compact representation. While the word or sentence embeddings capture basic semantic information their performance in additional tasks is often quite limited. Autoencoders do create a dense representation but the learned representation space is not smooth [8].

For this task, we will use an ARAE [8] to construct smooth, continuous representations of the sentences. Such representations have been shown to lie in a smoother contracted codespace than a typical autoencoder, with the benefit that similar inputs are mapped to nearby codes. The ARAE combines the training of a generator (G) and critic (C) of a Generative Adversarial Model (GAN) [2] as well as an encoder (E) and decoder (D) of a regular discrete input autoencoder. In this setup, E creates a continuous text representation $\hat{t}$ of the input text, while D uses a cross-entropy loss to try to recreate the original sentence. Additionally, G is a 2-layer feedforward network that learns how to generate realistic representations $\hat{t}$. C estimates the Wasserstein distance between the generated and real distribution as defined in the Wasserstein GAN (WGAN) [1], such that G is explicitly trained to minimize that distance. As a side-result of this setup G eventually learns to

create diverse texts with a low perplexity score [11]. To keep the overview concise, only E and D are shown for the ARAE model in figure 1.

For this competition, the ARAE model of [8] is modified by passing the discrete integer list of text inputs to both the generator and critic after normalization between -0.5 and 0.5. This is done to encourage the ARAE to learn an even smoother version of the code space as the critic, C, learns to identify when a representation doesn't match a text input. We also slightly adapt the softmax-temperature parameter which influences the extent to which a code of a text is different than that of another. To increase variability we use a softmax temperature of 0.1 rather than 1.0 when calculating the cross-entropy loss. In order to obtain good generalization we perform early stopping and select the model where the reconstruction error on the validation set is minimal.

Using only the captions in the training set, a smooth manifold for the captions is thus created with the above model. In essence, each caption now has an equivalent continuous representation $\hat{t}$, from which the original caption can be reconstructed. With such a representation, an alignment can be learned between the visual features and the relevant captions for each image, as explained in the next subsection.

2.3 Image-to-Text

In order to map the visual features to the continuous space we created for the captions, the input images are passed through a deep neural network that consists of 8 convolutional blocks. One such block contains a convolutional layer followed by a batch norm layer and a LeakyReLU activation function. At the end of the network another convolutional layer and two fully connected layers are added.

The output is then compared to the continuous textual representation of the caption as constructed in section 2.2. In our experiments we devised two methods to determine how suitable a text is for each image.

The first method simply uses a loss function derived from the cosine similarity between two embeddings. The output of the CNN network is then trained to be as similar to the continuous text code, $\hat{t}$, as possible.

The second method essentially does the same but runs the output of the CNN through the decoder, D, that was trained before (see section 2.2). The generated output distribution is subsequently compared to that of the one generated by the continuous representation of the original caption. For the comparison we use the same cosine similarity metric as before. The reasoning behind this approach is that performance might improve after decoding the representation to individual time-steps as more information for alignment is available.

For both approaches, the weights of D are not updated during this stage.

3 Results

A first important task in our system is to create textual representations that can be decoded to match the original text with enough accuracy. In order to

Table 1. Examples of preprocessed captions and their reconstruction from the autoencoder of the ARAE. EOS indicates the end-of-sentence marker while OOV indicates the out-of-vocabulary marker.

Original	**Reconstruction**
computed tomography of the abdomen showing enlarged adrenal glands , the left gland EOS	computed tomography of the abdomen showing a left gland with with liver kidney EOS
microscopic examination of the tumor specimen by hematoxylin and eosin stain revealed that EOS	microscopic findings of the tumor showed showed hematoxylin and eosin staining ; that EOS
ultrasound of the right upper quadrant showing the gallbladder free of stones blue EOS	axial image the right kidney quadrant showing a common with wall the . EOS
secondary electron image OOV of a fractured surface of an OOV lingual bar EOS	sem structure photomicrograph of of a representative surface showing the OOV showing view EOS

do so, we perform preprocessing on the text as detailed in section 2.1 and train an ARAE model to encode and decode the sentences as detailed in section 2.2. For a trained ARAE, we demonstrate a range of good and bad examples of the original and decoded sentences in table 1.

The captions are subsequently encoded and the convolutional net is tasked with finding the optimal caption for each image. The captions are generated from the output representations with a greedy method and are evaluated using the script provided by ImageCLEF which measures the result as a percentage of the maximum BLEU score over all sentences. Before calculating the score, stemming is performed and stopwords and punctuation are removed.

The evaluation score is the percentage of the obtained BLEU score over all sentences compared to the maximum possible BLEU score. Thus if one would achieve the best possible BLEU score for each sentence, the score would be 100%. Using a cosine similarity metric, we reach an accuracy of roughly 13.5% on the validation set when comparing the continuous embeddings directly (method 1 in section 2.3). If we use method 2, i.e. after passing the embeddings through the decoder, we obtain an accuracy of roughly 12.4%. For the ImageCLEF submission we only submitted the results of method 1, which obtained a score 13.76% on the test set, indicating that the system didn't overfit on the validation set.

Note that since we are using a cutoff of 15 words per caption, the maximum obtainable score is roughly 36.4% for such captions, as measured on our ground truth validation set. While the performance does improve over the first epochs, it turns out that the network is not able to line up the different embeddings with high accuracy. In fact the output sentences evolve to quite similar output where only some details are modified as illustrated in table 2.

This research provides an interesting direction for new image-to-text systems as there are several possible avenues for improvements. In a first step, training a stable model for larger captions might provide an immediate boost in perfor-

Table 2. Examples of output texts for different input images. While several captions are quite similar overall, some details are usually slightly modified.

Output examples
figure 2 a fundus photograph of the right eye showing a large and
figure showing a mass in the right and the uterus and ovaries
computed tomography scan showing a large mass in the right kidney and ureter
computed tomography scan of the abdomen showing a large mass in the right

mance as more relevant textual output can be aligned with the images, thus obtaining higher BLEU scores. Another possibility to improve the results is to investigate different distance functions. While in this paper, a simple cosine embedding loss was used, this type of alignment might benefit from a measure that expresses a distributional divergence, such as the Wasserstein distance [1]. Finally, besides using an ARAE, other methods that create continuous representations might be more suitable for this type of alignment. For example, one might consider building a representation that includes concept labels or is constructed with image alignment in mind, such as the char-CNN-RNN representation [7].

4 Conclusion

We present an alternative approach to caption generation by leveraging continuous representations for text that were learned with an ARAE model. Images are aligned to the continuous representations rather than discrete natural language. Measured as a percentage of the obtained BLEU scores over all sentences compared to the maximum possible BLEU score, this methodology achieves 13.76% on the submitted run and offers a promising avenue for follow-up research. The proposed setup can be a starting point for implementations with alternative network configurations and text representations that aim to enhance and exceed the obtained results.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. arXiv preprint arXiv:1701.07875 (2017)
2. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. pp. 2672–2680 (2014)
3. Hasan, S.A., Ling, Y., Liu, J., Sreenivasan, R., Anand, S., Arora, T.R., Datla, V., Lee, K., Qadir, A., Swisher, C., et al.: PRNA at ImageCLEF 2017 caption prediction and concept detection tasks (2017)
4. Hellerhoff: Leberabszess - CT axial PV.jpg, CC by 3.0
5. García Seco de Herrera, A., Eickhoff, C., Andrearczyk, V., , Müller, H.: Overview of the ImageCLEF 2018 caption prediction tasks. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)

6. Ionescu, B., Müller, H., Villegas, M., de Herrera, A.G.S., Eickhoff, C., Andrearczyk, V., Cid, Y.D., Liauchuk, V., Kovalev, V., Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Dang-Nguyen, D.T., Piras, L., Riegler, M., Zhou, L., Lux, M., Gurrin, C.: Overview of ImageCLEF 2018: Challenges, datasets and evaluation. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), LNCS Lecture Notes in Computer Science, Springer, Avignon, France (September 10-14 2018)
7. Kim, Y., Jernite, Y., Sontag, D., Rush, A.M.: Character-aware neural language models. In: AAAI. pp. 2741–2749 (2016)
8. Kim, Y., Zhang, K., Rush, A.M., LeCun, Y., et al.: Adversarially regularized autoencoders for generating discrete structures. arXiv preprint arXiv:1706.04223 (2017)
9. Liang, S., Li, X., Zhu, Y., Li, X., Jiang, S.: ISIA at the ImageCLEF 2017 image caption task (2017)
10. Lyndon, D., Kumar, A., Kim, J.: Neural captioning for the ImageCLEF 2017 medical image challenges (2017)
11. Spinks, G., Moens, M.F.: Generating continuous representations of medical texts. In: Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2018)