Evaluating and interpreting caption prediction for histopathology images

Renyu Zhang ZHANGR@UCHICAGO.EDU
Department of Computer Science
University of Chicago
Chicago, IL, U.S.

Christopher Weber CHRISTOPHER.WEBER@UCHOSPITALS.EDU
Department of Pathology
University of Chicago
Chicago, IL, U.S.

Robert Grossman ROBERT.GROSSMAN@UCHICAGO.EDU
Department of Medicine & Computer Science
University of Chicago
Chicago, IL, U.S.

Aly A. Khan AAKHAN@UCHICAGO.EDU
Department of Pathology
University of Chicago
Chicago, IL, U.S.

Abstract

The automatic generation of captions from medical images can provide for an efficient way to annotate histopathology images with natural language descriptions. Such large-scale annotation of medical images may help facilitate image retrieval tasks and standardize clinical ontologies. In this work, we focus on developing and methodically evaluating a new caption generation framework for histopathology whole-slide images. We introduce PathCap, a deep learning multi-scale framework, to predict captions from histopathology images using multi-scale views of whole-slide images. We demonstrate that our framework outperforms a standard baseline caption model on a diverse set of human tissues and provides interpretable contextual cues for understanding predicted captions. Finally, we draw attention to a novel dataset of histopathology images with captions from the Genotype-Tissue Expression (GTEx) project, providing a valuable dataset for the machine learning and healthcare community to benchmark future caption prediction and interpretation methods.

1. Introduction

In the last century, advances in clinical pathology, such as biospecimen fixation, staining, and digital microscopy, have enabled the routine digitization of histopathology slides Bera et al. (2019). Histopathological images contain rich clinical diagnostic information. For example in colonic biopsies, there is architectural information, including crypt abnormalities, and distribution of inflammatory cells, providing insight to disease processes. Anatomic pathologists have developed their own specialized language and lexicon to communicate these descriptive findings. Automatically describing the content of an image is a grand challenge in machine learning, requiring integration of computer vision and natural language processing disciplines. Accurate machine learning methods for generating and visualizing captions from histopathology images have several important potential applications, including (1) supporting pathologists by providing caption prompts and visual cues to help facilitate clinical review, and (2) enabling image retrieval tasks, for example, on archival histopathology slide images missing specific labels or descriptions.

The characterization of fine-grained features that distinguish various morphological and pathological classifications are primarily obtained through expert visual assessment of histopathology images, often requiring experts to spend a significant number of years training and refining their visual skills Brugnara et al. (1994). At the same time, new machine learning techniques for automatically generating natural language descriptions from histopathology images have received limited attention, nor have publicly available benchmark datasets been established for histopathology caption prediction tasks. In this work, we have sought to methodically evaluate if it is possible to generate short, clinically relevant descriptions (captions) from H&E histopathology whole-slide images automatically and propose a benchmark dataset for the machine learning community (Figure 1).

Deep neural networks can learn fine-grained features directly from raw images in a supervised machine learning setting and have already achieved great success in several complex tasks involving histopathology images, including tissue classification Bejnordi et al. (2017), disease outcome prediction Mobadersany et al. (2018), and prediction of genetic alterations Coudray et al. (2018). However, standard machine learning techniques for caption prediction present two non-trivial obstacles: First, histopathology images are often composed of more than 1 billion pixels (gigapixel), which limits most off-the-shelf deep neural networks models due to memory limitations. The rescaling of high-resolution images to overcome memory limitations can result in loss of contextual and spatial information, impeding the generation of relevant descriptions from whole-slide histopathology images. Second, methods for evaluating and visually interpreting predicted captions are needed to facilitate wide-spread clinical adoption. Thus, the combined task of predicting and then interpreting caption generation in the context of gigapixel sized images is a technically challenging problem that is largely unique to the healthcare domain.

We introduce a new multi-scale view framework called PathCap that initially clusters high-resolution tiles from histopathology whole-slide images, and then combines a single low-resolution thumbnail view of the whole-slide image with tiles randomly sampled from high-resolution clusters. Our experiments show that PathCap can effectively access and integrate information from both high-resolution and low-resolution views. We tested our framework on data from the Genotype-Tissue Expression (GTEx) project Consortium et al. (2017), and evaluated our predictions with a pathologist to uncover limitations and opportunities in caption prediction from histopathology images.

Generalizable Insights about Machine Learning in the Context of Healthcare

In this work, we developed a novel model that integrates multiple resolution views of gigapixel histopathology images in order to generate short, clinically relevant natural language

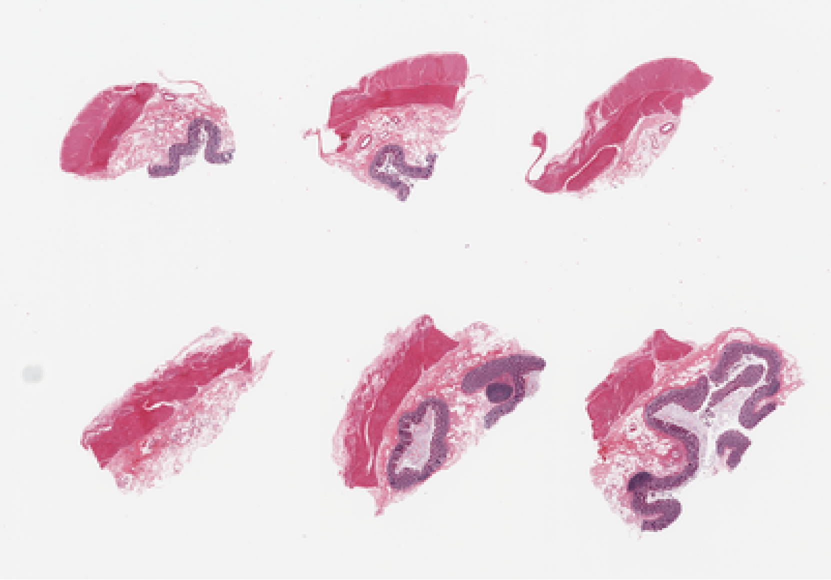


Figure 1: Example slide and caption from GTEx sample GTEX-131XE-0826: *6 pieces; 4 pieces have full thickness elements with well preserved mucosa; 2 have no mucosa (in this section).*

descriptions. We also present a method to produce visually interpretable predictions. Thus, the primary contribution of this study is three fold:

1. We develop a simple framework for harnessing different resolution views of histopathology images for various machine learning tasks, such as caption generation.
2. We demonstrate how tile-level clustering can be harnessed for interpreting predicted captions and obtaining visual cues.
3. We apply our model to a novel dataset from GTEx for caption prediction which contains > 9000 histopathology images and captions from diverse tissues and propose it as a benchmarking dataset for the machine learning and healthcare community.

Our code for this project is publicly available.[1]

2. Related Work

2.1. Image Captioning Models

Early image caption generation focused on detection Kulkarni et al. (2013) followed by template filling. Since the rise of deep learning, most caption generation models have adopted the encoder-and-decoder paradigm Vinyals et al. (2014a), Vinyals et al. (2014b), and Xu et al. (2015). These methods typically use non-medical images, such as real world

[1] https://github.com/zhangrenyuuchicago/PathCap

scenes found in ImageNet Russakovsky et al. (2015). Typically, the encoder is a CNN that extracts features from input images, and the decoder uses an LSTM Hochreiter and Schmidhuber (1997) to generate words step by step. Notably, Xu et al. (2015) incorporated the attention mechanism into the encoder-and-decoder paradigm by feeding an attention weighted combination of features (instead of a CNN extracted single feature) to the LSTM. This approach turned out to be very effective in terms of performance and now defines the standard baseline caption model. However, the visualization and interpretation of the attention weight on the input images can be very ambiguous and non-specific. Subsequent work in the field has focused on further exploiting attention, for example, You et al. (2016) plugged the attention weighted features over semantic concepts into hidden states of LSTM and words generation layers, and Liu et al. (2016) proposed to use instance segmentation to improve the correctness of attention.

More closely related to medical imaging, Zhang et al. (2017) aimed at generating semi-structured pathology descriptions. In order to gain effective gradient flow for training, they utilized a predefined subset of descriptions extracted from the reports. They demonstrated slightly better performance in their experiments compared to a standard baseline caption model. Jing et al. (2017) also adopted a encoder-and-decoder paradigm for X-ray images and developed a hierarchical LSTM model to specifically overcome the challenges of long paragraphs in clinical reports.

Collectively, these methods all require non-trivial changes to adopt to histopathology images due to the lack of instance segmentation information in routine imaging data and robust clinical pathology instance detectors. Furthermore, these methods require rescaling whole-slide images for implementation, causing loss of high-resolution information about the sample tissue and morphology and thus, limiting their ability to utilize full resolution data for generating salient captions.

2.2. Metric Learning

A key step in PathCap involves clustering semantically similar high-resolution tiles from histopathology whole-slide images. In order to cluster tiles within a whole-slide image, we sought to learn embeddings for arbitrary image tiles such that similar tiles have similar embeddings. To accomplish this we used metric learning, which aims to produce a feature space $\mathcal{F}$ with a certain metric structure, where similarity can be captured by some distance function, typically the Euclidean distance Ho et al. (2019). In the context of deep learning, classic metric learning uses no additional layers. Several variants have been proposed Movshovitz-Attias et al. (2017), Sohn (2016), Hadsell et al. (2006), and Chopra et al. (2005); among which triplet loss Schroff et al. (2015), Wang et al. (2014), Bell and Bala (2015), and Weinberger and Saul (2009) is the most popular. In practice, triplets make the training difficult by increasing the sample number cubically. Many methods have sought to accelerate the training Wang et al. (2014), Bell and Bala (2015), Schroff et al. (2015), and Song et al. (2015). In this paper, we follow a simple sampling strategy and objective function (Section 3.2). Qualitative evaluation shows that triplet loss produces embeddings that are consistent with semantic components of tissues inside histopathology images (Figure 2).

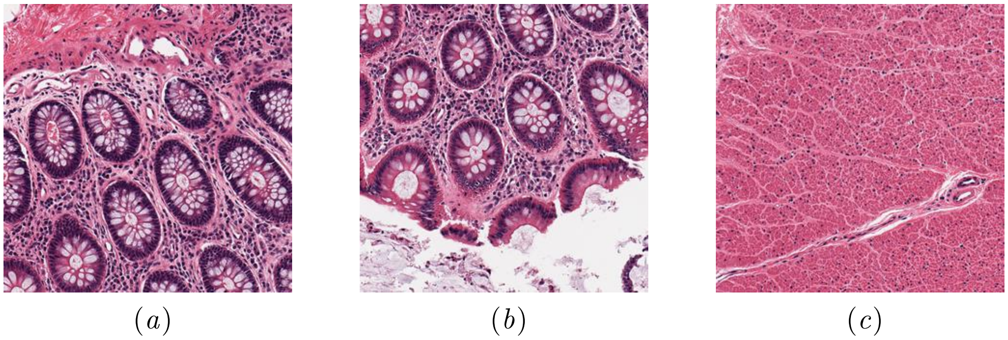


Figure 2: Example tiles used for triplet loss. (a) is the anchor tile showing colonic mucosa, (b) shows predominantly colonic mucosa, and (c) shows mostly smooth muscle (from muscularis propria). (b) and (c) correspond to positive and negative samples respectively for triplet loss.

3. Methods

3.1. Overview

The tissue regions within H&E whole-slide images $\{s^i\}_{i=1}^{M}$ are tiled into non-overlapping sections (1000x1000px) $\{t_j^i\}_{j=1}^{N^i}$. Here M is the number of whole-slide images in the dataset. The N^i is the number of tiles that contain tissues and are extracted from slide s^i. The tissue region is deduced by selecting tiles with an average grayscale pixel value in the range [0.2, 0.7]. An autoencoder is trained on tissue containing tiles $\{t_j^i\}_{j=1}^{N^i}$ using both reconstruction loss and triplet loss. We cluster the tiles $\{t_j^i\}_{j=1}^{N^i}$ extracted from each slide s^i based on the embeddings $\{e_j^i\}_{j=1}^{N^i}$ learned from the autoencoder. For simplicity, we focus our study on k-means, but other clustering approaches can be used as well. K-means takes a set of vectors as input, in our case the embedding produced by an autoencoder, and clusters $\{e_j^i\}_{j=1}^{N^i}$ into K distinct groups $\{C_k^i\}_{k=1}^{K}$ based on a Euclidean distance. Thus, if we fix the cluster number K as 5, the tiles from tissue regions in each histopathology image are clustered into 5 groups.

Next, a rescaled thumbnail b^i and tiles $\{t_k^i\}_{k=1}^{K}$ sampled from each cluster $\{C_k^i\}_{k=1}^{K}$ of a slide s^i are fed into our caption generation model (PathCap) during training and testing. If the cluster number is set to $K = 5$, five tiles, 1 from each cluster, are sampled randomly. The thumbnail b^i is used to initialize the LSTM, and tiles $\{t_k^i\}_{k=1}^{K}$ are fed to the LSTM step by step. Our attention module is based on the sampled tiles. To enable visualization of the attention across the whole slide image, we can show the attention weights over all tiles $\{t_k^i\}_{k=1}^{K}$ from a given cluster $\{C_k^i\}_{k=1}^{K}$. We used PyTorch to implement our model Vinodababu (2019).

3.2. Metric Learning with Triplet Loss

An autoencoder is trained on all tissue containing tiles $\{t_j^i\}$ extracted from all slides $\{s^i\}_{i=1}^{M}$ in the dataset. An autoencoder is an unsupervised method that generates a small com-

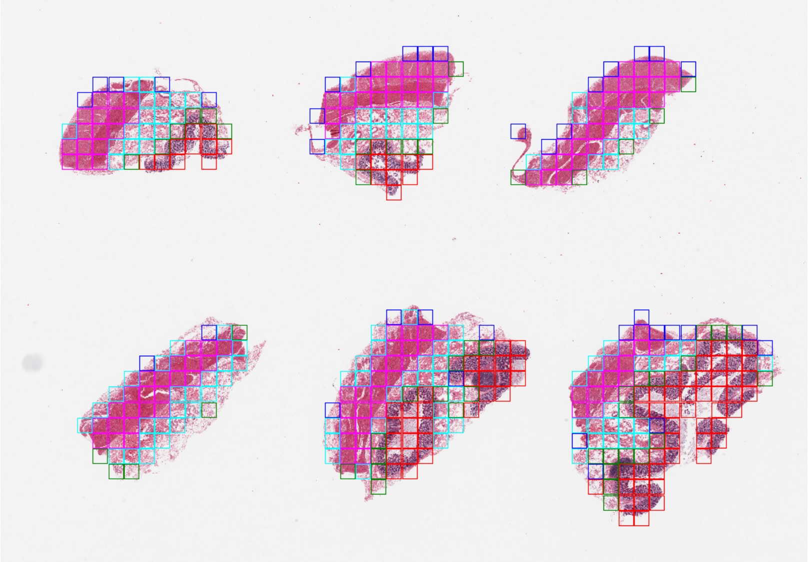


Figure 3: Example clustering visualization. Box color of each tile represents the cluster membership ($K = 5$). The tile cluster colors demonstrate that tiles in a cluster are semantically coherent across and within pieces.

pressed feature representation or embedding for each input sample. These features can capture the variance of the whole dataset while exhibiting a small amount of reconstruction loss. The large amount of tiles extracted from gigapixel histopathology slides make it computationally expensive to process all the tiles from a slide within one single pass. Instead, we randomly sample a limited number of tiles for each slide.

To learn a more robust embedding, in addition to the reconstruction loss, we use triplet loss. Specifically, during the training of the autoencoder, the data loader returns a set of triplet (t^i_j, t^i_k, t^i_l) tiles from each slide s^i. t^i_j is the anchor tile. t^i_k is a positive example of t^i_j. Here we define positive examples as an adjacent tile. t^i_l is a negative example of t^i_j, which means t^i_l is not adjacent to t^i_j (Figure 2). The loss is as follows:

$$L(t^i_j, t^i_k, t^i_l) = \mu \cdot \max(d(e^i_j, e^i_k) - d(e^i_j, e^i_l) + m, 0) + d(t^i_j, D(e^i_j))$$

E is encoder and D is decoder. $e^i_j = E(t^i_j)$. $d(\cdot, \cdot)$ represents the distance. m is the margin and μ is the factor for triplet loss. We use mean squared deviation as the distance.

We train the autoencoder with the Adam method Kingma and Ba (2014). The autoencoder is trained for 4 epochs. After the training of the autoencoder is finished, we use the autoencoder to obtain representations for all the tiles. For each slide, we perform k-means clustering for all the tiles in the slide (Figure 3).

3.3. Neural Network Architecture

Low-resolution thumbnail images are used to initialize the LSTM Hochreiter and Schmidhuber (1997). An attention mechanism on tiles is adopted for each step of generating captions,

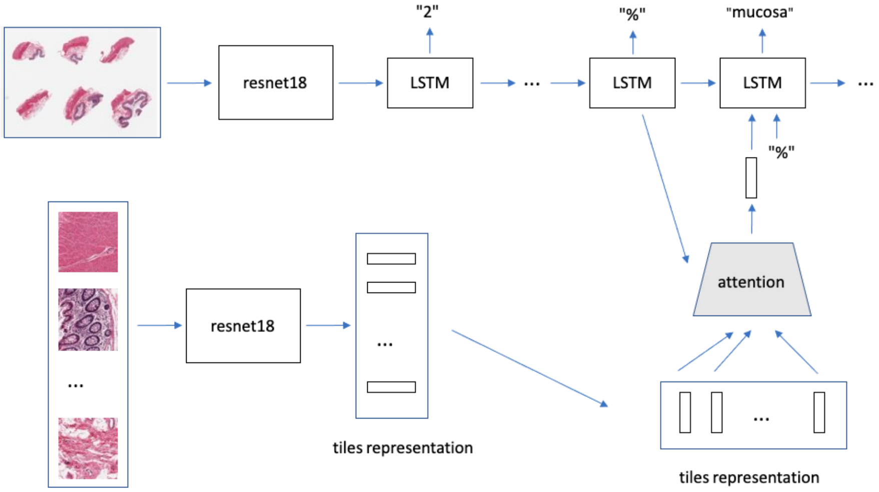


Figure 4: Overall architecture of PathCap. One ResNet-18 is used to extract visual features from the thumbnail of a histopathology image, and pass it to the LSTM. The other ResNet-18 extracts features from randomly sampled tiles from different clusters of the histopathology image, and passes them to the attention module and LSTM step by step.

following the approach from Ilse et al. (2018). Overall, PathCap contains three modules (Figure 4): the thumbnail encoder, tiles encoder, and decoder.

For the thumbnail encoder part, the standard ResNet-18 He et al. (2015) extracts the feature vector from a given input image thumbnail b^i. The feature vector is linearly transformed and then used to initialize LSTM.

The tile encoder contains another ResNet-18 to extract representations from tiles $\{t_k^i\}_{k=1}^K$. Let $H^i = \{h_k^i\}_{k=1}^K$ be a bag of K representations of K tiles from different clusters $\{C_k^i\}_{k=1}^K$ of a slide s^i. The attention-weighted representation z^t at step t for a slide s^i is

$$z^t = \sum_{k=1}^{K} \alpha_k^t h_k$$

where:

$$\alpha_k^t = \frac{\exp(w^T \tanh(V[h_k, m^t])}{\sum_{g=1}^{K} \exp(w^T \tanh(V[h_g, m^t])}$$

m^t is the hidden state of LSTM at step t, and w and V are parameters of two linear layers. $[\cdot,\cdot]$ is the concatenation operation.

For the decoder part of PathCap, source and target texts are predefined. For example, if the image description is “2 pieces, 15% vessel stroma, rep delineated”, the source sequence is a list containing [‘<start>’, ‘2’, ‘pieces’, ‘,’, ‘ 15%’, ‘vessel’, ‘stroma’, ‘rep’, ‘delineated’] and the target sequence is a list containing [‘2’, ‘pieces’, ‘,’, ‘ 15%’, ‘vessel’, ‘stroma’, ‘rep’, ‘delineated’, ‘<end>’]. Using these source and target sequences and the feature vector, the

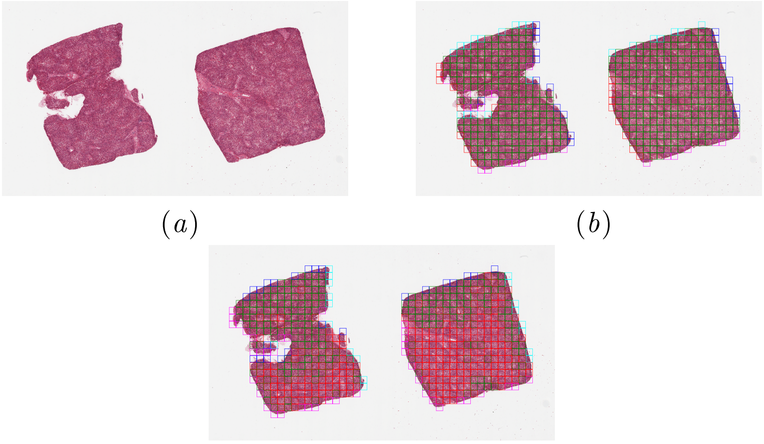


Figure 5: Example tile clustering (K =5) with triplet loss. (a) is the original slide. (b) and (c) show the tile clustering after we train the autoencoder with and without triplet loss respectively. Colors of the boxes show the cluster membership.

LSTM decoder is trained as a language model conditioned on the image feature vector. Notably, we can use the attention mechanism to extract features from sampled tiles and visualize the weights when generating each word of a caption for histology images.

3.4. Data Augmentation and Hyperparameter Settings

Each training slide contained between 10 to 1000 tiles (median 372). During the autoencoder and PathCap training we applied several data augmentations strategies similar to Liu et al. (2017) to improve model robustness. First, we randomly applied left-right and top-down flips. Second, we perturbed color: brightness with a maximum delta of 64/255, saturation with a maximum delta of 0.25, hue with a maximum delta of 0.04, and contrast with a maximum delta of 0.75. The Adam optimizer Kingma and Ba (2014) and validation data was used for parameter learning. Both the ResNet-18 for thumbnails and tiles were fine-tuned with learning rate = 1e-4. The decoder's learning rate was 4e-4. We decay learning rate with factor 0.8 if there is no improvement for 8 consecutive epochs, and terminate training if there is no improvement for 20 consecutive epochs.

4. Cohort

We downloaded all clinical slides from the Genotype-Tissue Expression (GTEx) portal.[2] The GTEx project aims to provide the scientific community a common resource with which to study human gene expression and regulation and its relationship to genetic variation. Notably, the GTEx Portal also provides open access to histopathology imaging data of donor tissue and histopathology notes describing the tissue sample quality (Figure 1).

[2] http://gtexportal.org/home/histologyPage

After selecting slides with captions and removing slides with sparse tissue content, we curated 9727 slide-caption pairs spanning 41 different tissue types. These pairs were randomly split into 7795 training, 948 validation, and 984 sized testing sets.

For the imaging data, we did not use any preprocessing methods on the whole-slide images. All histopathology slide images were subjected to digital tissue segmentation and segmented regions were clipped into non-overlapping 1000x1000px sized sections at 20x magnification. We removed tiles with intensity greater than 0.70 or less than 0.2 to remove the background. For the caption data, all the captions were converted to lowercase. Tokens with less than 5 frequency were removed from the captions, resulting in 971 tokens that cover 95.06% word occurrences in the dataset.

5. Results on Real Data

5.1. Results on Caption Generation

We first compared PathCap to a baseline model, which only takes low-resolution thumbnails as input and uses the Xu et al. (2015) approach (Table 1). For each step generating words, the model follows an attention mechanism and gives a weight for the spatial features extracted from thumbnails by ResNet-18 He et al. (2015). We used the Microsoft COCO Chen et al. (2015) tool to quantitatively compare the performance of models with different inputs. Here we used $beam\ size = 1$ and metrics including BLEU (columns labeled B-1, B-2, B-3, and B-4) Papineni et al. (2002), Meteor Denkowski and Lavie (2014), Rouge-L Lin (2004) and CIDEr Vedantam et al. (2014). We also examined a version of PathCap that only used tiles and without access to a thumbnail view, and found that using tiles alone performed slightly better than the baseline model. Taken together, PathCap, which combines information from high-resolution tile and low-resolution thumbnail views performed the best. All the metrics of PathCap are averaged over 20 rounds of testing.

Table 1: Performance on test set

Method	B-1	B-2	B-3	B-4	METEOR	ROUGE_L	CIDEr
Baseline	0.3822	0.2833	0.1996	0.1377	0.1958	0.4282	0.8936
PathCap	**0.4046**	**0.2986**	**0.2114**	**0.1455**	**0.2059**	0.4290	**0.9038**
Tiles-only	0.3944	0.2905	0.2040	0.1383	0.2032	**0.4312**	0.9003

5.2. Results on Metric Learning

In order to demonstrate the superiority of triplet loss on tile embeddings, we trained two autoencoders. One autoencoder was trained only with reconstruction (mean squared error, MSE) loss. The other autoencoder was trained with reconstruction loss and triplet loss. The encoder part of the autoencoder was composed of two convolutional layers and two maxpooling layers. The output of the encoder (embedding) is of length 460. The decoder part contained three convolutional layers. The μ was set to 0.1, and the margin 0.001. We trained two separate PathCap models with the clusters using the representations from each of the two different autoencoders.

We observed that the two different autoencoders produced qualitatively different tile clusterings (Figure 5). Next, we used the Microsoft COCO Chen et al. (2015) tool again to quantitatively compare the performance of models with different metrics, including BLEU Papineni et al. (2002), Meteor Denkowski and Lavie (2014), Rouge-L Lin (2004) and CIDEr Vedantam et al. (2014). Table 2 shows the performance of our models when we used different metric learning methods for clustering. As above, B-1, B-2, etc. refers to the BLEU score. Overall, we demonstrate both a qualitative improvement in tile-level clustering, and quantitative improvement in caption generation using metric learning.

Table 2: Influence of triplet loss

Loss	B-1	B-2	B-3	B-4	METEOR	ROUGE_L	CIDEr
MSE only	0.3944	0.2878	0.2011	0.1381	0.2005	0.4219	0.8703
MSE & triplet	**0.4046**	**0.2986**	**0.2114**	**0.1455**	**0.2059**	**0.4290**	**0.9038**

5.3. Results on Clustering

In order to explore the influence of cluster number K, we trained models with K from 2 to 5. An autoencoder was trained with reconstruction loss and triplet loss to generate embeddings for tiles extracted from each slide. After training the autoencoders, we generated representations for all tiles and performed k-means clustering using K from 2 to 5. In order to generate confidence intervals, we repeated this process 20 rounds.

For each PathCap trained model for each K, we evaluated our prediction on the testing dataset over 20 rounds. The average metrics over 20 rounds are reported in Table 3. The corresponding 95% confidence interval (CI) for each metric when cluster number = 3 are B-1 [0.3981,0.4111], B-2 [0.2938,0.3035], B-3 [0.2067,0.2162], B-4 [0.1406,0.1504], METEOR [0.2018,0.2100], ROUGE_L [0.4232,0.4348] and CIDEr [0.8598,0.9478]. Overall, our analysis suggests that PathCap is robust to cluster size changes, and demonstrates stable metrics across K from 2 to 5.

Table 3: Performance of PathCap with different cluster number (K)

$K =$	B-1	B-2	B-3	B-4	METEOR	ROUGE_L	CIDEr
2	0.3797	0.2814	0.1976	0.1334	0.1973	0.4249	0.8627
3	**0.4046**	**0.2986**	**0.2114**	**0.1455**	**0.2059**	0.4290	0.9038
4	0.3887	0.2863	0.2003	0.1355	0.1990	0.4280	0.8989
5	0.3885	0.2909	0.2084	0.1447	0.2015	**0.4367**	**0.9621**

5.4. Results on Visualization

PathCap has the advantage of visualizing the caption prediction based on the attention weight given to tiles from a cluster. As reference, visualization using the standard baseline model Xu et al. (2015) is depicted in Figure 6. The visualization and interpretation of the attention weight on the whole-slide images can be very ambiguous and non-specific.

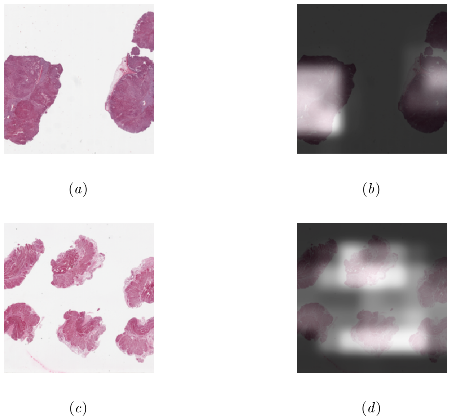


Figure 6: Example of visualizing caption tokens with a standard baseline model Xu et al. (2015). (a) and (c) are the input thumbnails to the model. (b) and (d) show the attention weights when the model generates the “myometrium” and “muscularis” tokens respectively. White/bright indicates more attention weight, black/dark indicates less attention weight.

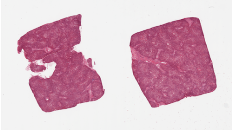
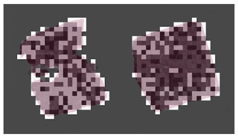
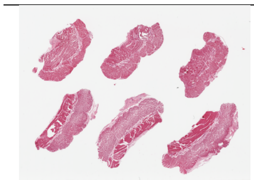
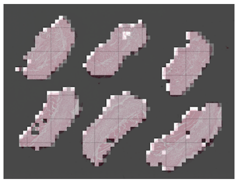
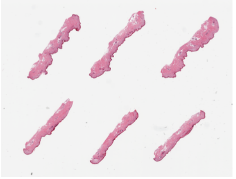
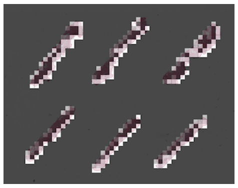
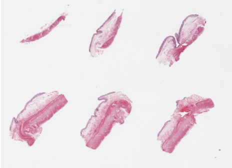
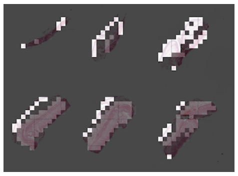
In contrast, with PathCap, an attention mechanism over tile features is deployed for our models. These tiles are sampled from different clusters. The clustering of tiles based on the embeddings learned using triplet loss underlies the potential of better separating the whole slides by small tiles. After the model is trained, weights on different clusters can be shown on the whole slide in the test dataset when the model predicts each word. We observe the model attends at word-level to both the inner parts of the tissue or texture and also the boundaries, depending on the caption context. Examples are shown in the Table 4.

Expert evaluation of the examples demonstrate broadly coherent and interpretable results. In the liver example, the predicted caption and visualization is appropriate for macrovesicular steatosis. Next, for the Esophagus example, the use of the phrase “good specimens” in the prediction is highly subjective and likely an atypical way to annotate specimens. However, the detection of muscularis propria provides improved context relative to the reference caption. For the skin example, the prediction of “5% dermal fat” is appropriate, however the tile clusters visualized for the “fat” token are instead squamous epithelium. Finally, for the colon example, the predicted caption and visualization is correct that the full thickness section contains about 1 mm thickness of colon, but it is mostly an irrelevant measure. Notably, the caption neglected to capture autolytic properties from the autopsy material.

6. Discussion

In this work, we present and examine the complex task of generating short, clinically relevant captions from gigapixel whole-slide histopathology images. We show that clustering tiles based on the embeddings learned using triplet loss allows for coherent segmentation of

Table 4: Visualization of the PathCap method on four test slides from four different tissues. The last column shows some examples of attention weights when the model generates the corresponding tokens. White/bright indicates more attention weight, black/dark indicates less attention weight.

Slide	PathCap Prediction	Reference	Example
 Liver[a] a. GTEx sample ID: 13FLV-0326	2 pieces , diffuse macrovesicular steatosis involves 70 % of parenchyma	2 pieces ; includes portion of capsule (target is 1 cm below capsule) , mild steatosis , passive congestion , focal portal chronic inflammation	 “macrovesicular”
 Esophagus[b] b. GTEx sample ID: 13FTW-1926	6 pieces , up to <unk> ; all muscularis , good specimens	6 pieces ; well trimmed	 “muscularis”
 Skin[c] c. GTEx sample ID: 13NYS-0126	6 pieces ; well trimmed ; 5 % dermal fat	6 pieces ; <unk> epidermis (<unk>) , solar elastosis ; well trimmed , 10 % dermal fat	 “fat”
 Colon[d] d. GTEx sample ID: 13O3P-2326	6 pieces , mucosa up to 1mm , <unk> % thickness	6 pieces ; mucosa autolyzed ; muscularis preserved	 “mucosa”

whole-slide images and results in improved visualization of attention. Thus, our specific technical contribution of clustering tiles within histopathlogy images in order to facilitate downstream tasks, such as caption generation and interpretation, suggests a promising strategy for other machine learning tasks in digital pathology. Finally, we demonstrate the relative effectiveness of PathCap compared to a standard baseline caption prediction approach, and propose the GTEx dataset as a novel benchmark for future caption prediction and interpretation methods.

Limitations We note some important limitations in our work. First, while PathCap achieves better performance over the standard baseline caption prediction method, there is significant room for improvement. Our results confirm that caption generation from histopathology images is a unique and technically challenging problem. Future work in caption prediction could benefit from considering this specific problem setting. Second, we trained and tested our model only on the GTEx data. Due to limitations in publicly available paired caption and histology images, we were unable to evaluate domain adaptation or consider other imaging datasets. Future work should consider evaluating PathCap generated captions on additional datasets as they become available. Third, our captions are short descriptions relating to specimen quality from GTEx (e.g., sample composition). We did not test our model on text from large reports or clinical notes. We hypothesize that integration of a hierarchical LSTM model, such as one proposed by Jing et al. (2017), may be useful for these scenarios.

References

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22): 2199–2210, 2017.

Sean Bell and Kavita Bala. Learning visual similarity for product design with convolutional neural networks. *ACM Trans. Graph.*, 34(4):98:1–98:10, July 2015. ISSN 0730-0301. doi: 10.1145/2766959. URL http://doi.acm.org/10.1145/2766959.

Kaustav Bera, Kurt A Schalper, David L Rimm, Vamsidhar Velcheti, and Anant Madabhushi. Artificial intelligence in digital pathology-new tools for diagnosis and precision oncology. *Nature reviews. Clinical oncology*, 2019.

Carlo Brugnara, Terry Fenton, and James W Winkelman. Management training for pathology residents: I. results of a national survey. *American journal of clinical pathology*, 101 (5):559–563, 1994.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. URL http://arxiv.org/abs/1504.00325.

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1, June 2005. doi: 10.1109/CVPR.2005.202.

GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204, 2017.

Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10):1559, 2018.

Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014.

R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, June 2006. doi: 10.1109/CVPR.2006.100.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Chih-Hui Ho, Pedro Morgado, Amir Persekian, and Nuno Vasconcelos. Pies: Pose invariant embeddings. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9 (8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. *CoRR*, abs/1802.04712, 2018. URL http://arxiv.org/abs/1802.04712.

Baoyu Jing, Pengtao Xie, and Eric P. Xing. On the automatic generation of medical imaging reports. *CoRR*, abs/1711.08195, 2017. URL http://arxiv.org/abs/1711.08195.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, Dec 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.162.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

Chenxi Liu, Junhua Mao, Fei Sha, and Alan L. Yuille. Attention correctness in neural image captioning. *CoRR*, abs/1605.09553, 2016. URL http://arxiv.org/abs/1605.09553.

Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E. Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q. Nelson, Gregory S. Corrado, Jason D. Hipp, Lily Peng, and Martin C. Stumpe. Detecting cancer metastases on gigapixel pathology images. *CoRR*, abs/1703.02442, 2017. URL http://arxiv.org/abs/1703.02442.

Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018.

Yair Movshovitz-Attias, Alexander Toshev, Thomas K. Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. *CoRR*, abs/1703.07464, 2017. URL http://arxiv.org/abs/1703.07464.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, June 2015. doi: 10.1109/CVPR.2015.7298682.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1857–1865. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6200-improved-deep-metric-learning-with-multi-class-n-pair-loss-objective.pdf.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. *CoRR*, abs/1511.06452, 2015. URL http://arxiv.org/abs/1511.06452.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726, 2014. URL http://arxiv.org/abs/1411.5726.

Sagar Vinodababu. a-pytorch-tutorial-to-image-captioning. https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning, 2019.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014a. URL http://arxiv.org/abs/1411.4555.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014b. URL http://arxiv.org/abs/1411.4555.

Jiang Wang, Yang song, Thomas Leung, Chuck Rosenberg, Jinbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 04 2014. doi: 10.1109/CVPR.2014.180.

Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, 10:207–244, June 2009. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1577069.1577078.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015. URL http://arxiv.org/abs/1502.03044.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. *CoRR*, abs/1603.03925, 2016. URL http://arxiv.org/abs/1603.03925.

Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. *CoRR*, abs/1707.02485, 2017. URL http://arxiv.org/abs/1707.02485.