MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network

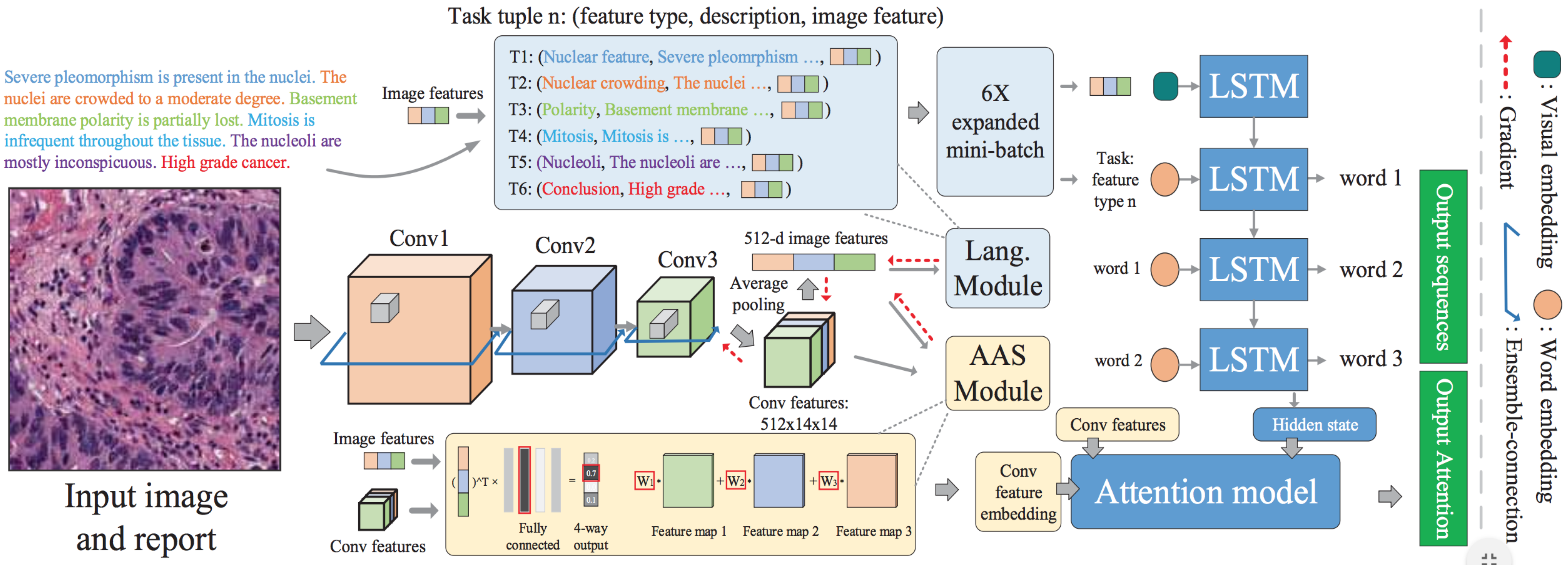
Authors: Zizhao Zhang, et al. from University of Florida

Presented by Qingsong Yang

June 28, 2018

Purpose: A direct multimodal mapping between medical images and diagnostic reports

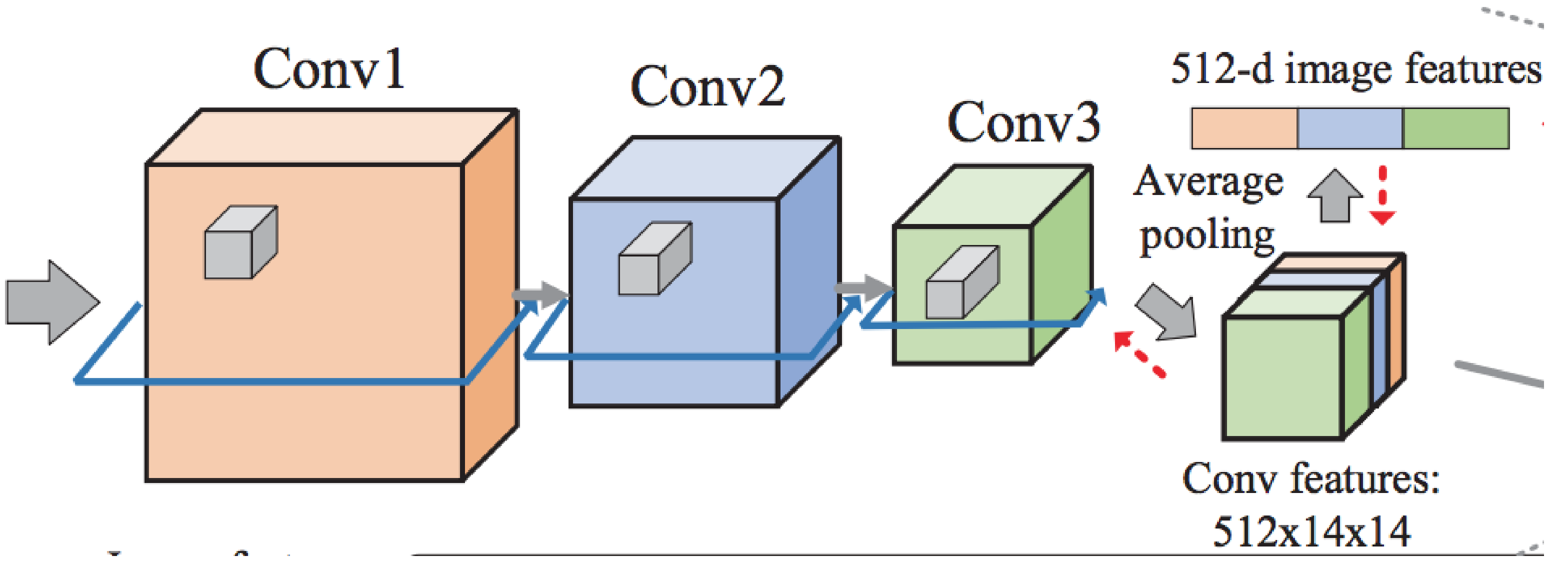
- Read images
- Generate diagnostic reports
- Visualize attention



- Image Module
- Language Module

Image Model

Input image



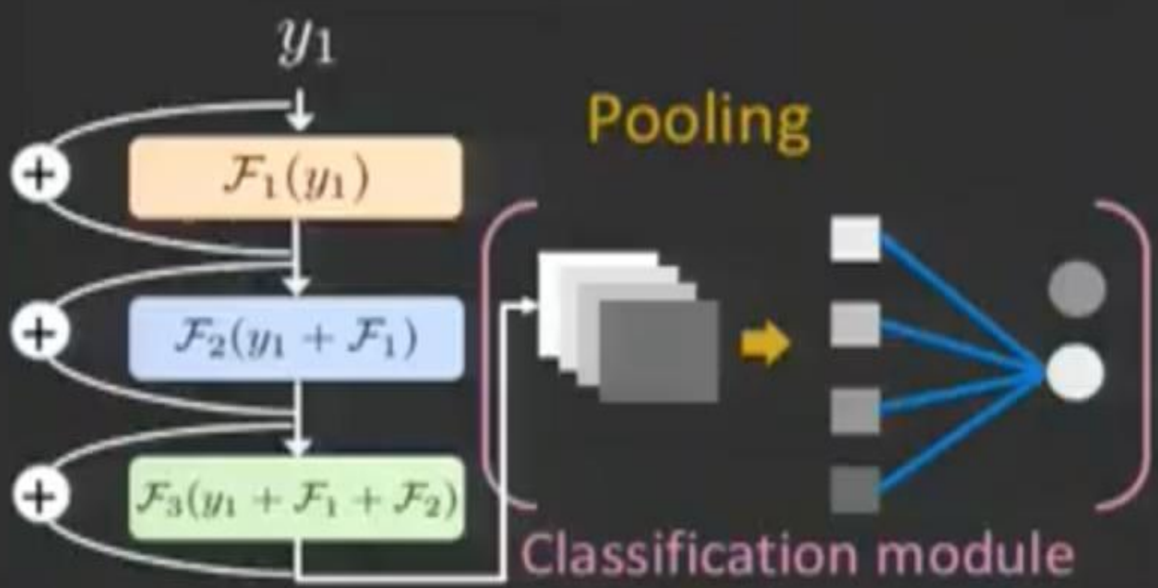
- Four ResNet blocks
- Replace the addition shortcut by concatenation shortcut
- 512-d image features are used for LSTM
- Convolutional features and the 512-d image features are used for attention model

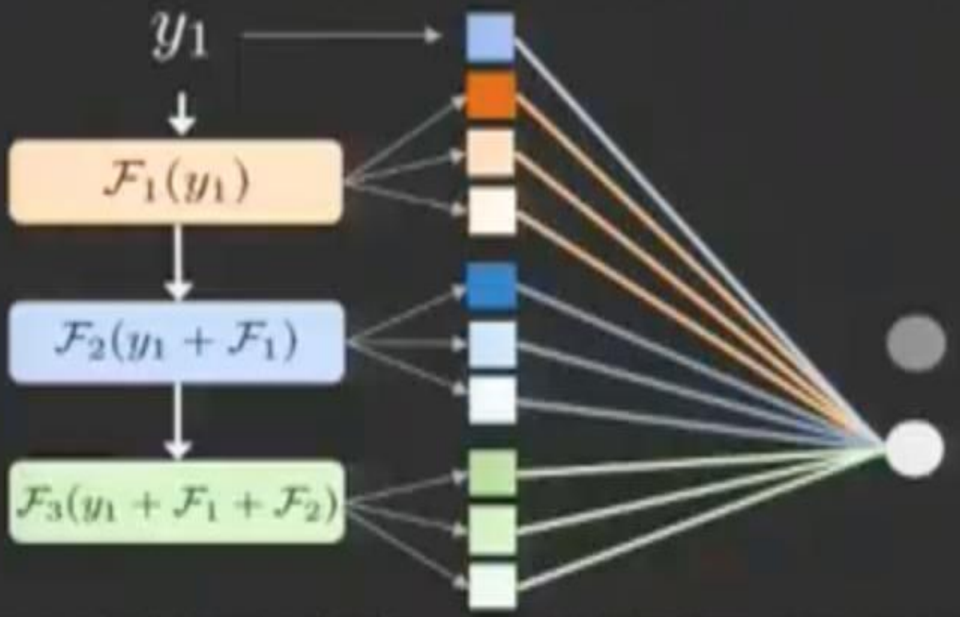
Standard approach: Shared weights

Our approach: Un-shared weights

Math Formulation:

$$\sum_{\text{spatial}} \left(y_1 + \sum_{i=1}^{3} \mathcal{F}_i \right) \boxed{w^c} \Rightarrow \sum_{\text{spatial}} \left(y_1 \boxed{w_1^c} + \sum_{i=1}^{3} \mathcal{F}_i \boxed{w_{i+1}^c} \right)$$

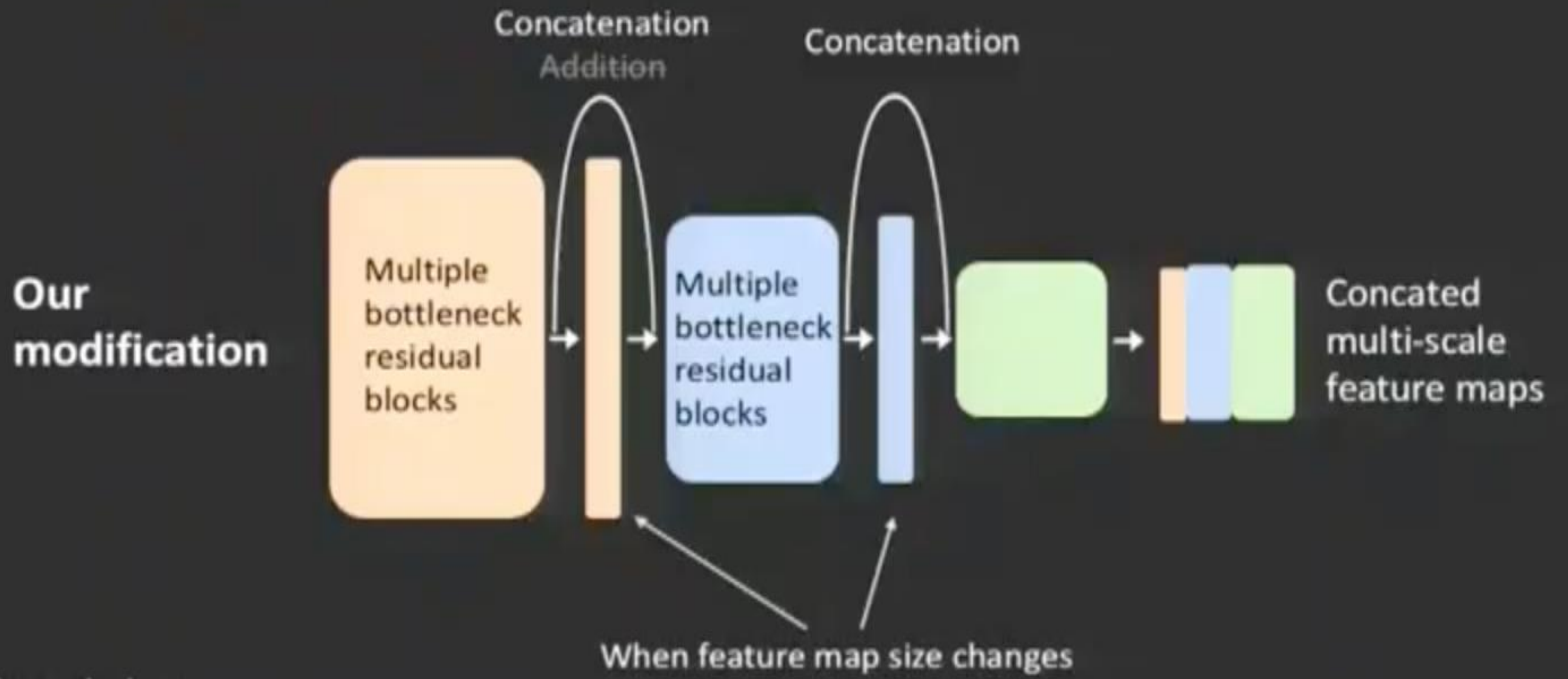




Problems: All layers sharing the same weights undermines the layer feature effects

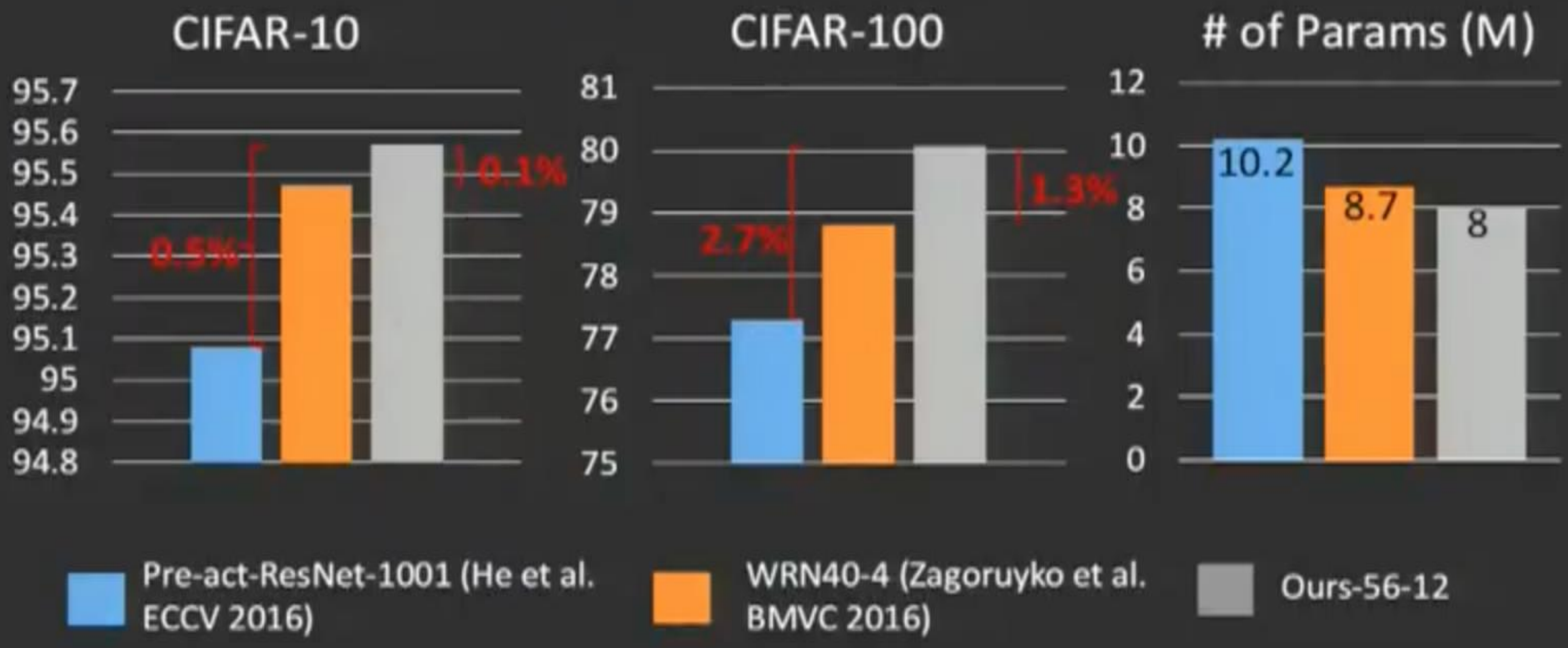
Solution: Using individual weights for different layers

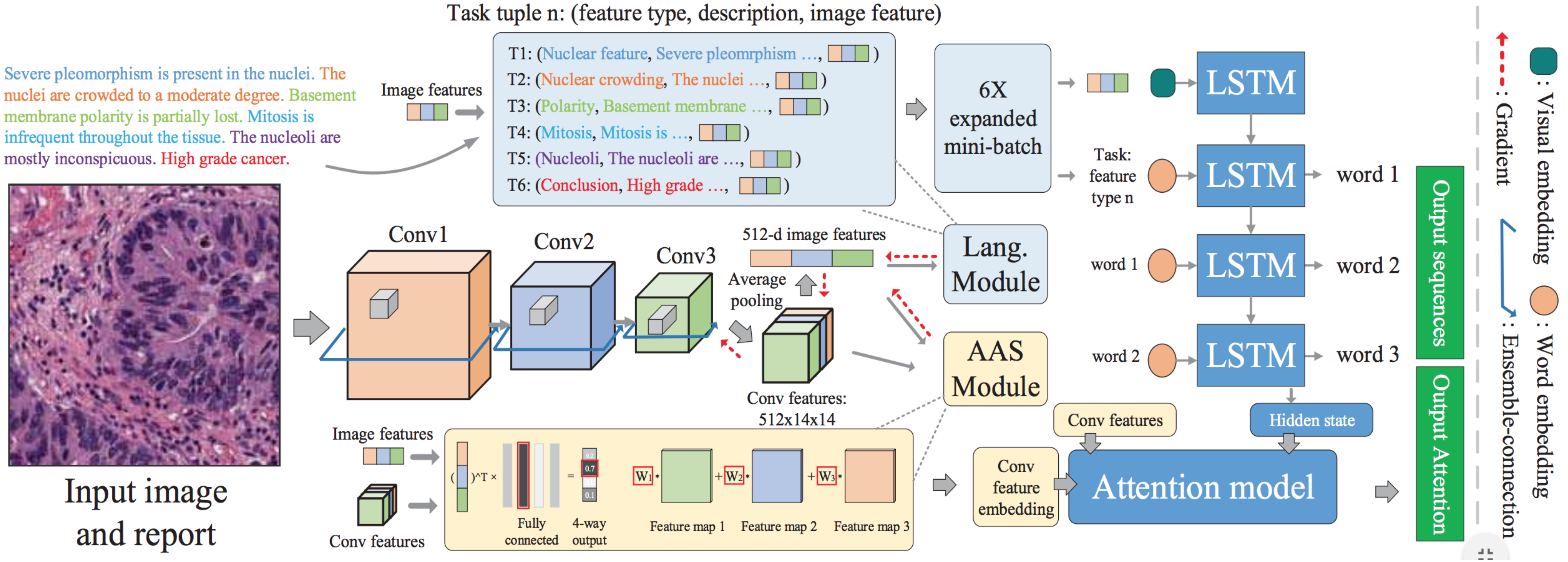
In-network Design



Acknowledge:
DenseNet (Huang&Liu et al, CVPR 2017) has similar modifications but a different motivation to ours.

Performance on CIFARs





- Image Module
- Language Module

Auxiliary Attention Sharpening (AAS) Module

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Kelvin Xu
Jimmy Lei Ba
Ryan Kiros
Kyunghyun Cho
Aaron Courville
Ruslan Salakhutdinov
Richard S. Zemel
Yoshua Bengio

Learning Deep Features for Discriminative Localization

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba
Computer Science and Artificial Intelligence Laboratory, MIT
{bzhou,khosla,agata,oliva,torralba}@csail.mit.edu

Abstract

In this work, we revisit the global average pooling layer proposed in [13], and shed light on how it explicitly enables the convolutional neural network (CNN) to have remarkable localization ability despite being trained on image-level labels. While this technique was previously proposed as a means for regularizing training, we find that it actually builds a generic localizable deep representation that exposes the implicit attention of CNNs on an image. Despite the apparent simplicity of global average pooling, we are able to achieve 37.1% top-5 error for object localization on ILSVRC 2014 without training on any bounding box annotation. We demonstrate in a variety of experiments that our network is able to localize the discriminative image regions despite just being trained for solving classification task[1].

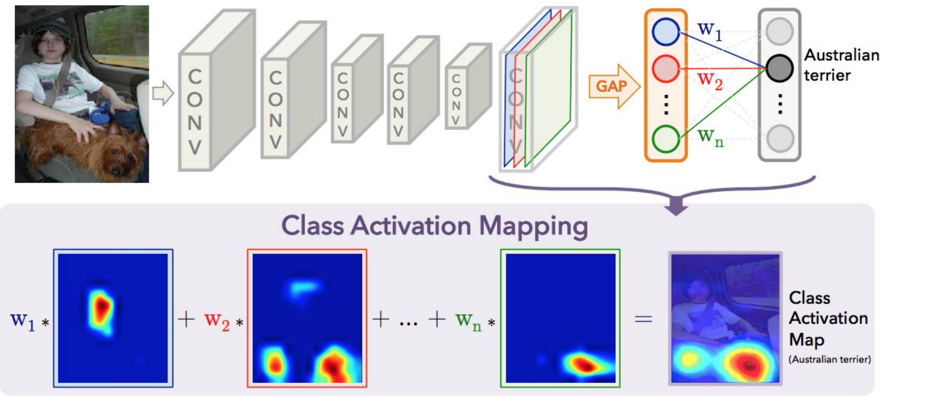


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

Auxiliary Attention Sharpening (AAS) Module

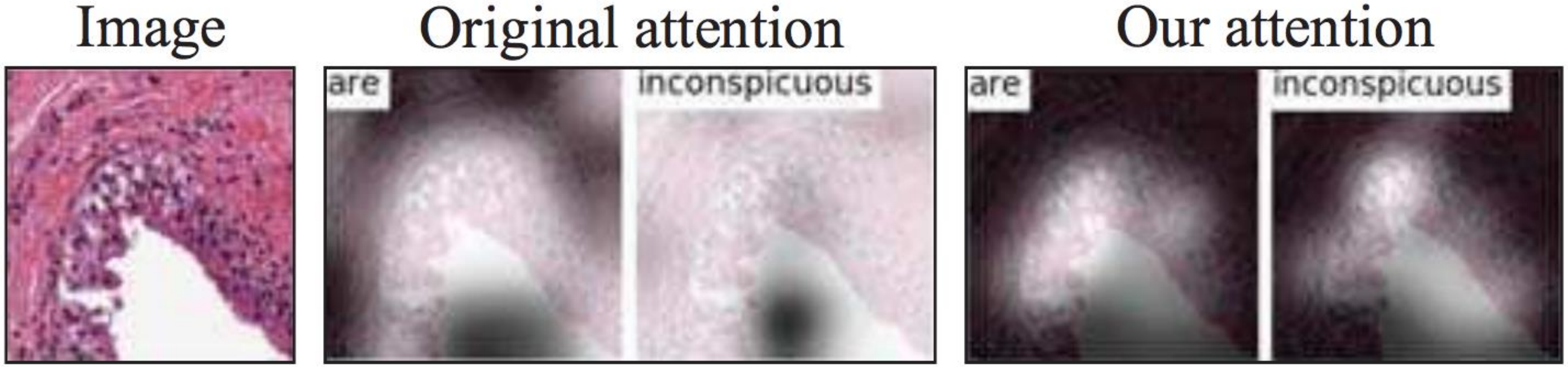
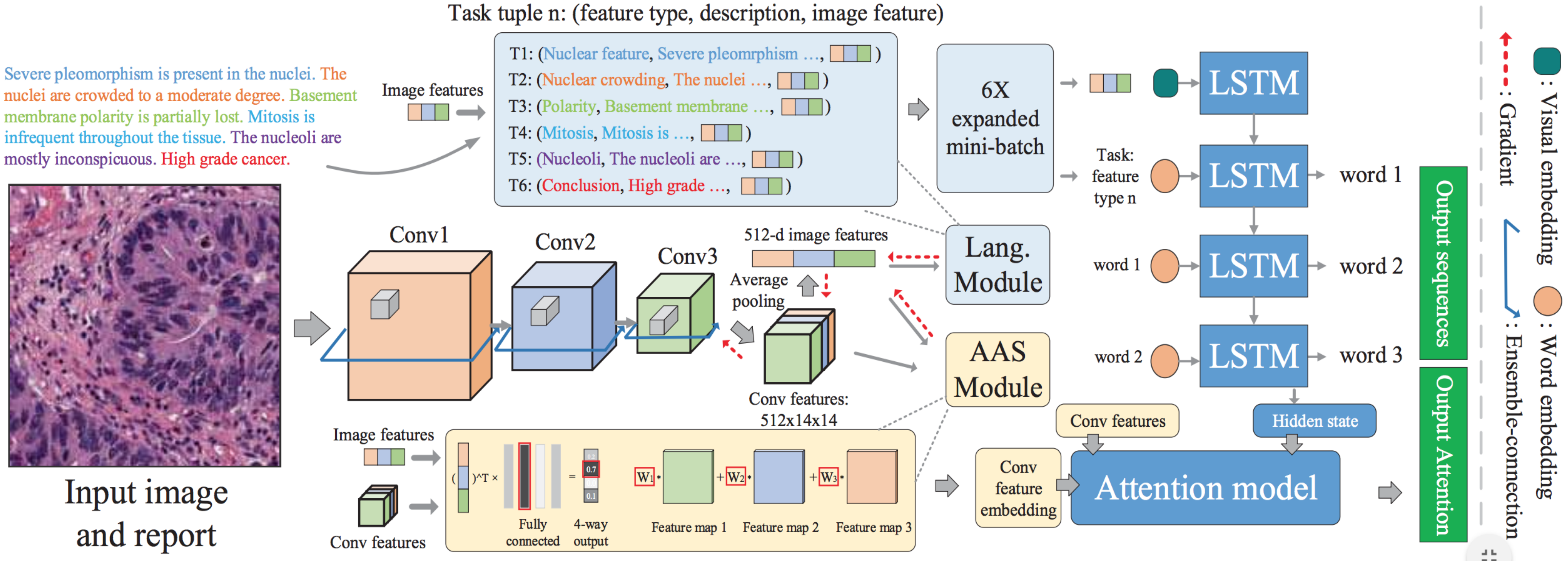
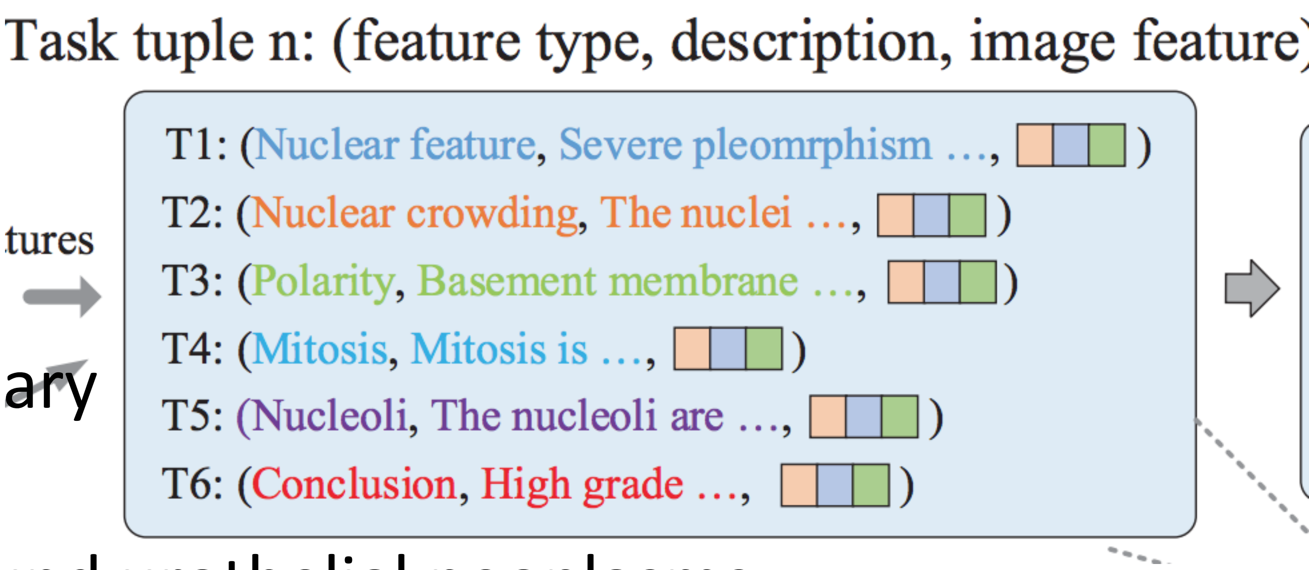


Figure 3: The attention maps of the original method (middle) and our method (right). Our method generates more focal attention on informative (urothelial) regions.



- Image Module
- Language Module

Bladder Images



- 32 patient slides at risk of a papillary
- 1000 images (500x500pixels) around urothelial neoplasms
- 5 reports are provided per images
  - 5 types of cell appearance features: state of nuclear pleomorphism, cell crowding, cell polarity, mitosis, prominence
  - Conclusion comes in 4 classes: normal, low-grade carcinoma, high-grade carcinoma, and insufficient data
- 5000 pairs of data in total (5-fold cross-validation)

Training

- The five descriptions and one conclusion are treated as K=6 separate tasks for LSTM training
- The conclusion is used as a four-way label for CNN training

The overall model has three sets of parameters: θ_D in the image model D, θ_L in the language model L, and θ_M in the AAS module M. The overall optimization problem in MDNet is defined as

$$\max_{\theta_L,\theta_D,\theta_M} \mathcal{L}_M(l_c, M(D(I;\theta_D);\theta_M)) + \mathcal{L}_L(l_s, L(D(I;\theta_D);\theta_L)), \quad (11)$$

where $\{I, l_c, l_s\}$ is a training tuple: input image I, label l_c and groundtruth report sentence l_s. Modules M and L are supervised by two negative log-likelihood losses $\mathcal{L}_M$ and $\mathcal{L}_L$, respectively.

$$\theta_D \leftarrow \theta_D - \lambda \cdot \Big((1-\beta) \cdot \frac{\partial \mathcal{L}_M}{\partial \theta_D} + \beta \cdot \eta \frac{\partial \mathcal{L}_L}{\partial \theta_D}\Big), \quad (12)$$

where λ is the learning rate, and β dynamically regulates two gradients during the training process. We also introduce another factor η to control the scale of $\frac{\partial \mathcal{L}_L}{\partial \theta_D}$, because $\frac{\partial \mathcal{L}_L}{\partial \theta_D}$ often has smaller magnitude than $\frac{\partial \mathcal{L}_M}{\partial \theta_D}$. We will analyze

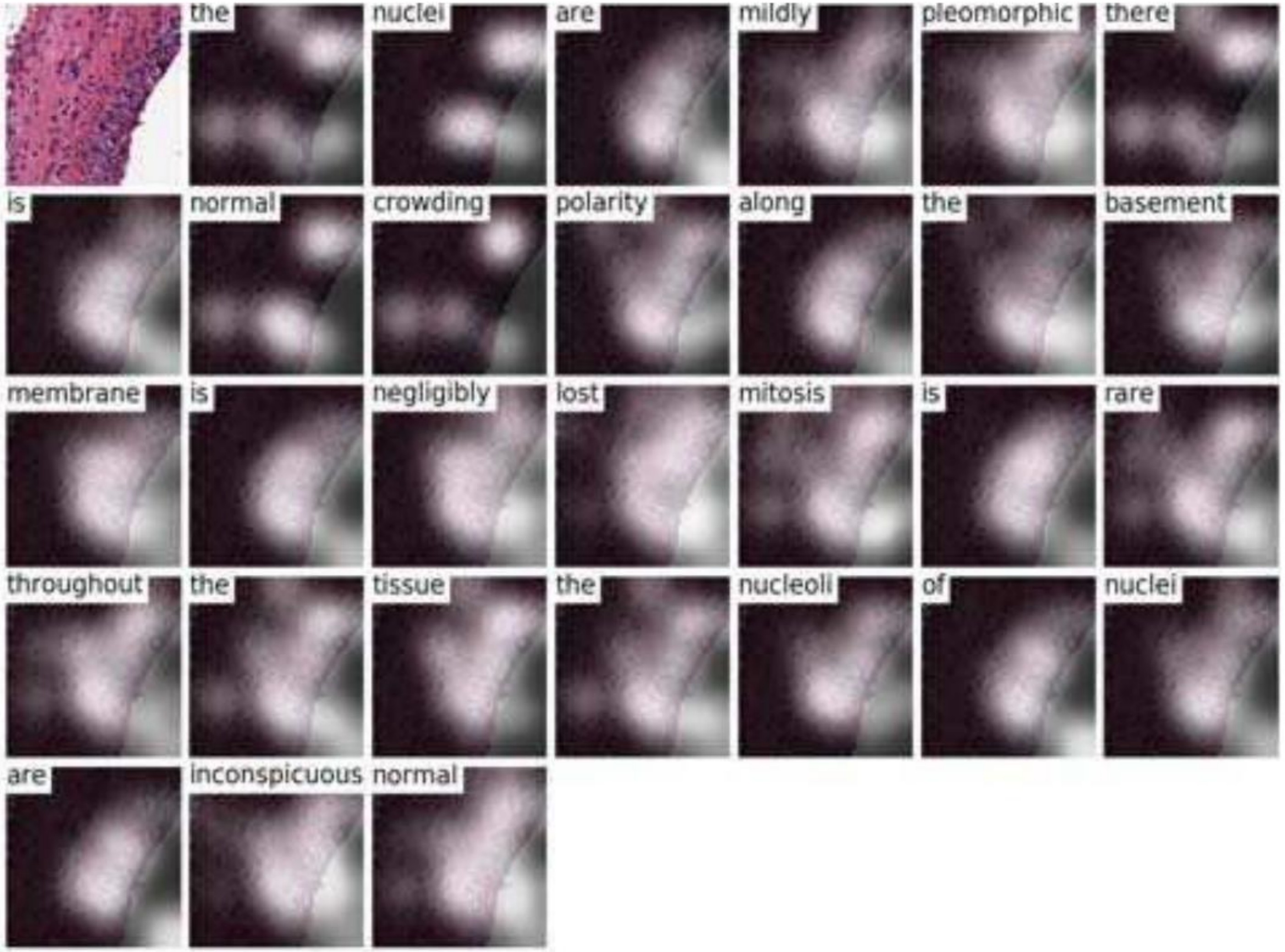


Figure 5: The image model predicts diagnostic reports (left-up corner) associated with sentence-guided attention maps. The language model attends to specific regions per predicted word. The attention is most sharp on urothelial neoplasms, which are used to diagnose the type of carcinoma.

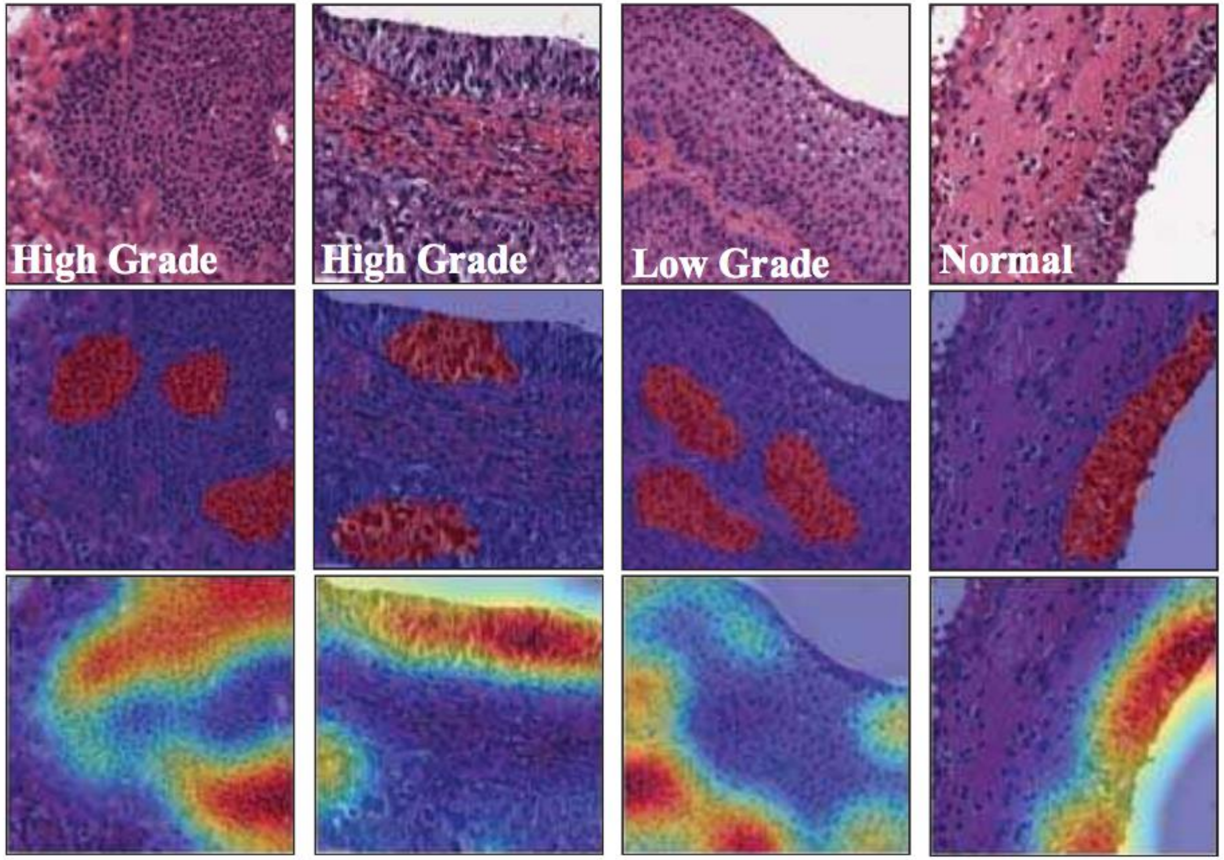


Figure 6: The illustration of class-specific attention. From top to bottom, test images, pathologist annotations, and class attention maps. Like the pathologist annotations, the attention maps are most activated in urothelial regions, largely ignoring stromal or background regions. Best viewed in color.

Quantitate Evaluation

	CNN		BLEU-4	METEOR	DCA(%)±std
	Pre-trained	Fine-tuning			
Baseline (GoogleNet)	✔	✔	66.9	39.5	74.2±3.8
Ours	Train end-to-end from scratch		**67.7**	**39.6**	**78.4±1.5**

- Baseline: NeuralTalk2 (Karpathy et al, CVPR, 2015)
- **DCA** - Diagnostic conclusion accuracy

- https://youtu.be/yy7NUrc3KI0
- https://www.youtube.com/watch?v=DiNUcYi3Oxs