2nd ADNE assignement
Task 2

João Camacho, N. 56861

I. Summary of the models

A. Road side Parking Net

Based on [Xu+18], this model was made for the task of licence plate recognition. Here the model is composed of two modules:

- **Detection Module**
  Mainly a CNN composed of 5 pairs of convolutional layers, each followed by a max pooling layer, these convolutional layers feed into a "box predicter", this box predictor is can be seen as a secondary output that tries to predict the bounding box of the license plate, this was a pre-trained task. The predicted bounding box is used to extract the features from the feature maps obtained in the covolutional layers that correspond to the license plate section of the image.
- **Recognition Module**
  This module uses the location predicted by the box predicter and the convolution features to do Region of interest pooling, this pooling extracts only the features from the license plate from the feature maps that result from the first 3 pairs of convolution layers. Using this box regression to extract only the region of interest from the convolution maps can be seen as the attention mechanism of this model, it discards features from outside the bounding box and considers only the ones inside it that correspond to the license plate.
  Now using only relevant features the model can easily classify the license plate numbers, the model does this by using a license plate number classifier that is implied to be a group of fully connected dense layers.

At the end the model performs relatively well by scoring a average precision of 95.5% at 61 frames per second, compared with other models for license plate recognition the RPnet scores the highest average precision at the highest frames per second.

B. Attention-based Extraction of Structured Information from Street View Imagery

This model based on [Woj+17] was developed for the task of image text recognition and was tested on the challenging French Street Name Signs (FSNS) and a more challenging dataset composed of images from store fronts with a business name, this more challenging dataset was composed of 1M images extracted from google street view images.

Model description

The model consists of a CNN for feature extraction of the images, then the features are used to calculate a weighted sum, that uses the spatial attention mask for the weights. Then we use a RNN to transform a the image features into text, the input of the RNN at time t is defined as a sum of the index of the previous letter (ground truth during training, predicted during test time) multiplied by a weight Wc and the weighted sum of the attention mask and the CNN image features multiplied again by a weight Wu.

Using this input and the previous hidden state of the RNN we pass is through a LSTM layer to compute the output of the layer $\hat{o}_t$ and the next hidden state s_t, this new output is an embedding so it has to be passed trough a softmax activation to be comparable with the ground-truth, but the softmax takes in the output of the LSTM layer and the weighted sum of the attention mask and the CNN image features, each multiplied with new weights. Finally with the softmax activation output we can compute the most likely letter with the highest probability.

The spatial attention is computed with the $tanh$ of the weighted sum of the hidden state and the CNN features, this is then passed by a softmax activation to compute the attention weights. The weighted hidden state serves as a time-varying offset for the weighted features of the image, to make the model location aware we add a one hot encoding of the selected coordinate for the CNN features, this combination produces the attention weight vector that is temporal and spatially aware and is the baseline attention mechanism of the model. The spatial awareness of the model was added later after seeing it was needed for good results to be achieved.

Using this model with 3 kinds of feature extractors (CNNs) called inception-v2, inception-v3 and inception-resnet-v2, the best results was achieved by the inception-resnet-v2 with Size per view 7x7x1088 and dept 51, they achieved 83.3% accuracy on the FSNS using this feature extractor.

Only qualitative results where shown in the google street view dataset, they compute the saliency of a pixel as the partial derivative of the logit of the attention weights vector with respect to the input image, They show how the saliency maps and the attention masks focus on the regions of the image.

C. Attention-Based Models for Speech Recognition

Based on [Cho+15], this model was created with the task of speech recognition, the model was tested on TIMIT phoneme recognition task and achieved a 18.7 phoneme error (PER), since some utterances where too short some of them had to be extended by repeating past utterances. This model was not

created for the task of image recognition but it's attention mechanism could be adapted to work with image recognition.

Model description

The model consists of an encoder for feature extraction that extracts a representation more suitable for the attention mechanism to work with, the encoder in this case is a deep bidirectional recurrent network (BiRNN), followed by the encoder we apply the attention mechanism called attention-based recurrent sequence generator (ARSG) with convolutional features, this attention mechanism is an improvement on a previously studied mechanism where we scored each element of the encoder output h_j with the $tanh$ of the encoder output with the previous hidden state and a bias, this previous model had the limitation that identical or very similar elements of h_j are scored equally regardless of their position in the sequence (in other words the model has limited contextual disambiguation due to limited capacity of the elements h_j to transfer contextual information), this mechanism is called content-based. To solve this problem a location-based and content-based hybrid attention mechanism is made, this new solution uses the previous alignment to compute the new score, making the model location-based, it does this by convoluting the previous alignment with a trainable matrix F and adding the result of this convolution to the previous input of the $tanh$ with a multiplied weight matrix.

Then the score vector needs to be normalized because if the input sequence h_j that is the output of the encoder is long then it is likely the glimpse witll have noisy information form irrelevant features, also due to long h_j the computations may become prohibitively expensive, to solve this Sharpening and Smoothing is used. In Sharpening many methods can be used like adding inverse temperature to the softmax activation of the scores to calculate the attention weights vector, another way is to use windowing and this ends up being the method that produces better results and is theoretically more computationally efficient. Sharpening overall made the long utterances more simple to compute and provided better results for them.

Smoothing was used because after applying the Sharpening for the long utterances the model lost accuracy at predicting the short utterances from the training, this leads to the hypothesis that the model preforms best if he considers a collection of top scored utterances because only considering the top scored one, removes diversity from the model. To solve this a tweak was made to the activation function used on the scores to output the attention weights, instead of the unbounded exponential function in the softmax function a bounded sigmoid function was used, this had the affect of smoothing the output probabilities and allowed the model to focus on more than one top scored utterance. At the end the model scored 17.6% Phoneme error rate (PER) using smoothing and the convolution features of the previous alignment α_{j-1}

D. Recurrent Attention Model (RAM)

Based on [Mni+14], this model was made for the task of image classification but was adapted to play a simple game so it's limits could be tested. On the MNIST dataset the model manages a average error of 1.07% and on the Translated MNIST a average error of 1.2%. In cluttered enviroments like the 60x60 Cluttered Translated MNIST the model has 4.04% error and the 100x100 Cluttered Translated MNIST the model has 8.11%. This model works better than Fully connected dense Networks and CNNs, because of it's attention mechanism that allows it to focus on the important aspects of the image and ignore clutter.

Model description

This model is composed of a glimpse sensor that extracts representations of the image via the retina encoding that extracts k square patches centered at location l, with the first patch being $g_w g_w$ pixels in size, and each successive patch having twice the width of the previous. It is also important to note that patched away from the selected location l are processed with lower resolution allowing for lower dimensionality than the actual image, this is the main focus of the attention, since higher resolution zones have more pixels this zones have more parameters and therefore more focus. The glimpse sensor is incorporated in a glimpse network that takes the location l to focus on and gives it to the glimpse sensor, then trough a fully connected dense layers with Relu activations, the glimpse and the location l are encoded and added together so they can be passed trough another fully connected dense layers with Relu activation to be encoded together, these dense layers a trainable and named $\{\theta_g^0, \theta_g^1, \theta_g^2\}$, the glimpse network outputs the new glimpse.

This glimpse is used in the core network of the model that is an RNN, the glimpse is inputed to the RNN layer and used to compute the new internal state h_t, this new state h_t has information on the history of past observations; it encodes the agent's knowledge of the environment and is instrumental to deciding how to act and where to deploy the sensor. Based on the internal state many actions can be preformed, in the case of image processing there are two actions used, the decision of the next location for the glimpse network to focus on and the environment action α_t that is used for classification. Each of the actions are passed trough a trainable fully connected dense layer but the training rule changes for each one, for the environment action the training rule is the standard optimization of the cross entropy loss by back-propagation of the gradients, for the location the reinforcement rule is used that is a rule that scores the output of the layers based on a cumulative reward given based if the object is classified correctly or not, we train the linear network by reducing the squared error between R_t i 's and b_t, b_t is a baseline.

II. Comparisons between attention mechanisms

A. Model 1 and 2

Advantages between model 1 and 2 Model 1 can be seen as a more sharp attention mechanism because it does not consider weights for it's attention vector it instead bisects the convolution features of the defined Region of interest, this can provide stronger results in the task of license plate recognition

since it completely disregards any background clutter (like a random number appearing in the background outside the license plate).

Disadvantages between model 1 and 2

The first Disadvantage is that model 1 requires positional labels, it can not learn the position to focus it's attention without positional labels and requires pre-training of the regression box with those labels.

Model 1, on contrary of model 2, does not include information from features outside it's focus, this makes the training of the bounding box even more important because the model depends entirely on the correct positioning of the box.

Model 2 also adds more diversity to the features it includes outside it's focus and that might be useful in the RNN.

Model 2 also has more vocabulary capabilities than the model 1 that was trained to predict only license plate numbers, the fact model 2 was mode for the task of reading street signs it can then output more characters than model 1.

Model 1 can also be seen as a time invariant model meaning the model does not consider the previous outputted character, while model 2 considers this.

B. Model 1 and 3

Advantages between model 1 and 3

Again the main advantage of model 1 is it's sharp focus on the bounding box that might be useful in certain applications, but since model 3 uses sharpening methods to select the best utterances it could become competitive in the same tasks that model 1 preforms well.

Disadvantages between model 1 and 3

Many of the disadvantages of model 1 here, are the same as comparing with model 2 in:II-A.

Model 3 considers top scored utterances adding to the learning capability that model 1 lacks.

C. Model 1 and 4

Advantages between model 1 and 4

Again the main advantage of model 1 is it's sharp focus on the bounding box that might be useful in certain applications, but since model 4 uses an adaptive sensor method to select the best area of focus it could become competitive in the same tasks that model 1 preforms well.

Model 4 can also adjust the size of it's focus but need's to be adapted for that action.

Disadvantages between model 1 and 4

Many of the disadvantages of model 1 here, are the same as comparing with model 2 in:II-A.

Model 4 considers top scored utterances adding to the learning capability that model 1 lacks.

D. Model 2 and 3

Advantages between model 2 and 3

Model 2 is capable of generalizing really well, since it was tested and proven to work on google street images that have a lot of background noise and distortion, while model 3 had problems processing utterances that differ from the ones used in training (in size) and had to be adapted.

Model 2 and 3 both use a location-aware attention mechanism but in model 2 they had to adapt this location aware mechanism to work with multi-line text, model 3 did not received this adaptation and used the one that did not work in model 2 and might not work so well in it too due to the similarities of the models.

Disadvantages between model 2 and 3

Model 3 has many methods of Sharpening and Smoothing that can be used, the best performing one is windowing that can not only include many top scored features (that add diversity to the possible outputs of the model) but can also make computations faster since we are just computing the score for the evaluated features in the window.

Overall model 3 has more capability than model 2 of using past outputs to achieve better results.

E. Model 2 and 4

Advantages between model 2 and 4

Model 2 is overall a simpler model to implement, due to only using CNN to extract the features from the images and not a glimpse network like in model 4.

Disadvantages between model 2 and 4

Model 4 has a more complex method of extracting features from the images through it's glimpse network, this method is proven to outperform a CNN and in model 2 a CNN is used for feature extraction.

Model 4 uses a more advanced way of treating attention that limits the amount of features used, in contrast model 2 uses a weight vector to determine the more relevant features out of all of them, making model 2 relatively slower due to considering all features.

Model 4 has various trainable fully connected layers that might make the model relatively more powerfull than model 3.

F. Model 3 and 4

Advantages between model 3 and 4

Model 3 uses sharpening and smoothing methods that allow it to consider k top scored outputs and consider that before outputing a response(making it have more diversity in the considered responses).

Disadvantages between model 3 and 4 Model 3 doesn't use a method to limit the extracted features of images, like model 4 does, so model 4 might be computationally faster due to having fewer features extracted from the image.

Model 4 has various trainable fully connected layers that might make the model relatively more powerfull than model 3.

Model 4 can implement actions that allow it to preform secondary tasks likes adjusting the size of the sensor or playing a game, this kind of versatility doesn't exist in model 3.

References

[Mni+14] Volodymyr Mnih et al. "Recurrent Models of Visual Attention". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., 2014, pp. 2204–2212. URL: http://papers.nips.cc/paper/5542-recurrent-models-of-visual-attention.pdf (visited on 05/23/2020).

[Cho+15] Jan K Chorowski et al. "Attention-Based Models for Speech Recognition". In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes et al. Curran Associates, Inc., 2015, pp. 577–585. URL: http://papers.nips.cc/paper/5847-attention-based-models-for-speech-recognition.pdf (visited on 05/23/2020).

[Woj+17] Zbigniew Wojna et al. "Attention-Based Extraction of Structured Information from Street View Imagery". en. In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Kyoto: IEEE, Nov. 2017, pp. 844–850. ISBN: 978-1-5386-3586-5. DOI: 10.1109/ICDAR.2017.143. URL: http://ieeexplore.ieee.org/document/8270074/ (visited on 05/23/2020).

[Xu+18] Zhenbo Xu et al. "Towards End-to-End License Plate Detection and Recognition: A Large Dataset and Baseline". en. In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari et al. Vol. 11217. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 261–277. ISBN: 978-3-030-01260-1 978-3-030-01261-8. DOI: 10.1007/978-3-030-01261-8_16. URL: http://link.springer.com/10.1007/978-3-030-01261-8_16 (visited on 05/18/2020).