On the Learning Dynamics of Attention Networks

Rahul Vashisht[1,*] **and Harish G. Ramaswamy**[1]

[1]Department of CSE, Indian Institute of Technology Madras

Abstract. Attention models are typically learned by optimizing one of three standard loss functions that are variously called – soft attention, hard attention, and latent variable marginal likelihood (LVML) attention. All three paradigms are motivated by the same goal of finding two models– a 'focus' model that 'selects' the right *segment* of the input and a 'classification' model that processes the selected segment into the target label. However, they differ significantly in the way the selected segments are aggregated, resulting in distinct dynamics and final results. We observe a unique signature of models learned using these paradigms and explain this as a consequence of the evolution of the classification model under gradient descent when the focus model is fixed. We also analyze these paradigms in a simple setting and derive closed-form expressions for the parameter trajectory under gradient flow. With the soft attention loss, the focus model improves quickly at initialization and splutters later on. On the other hand, hard attention loss behaves in the opposite fashion. Based on our observations, we propose a simple hybrid approach that combines the advantages of the different loss functions and demonstrates it on a collection of semi-synthetic and real-world datasets.

1 Introduction

Attention models have emerged as one of the most successful architectures in deep learning [1, 6, 16, 20]. These models offer a natural way to interpret the intermediate outputs by introducing an *attention vector*, which identifies the relevant part of the input responsible for the output. Along with improving the model's performance, attention models also provide a transparent mechanism to study intermediate outputs in neural networks . As a result, attention models have become an essential tool for explainability in downstream tasks, thereby emphasizing the need for a comprehensive understanding of their working mechanisms.

Recent advances in the field have focused on softly simulating alignments, also known as *soft attention*, where the model uses a convex combination of features based on attention weights calculated using a deterministic function. Soft attention allows the model to focus on multiple segments of the input for downstream tasks, providing better accuracy than traditional neural networks. In contrast, *hard attention* based approaches select one of the input segments based on the attention weight distribution [3, 13, 20]. Typically, a hard attention model is trained either by directly maximizing the log-marginal-likelihood or maximizing a lower bound objective obtained using Jensen's inequality. Both these methods are computationally expensive compared to the soft attention model.

Contributions: In this paper, we give insights into the learning dynamics of attention mechanisms under soft and hard attention

paradigms. We study the dynamics of attention mechanisms under a "fixed focus setting" and identify a distinct property of soft and hard attention that makes them act very differently at different points in the training. This property makes the final learned model through soft attention less interpretable and makes the training with hard attention particularly slow at initialization. We propose a hybrid approach that addresses some of these limitations. We also derive closed-form expressions for the parameter trajectory under gradient flow in a simple setting. Our work sheds light on the failure modes of different attention paradigms and enables the design of new algorithms with desirable properties.

Related Work: There has been a lot of research in the field of deep learning to understand the behavior of attention models. Translation tasks between sentence pairs have used hard attention [2, 3], while soft attention has been used in other recent applications [4, 7]. Several studies have investigated whether attention mechanisms can offer meaningful insights using empirical methods [8, 17, 15]. Even with these studies, a thorough understanding of attention models is still lacking in most research. Some research has focused on understanding attention weights in classification tasks and natural language settings such as visual question answering. In addition, other studies have discovered a mathematical relationship between attention scores and word embedding norms in topic classification [11]. There have been works that show latent variable attention performs better than soft attention. These works also propose variational inference based attention models [5]. Similar to this, there have been works focused on performing exact hard attention for the monotonic and non-monotonic sequence to sequence character-level transduction tasks [18, 19]. In this paper, we study the learning dynamics of various attention paradigms and explore how they produce different results.

2 Losses and inference methods for attention

In this section, we briefly summarise the three known paradigms/loss functions for attention in the context of a simple problem that we term selective dependence classification [14] (SDC). We also analyze the performance of the three paradigms on a semi-synthetic dataset based on CIFAR10 [9] and identify some characteristic signatures of these paradigms.

2.1 Attention and Latent variable alignment (LVA)

In the latent variable model for attention, we consider an instance $\mathbf{X}$ that is a mosaic object consisting of multiple segments(or parts or patches), each of which is represented by a fixed dimensional vector. The label variable $y \in \mathcal{Y}$ is generated based on the mosaic instance

---

* Corresponding Author, email: rahul@cse.iitm.ac.in

X and a latent variable **z**. The latent variable **z** (often called 'alignment' [5] in literature) indicates which segment (or segments) of **X** generates y. Directly maximizing the $\log p(y|\mathbf{X}; \theta)$ (log marginal likelihood) is complicated in general, and becomes even more intractable in cases where the alignment variable **z** can take a large number of values [18, 19].

2.2 Selective Dependence Classification

For the sake of studying the three standard paradigms of attention, we consider the following concrete version of the latent variable alignment (LVA) problem mentioned above – we call this a selective dependence classification(SDC) problem. Here the instance $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m] \in \mathbb{R}^{d \times m}$, contains m parts or segments each of which is represented by a vector in $\mathbb{R}^d$. The label Y takes values in $\mathcal{Y} = \{1, 2, \ldots, C\}$. The hidden alignment variable Z takes values in $\{1, 2, \ldots, m\}$, indicating that only one of the segments of **X** is responsible for the generation of y. The instance **X** is called a mosaic instance, and the segment identified by Z is called a 'foreground' segment, while the rest are called 'background' segments. The latent variable alignment probabilistic model is given as follows:

$$P(Y = k|\mathbf{X}, Z = i) = \sigma_k\left([\mathbf{g}_1^*(\mathbf{x}_i), \ldots, \mathbf{g}_C^*(\mathbf{x}_i)]\right)$$
$$P(Z = i|\mathbf{X}) = \sigma_i\left([f^*(\mathbf{x}_1), \ldots, f^*(\mathbf{x}_m)]\right)$$

where $f^* : \mathbb{R}^d \to \mathbb{R}$ and $\mathbf{g}^* : \mathbb{R}^d \to \mathbb{R}^C$ are the parameters of the data model defined above and σ is the softmax operator that transforms an arbitrary vector into probability vector of the same dimension, and σ_j represents the jth co-ordinate.

The training data for the task is the collection of pairs **X**, Y. The final goal in the SDC problem is simply to learn a model that predicts the label Y correctly from a mosaic instance **X**. This toy problem is analogous to an image classification problem where each image is labeled only based on an object occupying only a small (and unknown) portion of the image. The LVA model is a discriminative model and only gives $P(Z, Y|\mathbf{X})$. In our synthetic and theoretical arguments, we consider a corresponding full generative model by giving $P(Y, Z)$ (which is simply uniform over $[C] \times [m]$) and $P(\mathbf{X}|Y, Z)$. Conditioned on Z, Y, the Zth segment of **X**, denoted by $\mathbf{x}_Z$ is distributed as D_Y, a distribution over $\mathbb{R}^d$ which we call the foreground distribution for class Y. $\mathbf{x}_j$ for all $j \neq Z$ are drawn independently from a background distribution D_0. An illustration of SDC is given in the appendix[1]

2.3 Paradigms of Attention

Consider the following template attention model for the SDC task that we call a Focus-Classify Attention Model (FCAM). The FCAM is parameterised by a focus function $f : \mathbb{R}^d \to \mathbb{R}$ that scores each segment, and a classification function $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^C$ that classifies an 'aggregated segment'. The different modes of aggregation and penalties for misclassification give rise to the three loss functions used commonly in attention – soft attention (SA), latent variable marginal likelihood (LVML), and hard attention (HA).

$$L^{\text{SA}}(f, \mathbf{g}, \mathbf{X}, y) = -\log(\sigma_y(\mathbf{g}(\sum_j a_j(\mathbf{X})\mathbf{x}_j)))$$
$$L^{\text{LV}}(f, \mathbf{g}, \mathbf{X}, y) = -\log(\sum_j a_j(\mathbf{X})\sigma_y(\mathbf{g}(\mathbf{x}_j)))$$
$$L^{\text{HA}}(f, \mathbf{g}, \mathbf{X}, y) = -\sum_j a_j(\mathbf{X})\log(\sigma_y(\mathbf{g}(\mathbf{x}_j)))$$

where $\mathbb{R}^m \ni \mathbf{a}(\mathbf{X}) = \sigma(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_m))$ is the normalised score given to the m segments in the mosaic input **X** by the focus model f. It can easily be seen that L^{LV} corresponds exactly to the negative log-likelihood of the SDC model, and hence the minimizer of L^{LV} is the maximum-likelihood estimator for the SDC problem. The soft (hard) attention loss is usually motivated as an approximation (Jensen's inequality upper bound relaxation) of L^{LV} [5]. Also, it is important to note that in the case when the attention vector $\mathbf{a}(\mathbf{X})$ has only one non-zero entry, all three losses become equal and that is indeed a desirable property when training such models. However, this does not happen in practice as the initial focus network f is usually such that the entries of **a** are all approximately $1/m$.

In the training phase, the focus and classification models $f, \mathbf{g}$ are learnt by optimising one of the loss functions above, and the learnt model is used for making a prediction on a new mosaic instance **X** using the three corresponding inference procedures as (i) $s_k^{\text{SA}} = \sigma_k(\mathbf{g}(\sum_j a_j \mathbf{x}_j))$, (ii) $s_k^{\text{LV}} = \sum_j a_j \sigma_k(\mathbf{g}(\mathbf{x}_j))$, and (iii) $s_k^{\text{HA}} = \sigma_k(\mathbf{g}(\mathbf{x}_{j^*}))$ where, $j^* = \arg\max_j f(\mathbf{x}_j)$ and $\mathbf{a} = [\sigma(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_m))]$.

The vectors s^{SA}, s^{LV} and s^{HA} are the scores for the C classes given by the aggregation methods. The final class prediction is done by simply returning the maximising co-ordinate of the score vector. Given a focus model f and classification model $\mathbf{g}$ any of the above 3 inference procedures can be used on a test point for making a prediction, but the default choice is to use the inference procedure corresponding to the loss function that the model $f, \mathbf{g}$ minimises.

Despite the latent variable marginal likelihood loss having pride of place as the classic ML estimator, it is typically less preferred over the soft attention and hard attention paradigms. The soft attention loss has the advantage of being efficient in the number of calls to the computationally expensive **g** function, as the segment aggregation happens *before* passing it to **g**. This is especially advantageous in situations where the alignment variable Z takes an exponentially large number of values (a typical alignment example that generates a 5 word caption for an image with 196 segments has 196^5 values). While the hard attention loss L^{HA} has the same issue as L^{LV} it can be efficiently approximated as it can be expressed as an expectation. While the computational issues with the three paradigms are well known, the differences in learning dynamics and the final model learnt when using the different loss functions are not as well studied.

2.4 Comparative Empirical Analysis of the Attention Paradigms

In this section, we perform an empirical study on a synthetic SDC task based on the CIFAR10 dataset and identify some key characteristics of the final model learnt in all three paradigms. The dataset for the SDC task was generated as follows. The label y for each mosaic instance takes one of three possible values (car, plane, bird). The $m - 1$ background segments of the corresponding mosaic instance **X** are drawn randomly from images of the other 7 classes,

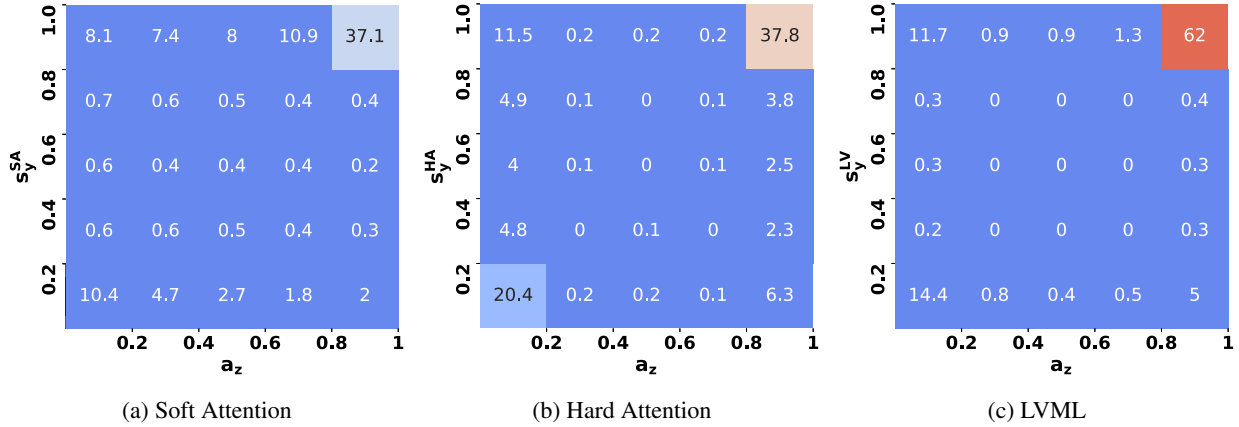(a) Soft Attention   (b) Hard Attention   (c) LVML

Figure 1: Focus-Prediction heat maps of test data for the three attention paradigms on the CIFAR10 SDC task with $n = 10000$ and $m = 5$.

the foreground segment is drawn from images corresponding to the foreground class given by the label y. Note that the position (or index) of the foreground segment can be arbitrary. We sample several such pairs (mosaic instances, labels), train an FCAM on a subset of this dataset, and evaluate it on the rest. The architecture of the focus model f and classification model $\mathbf{g}$ are both convolutional neural networks with three convolutional and four fully connected layers.

The results of the experiment (github link) with $m = 5$ segments and $n = 10000$ training points are given in Figure 1. (More results in other settings and also with CIFAR100 using large number of classes are in the appendix). The Focus-Prediction heat map of any given FCAM $f, \mathbf{g}$ simply gives the joint histogram of the normalised focus score of the true foreground a_z, and the model score of the true class s_y^{AP} where AP represents the attention paradigm used and takes values in {SA, LV, HA}. Here z corresponds to the true value of the hidden alignment variable Z and y corresponds to the true label Y. Note that computing this heat map requires access to the hidden alignment variable Z even though the models $f, \mathbf{g}$ have been trained without access to it.

A perfect model would have all instances in the top right corner of the heat map (we define this as the Strongly Accurate Interpretable Fraction, SAIF). Instances in the top left correspond to mosaic instances being classified correctly despite the focus model f giving a low score to the foreground segment. Instances in the bottom right correspond to mosaic instances where the focus model f scores the foreground patch much higher than the background patches, but the final score for the true class is low.

The Focus-Prediction heat maps in Figure 1 reveal some interesting patterns (that are also present in other settings, see supplementary material).

- The models trained with soft attention gives confident class label predictions (observe that row sums of the middle 3 rows in Figure 1 (a) are close to zero)

- The focus model in the FCAM trained with soft attention loss is often not very confident (observe the column sums of the middle 3 columns are non-negligible in Figure 1(a))

- The models trained with hard attention do not give confident class label predictions (observe that row sums of the middle 3 rows in Figure 1(b) are non-negligible)

- The focus model in the FCAM trained with hard attention loss is

often confidently wrong or right (observe the column sums of the middle 3 columns are close to zero in Figure 1(b))

- The FCAM models trained with latent variable marginal likelihood have both confident class label predictions and focus scores (observe that the row (columns) sums for the middle rows (columns) are close to zero in Figure 1(c))

- The top-right number in the heat map – corresponding to the fraction of instances classified correctly after being focussed correctly – in the LVML paradigm is often noticeably higher than the SA and HA paradigms.

The LVML paradigm is clearly superior to the other paradigms, however, computational complexity issues stand in the way of choosing it. The rest of the paper comes up with a simple explanation for why the three paradigms behave the way they do, and tries to come up with an approach that performs as well as the LVML paradigm, but without its computational issues.

3 Learning Dynamics of Attention Models

The loss functions for attention are typically simultaneously minimized over f and $\mathbf{g}$. However, it is instructive to analyze the dynamics of the evolution of $\mathbf{g}$ for certain fixed focus models f. We consider focus models that give a score of $\alpha \in [0, 1]$ to the foreground patch (i.e. $a_{j^*} = \alpha$ where j^* is the index of the foreground patch) and $\frac{1-\alpha}{m-1}$ for all the background patches (i.e $a_j = \frac{1-\alpha}{m-1}$ for all $j \neq j^*$).

3.1 Fixed Focus Loss Curves

For a given value of α, the problem of optimising for $\mathbf{g}$ is similar to the standard C-class classification problem with all three paradigms, with different data properties. As α increases, the optimisation problem over $\mathbf{g}$ becomes simpler for all three paradigms. When $\alpha = 1$, the optimisation problem over $\mathbf{g}$ for all three paradigms becomes equal to the classification problem of distinguishing between the foreground classes. However, the three paradigms simplify the optimisation problem over $\mathbf{g}$ in different ways for $\alpha \in [\frac{1}{m}, 1)$. An increase in α increases the margin between the C-classes in the case of soft attention, while it reduces the number of 'outlier points' in the case of hard attention, and is a hybrid of these two in the case of latent variable marginal likelihood.

(a) CIFAR10 Soft attention

(b) CIFAR10 Hard Attention

(c) CIFAR10 LVML

(d) Linear-Orthogonal Soft Attention

(e) Linear-Orthogonal Hard Attention
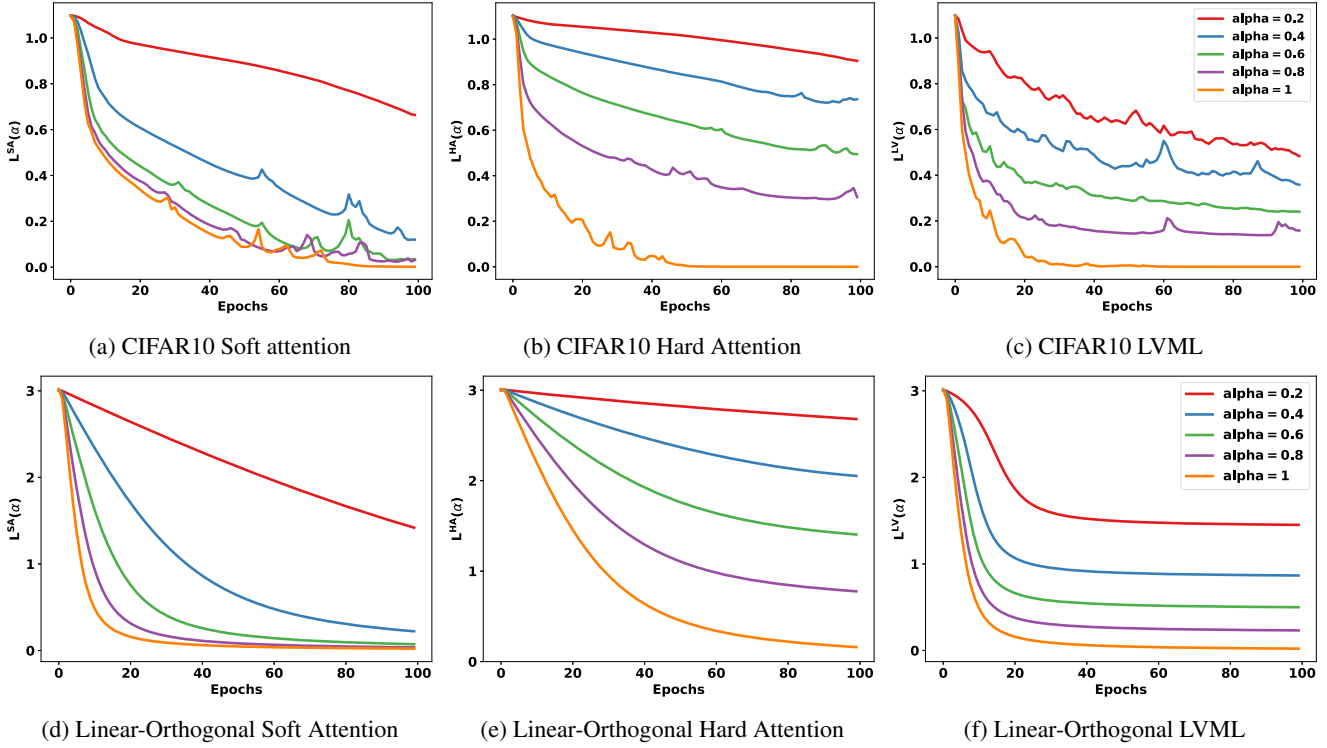
(f) Linear-Orthogonal LVML

Figure 2: Log-Loss curves for train data for SDC task under fixed focus setting. The top row contains CIFAR10 data for $n = 10000$ and $m = 5$ and the bottom row is for linear orthogonal setting for $C = 20$ and $m = 20$.

Concretely, the fixed focus loss values for a given classification model $\mathbf{g}$ and mosaic instance $\mathbf{X}$ with label y, and foreground index j^* are given below for the three attention paradigms.

$$L^{\text{FF,SA}} = -\log[\sigma_y(\mathbf{g}(\alpha\mathbf{x}_{j^*} + \tfrac{1-\alpha}{m-1}\sum_{j \neq j^*}\mathbf{x}_j))]$$

$$L^{\text{FF,LV}} = -\log[\alpha\sigma_y(\mathbf{g}(\mathbf{x}_{j*})) + \tfrac{1-\alpha}{m-1}\sum_{j \neq j^*}\sigma_y(\mathbf{g}(\mathbf{x}_j))]$$

$$L^{\text{FF,HA}} = -\alpha\log[\sigma_y(\mathbf{g}(\mathbf{x}_{j^*}))] + \tfrac{1-\alpha}{m-1}\sum_{j \neq j^*}\log[\sigma_y(\mathbf{g}(\mathbf{x}_j))]$$

From the above expressions, one can make the following observation. While it is possible to find a $\mathbf{g}$ that makes any of the above three expressions close to zero for any α and any given $(\mathbf{X}, y)$ pair, the population expectation of $L^{\text{FF, LV}}$ and $L^{\text{FF, HA}}$ is bounded away from 0 for any $\alpha < 1$. In the Hard Attention (HA) and Latent Variable (LV) paradigms, the value of 1-α represents the proportion of data coming from background segments that can have all possible labels. In the HA paradigm, $\alpha < 1$ indicates that a fraction of 1-α of the total data points are sampled from the set of background segments. Since background segments can appear with any foreground segment, they possess all possible labels, and achieving zero loss is not possible. In particular, this means that it is not possible to achieve low loss with the LVML and HA paradigms (on a large enough dataset) if the focus model f is such that $f(\mathbf{x}_j)$ is not large for a unique j (we call such a f as non-confident). On the other hand, low losses are possible in the soft attention paradigm even with non-confident focus models.

Figure 2 shows the evolution of fixed focus loss values when $\mathbf{g}$ is updated through gradient descent for varying values of α. Figure 2 gives such curves for both the CIFAR10 SDC data and a purely synthetic SDC setting (that we call the linear orthogonal setting) for

which we can derive the trajectory of the parameters when running gradient flow on the population loss. The details of this setting are in section 4. We denote the classifier $\mathbf{g}$ in the tth epoch when using the attention paradigm AP with fixed foreground focus score α as $\mathbf{g}^{\text{AP}}_{\alpha,t}$.

We make the following observations about the fixed focus losses based on Figure 2

- The soft attention loss eventually goes to zero for moderately large values of α.

- The hard attention and LVML losses flatten at a value above 0 for any $\alpha < 1$.

- The LVML fixed focus loss curves generally have a steeper decline at initialization than the other paradigms, and converges to a lesser loss value than the hard attention paradigm.

- The difference in the fixed focus loss curves for different α values is minimal at larger values of α in the case of soft attention, but remains significant for the other two paradigms.

- In the hard attention paradigm, the flat loss curve at small values of α indicates the difficulty faced by the classification module to improve when the focus model f is not confident and correct.

While the above trends are obvious in the synthetic linear-orthogonal setting, they are also clearly visible in the case of the CIFAR SDC dataset, indicating that this difference in behaviour is a characteristic of the paradigm used.

3.2 Focus Improvement Incentive Curves

In the previous section, we studied the evolution of the classification network, when the focus network f is (somehow magically) fixed to
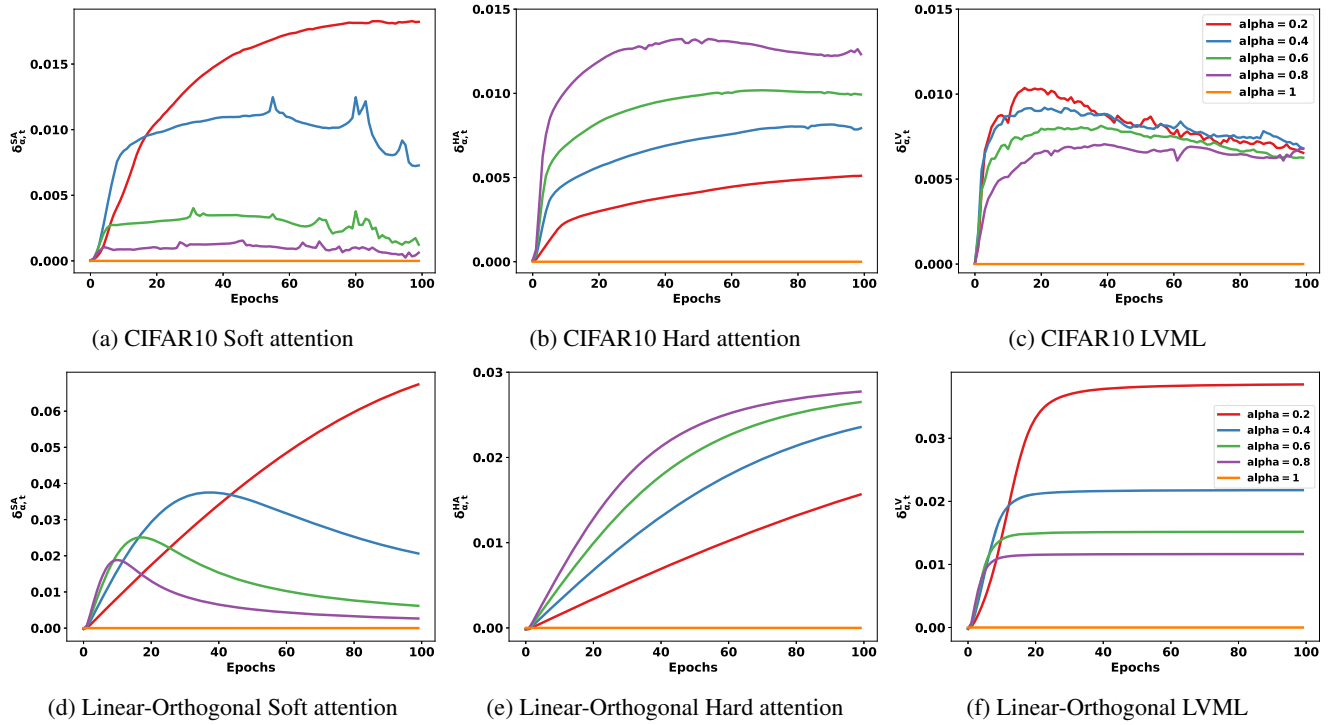
(a) CIFAR10 Soft attention　　　　(b) CIFAR10 Hard attention　　　　(c) CIFAR10 LVML

(d) Linear-Orthogonal Soft attention　　(e) Linear-Orthogonal Hard attention　　(f) Linear-Orthogonal LVML

Figure 3: Focus Improvement Incentive curves for train data for SDC task under fixed focus setting. The top row contains CIFAR10 data for $n = 10000$ and $m = 5$ and the bottom row is for linear orthogonal setting for $C = 20$ and $m = 20$.

give a score of α to the foreground and $\frac{1-\alpha}{m-1}$ the background segments. Note that this is trivially possible for $\alpha = \frac{1}{m}$ by setting the focus network to be identically equal to 0.

The prevailing intuition for learning dynamics of attention models is the following. The classification network $\mathbf{g}$ can capture a non-negligible signal distinguishing the C foreground classes even with this trivial focus network and improve. As $\mathbf{g}$ gets better, the incentive for the focus network to 'improve' by giving higher scores to the foreground segment increases, and thereby making the optimisation problem faced by the classification network $\mathbf{g}$ easier and kick-starts a virtuous cycle.

We quantify the above phenomenon for the three attention paradigms, by defining the 'incentive' for the focus model to improve as follows. For any given classification model $\mathbf{g}$, the focus improvement incentive at focus value α is given by $\mathbf{E}_{(\mathbf{x},y)} \left[-\frac{d}{d\alpha} L^{\text{FF, AP}}(\alpha, \mathbf{g}, \mathbf{X}, y) \right]$. This captures the reduction in the loss for a given classification model $\mathbf{g}$, as the focus model gets better. In particular, we study this quantity for the classifiers $\mathbf{g}^{\text{AP}}_{\alpha,t}$ obtained while minimising the fixed focus losses, where we approximate the derivative with a finite difference.

$$\delta^{\text{AP}}_{\alpha,t} = L^{\text{FF,AP}}(g^{\text{AP}}_{\alpha,t}, \mathbf{X}, y, \alpha) - L^{\text{FF,AP}}(g^{\text{AP}}_{\alpha,t}, \mathbf{X}, y, \alpha')$$

where, $\alpha' = \min(\alpha + 0.01, 1)$ and AP takes values in $\{\text{SA}, \text{HA}, \text{LV}\}$.

We make the following observations based on the focus improvement incentive curves in Figure 3.

- The focus improvement incentive for the soft attention paradigm diminishes with increase in α (when considering $\mathbf{g}^{\text{SA}}_{\alpha,t}$ for moderately large t), and becomes quite negligible as α approaches 1. (See Figure 3(a, d)).

- The focus improvement incentive for the hard attention paradigm increases with increase in α, and remains quite small till very large t for small α (See Figure 3(b, e)).

- The focus improvement incentive for the LVML paradigm also diminishes with increase in α similar to the soft attention paradigm, but the fall in incentive is not nearly as steep and remains bounded away from zero even for large values of α (See Figure 3(c, f)).

3.3 Explanation for the Attention Paradigm Behaviour

Based on the observations in Sections 3.1 and 3.2, we now attempt an explanation of the behaviour of the attention paradigms when the entire model (i.e. both f and $\mathbf{g}$) is optimised simultaneously. In particular, we explain the findings in Section 2.4 that comment on the results in Figure 1.

The final soft attention Focus-prediction heat maps in Figure 1 have a large number of mosaic instances with medium values for the focus score a_z, because the need for the focus network to be confident and correct is not present – i.e. the soft attention loss $L^{\text{FF, SA}}_{\alpha}$ can be driven close to zero even for $\alpha < 1$ (See green, purple and orange curves in Figure 2 (a,d)). The incentive for increasing α also falls rapidly with increase in α through the optimisation process (See Figure 3 (a,d).

The final hard attention model has a large fraction of instances where the focus score on the foreground segment a_z is small (< 0.2) or large (> 0.8). This is likely due to the fact that the fixed focus loss curves are almost flat for small values of α (See red curve in Figure 2(b,d)) and the incentive for the focus network to improve is significantly lesser (see red curve in Figure 3(b,d)). However, the large incentive for the focus to improve when α is moderately large (see green, blue and red curves in Figure 3(b,d)) ensures that very

few instances remain in the moderate a_z region: i.e. the focus scores go all the way to 1 or stay around $\frac{1}{m}$.

The latent variable marginal likelihood model seems to combine the best of both the soft and hard attention paradigms: a significant incentive for the focus network to improve for small α that does not decay too rapidly even for large α, and hence it appears to achieve the best result while training FCAMs for the SDC task.

4 The Linear Orthogonal Setting

In this section, we define a simple (toy-like) SDC task for which the gradient flow with the expected population gradient can be computed in closed form. The generative model under this setting has following assumptions.

Assumption 1 *The foreground distributions are all supported on a single point, i.e. $D_y = \delta_{s_y}$, where $s_1, s_2, \ldots, s_C$ are all orthogonal vectors in $\mathbb{R}^d$.*

Assumption 2 *The background distribution D_0 has zero mean.*

Assumption 3 *The support of the background distribution D_0 is orthogonal to the vectors $s_1, s_2, \ldots, s_C$.*

The FCAM model we use for this analysis is simply a linear model for $f : \mathbb{R}^d \rightarrow \mathbb{R}$ parameterised by a d-dimensional vector u and a linear model for $g : \mathbb{R}^d \rightarrow \mathbb{R}^C$ parameterised by a $C \times d$ weight matrix $[w_1, w_2, \ldots, w_C]^\top$. While these assumptions are clearly very restrictive, the simple nature of this setting allows us to clearly see the fundamental difference between the three paradigms.

The ideal solution to the above problem would be such that the focus model f has the background subspace in its null-space thus ensuring that in every mosaic instance $\mathbf{X}$, the foreground segment $\mathbf{x}_{j^*}$ would get a positive attention score $f(\mathbf{x}_{j^*})$ while all the background patches evaluate to 0. The ideal classification model $\mathbf{g}$ would then be a model that classifies s_y as y for all $y \in [C]$. We show that gradient flow on the population in all three paradigms do converge to the same ideal solutions (up to a scalar multiple) but at differing rates.

4.1 Fixed focus losses

In this section, we first derive the dynamics of the classification model $\mathbf{g}$, when the focus model is fixed to give a score of α to the foreground segment and $\frac{1-\alpha}{m-1}$ to all the background segments.

Theorem 1 *In the linear-orthogonal setting, when the focus network f is fixed to score the foreground segment with value $\alpha \in [0, 1]$ the parameter w of the classification model $\mathbf{g}$ varies as follows under gradient flow from a zero initialisation:*

$$w_k^{\text{AP}}(t) = \mu^{\text{AP}}(t) \left[s_k - \frac{1}{C} \sum_{k'=1}^{C} s_{k'} \right]$$

where AP *takes values in* $\{\text{SA}, \text{HA}, \text{LV}\}$ *and* $k \in [C]$. *The scalar* $\mu^{\text{AP}}(t)$ *varies as follows:*

$$\frac{\mathrm{d}\mu^{\text{SA}}(t)}{\mathrm{d}t} = \frac{\alpha}{\exp(\alpha\mu^{\text{SA}}(t)) + C - 1} \tag{1}$$

$$\frac{\mathrm{d}\mu^{\text{HA}}(t)}{\mathrm{d}t} = \frac{\alpha\beta^{\text{HA}}(t)}{\exp(\mu^{\text{HA}}(t))} \tag{2}$$

$$\frac{\mathrm{d}\mu^{\text{LV}}(t)}{\mathrm{d}t} = \frac{\alpha(\beta^{\text{LV}}(t))^2}{Z(t)\exp(\mu^{\text{LV}}(t))} \tag{3}$$

where $Z(t) = \alpha\beta^{\text{LV}}(t) + \frac{1-\alpha}{C}$, $\beta^{\text{AP}}(t) = \frac{\exp(\mu^{\text{AP}}(t))}{\exp(\mu^{\text{AP}}(t)) + C - 1}$

The losses of the corresponding classification model and the focus improvement incentive are illustrated in Figure 2(d,e,f) and Figure 3 (d,e,f) respectively. The similarity of linear orthogonal setting plots with CIFAR10 plots in Figures 2 and 3 suggests that differences in behaviour among the attention paradigms is due to an intrinsic property of the paradigms and not an artifact of the data.

4.2 Gradient flow trajectory for the full FCAM

The extreme simplicity of the linear orthogonal setting allow for the fixed focus model used in the previous section to actually manifest during training. In fact, when f and $\mathbf{g}$ are optimised simultaneously, the focus model exactly takes the trajectory of the fixed focus model with α that increases from $\frac{1}{m}$ to 1. Based on this we can derive the joint trajectory of the focus network parameter u and classification network parameter w as follows.

Theorem 2 *Under the linear-orthogonal setting, when the initial values of u and w are set to zero, the parameters evolve under population gradient flow as follows*

$$w_k^{\text{AP}}(t) = \mu^{\text{AP}}(t) \left[s_k - \frac{1}{C} \sum_{k'=1}^{C} s_{k'} \right]$$

$$u^{\text{AP}}(t) = \nu^{\text{AP}}(t) \sum_{k=1}^{C} s_k$$

where AP *takes values in* $\{\text{SA}, \text{HA}, \text{LV}\}$ *and* $k \in [C]$. *The scalar* $\nu^{\text{AP}}(t)$ *varies as follows:*

$$\frac{\mathrm{d}\nu^{\text{SA}}(t)}{\mathrm{d}t} = \frac{\mu^{\text{SA}}(t)(C-1)(\alpha^{\text{SA}}(t) - (\alpha^{\text{SA}}(t))^2)}{C(\exp(\alpha^{\text{SA}}(t)\mu^{\text{SA}}(t)) + C - 1)} \tag{4}$$

$$\frac{\mathrm{d}\nu^{\text{HA}}(t)}{\mathrm{d}t} = \frac{\log[C\beta^{\text{HA}}(t)]}{C}(\alpha^{\text{HA}}(t) - (\alpha^{\text{HA}}(t))^2) \tag{5}$$

$$\frac{\mathrm{d}\nu^{\text{LV}}(t)}{\mathrm{d}t} = \frac{\alpha^{\text{LV}}(t)}{C}\left[\frac{\beta^{\text{LV}}(t)}{Z(t)} - 1\right] \tag{6}$$

where, $Z(t) = \alpha^{\text{LV}}(t)\beta^{\text{LV}}(t) + \frac{1 - \alpha^{\text{LV}}(t)}{C}$,

$\alpha^{\text{AP}}(t) = \frac{\exp(\nu^{\text{AP}}(t))}{\exp(\nu^{\text{AP}}(t)) + m - 1}$,

$\beta^{\text{AP}}(t) = \frac{\exp(\mu^{\text{AP}}(t))}{\exp(\mu^{\text{AP}}(t)) + C - 1}$, *and*

$\frac{\mathrm{d}\mu^{\text{AP}}(t)}{\mathrm{d}t}$ *follows the expression in Theorem 1.*

Theorem 1 and 2 implies that under linear-orthogonal setting, the three algorithms trajectory differs only in terms of scalar multiples $\mu(t)$ and $\nu(t)$. Figure 4 shows the simulated evolution of $\mu(t)$ and $\nu(t)$ for two different values of m, C. The LVML model clearly converges faster than both soft and hard attention. The hard attention model converges to larger value for $\nu(t)$ corresponding to a more confident focus model. The proofs for theorems are in given in the appendix.
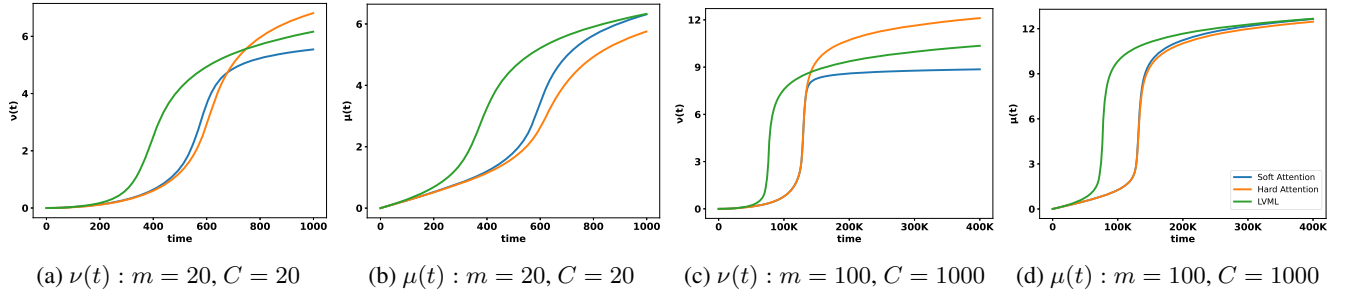
(a) $\nu(t) : m = 20, C = 20$ (b) $\mu(t) : m = 20, C = 20$ (c) $\nu(t) : m = 100, C = 1000$ (d) $\mu(t) : m = 100, C = 1000$

Figure 4: Evolution of the focus model ($\nu(t)$ in Theorem 2) and classification model ($\mu(t)$ in Theorem 2) for linear orthogonal setting

-	SAIF (%)				Accuracy (%)			
Dataset	Hard	Soft	Hybrid	LVML	Hard	Soft	Hybrid	LVML
CIFAR10 ($m = 5$)	37.8	37.1	59.0	62.0	63.3	75.3	76.6	77.9
CIFAR10 ($m = 20$)	1.9	23.8	38.2	47.2	42.1	65.2	66.0	71.0
CIFAR100 ($m = 5$)	13.6	79.2	83.2	77.5	46.6	83.2	80.0	79.7
CIFAR100 ($m = 20$)	1.7	67.0	66.8	66.5	8.2	72.2	71.0	69.0
HateXplain-1	2.7	4.6	21.6	22.6	54.3	51.6	54.6	54.6
HateXplain-2	23.4	9.2	24.6	19.2	49.8	49.63	49.0	44.7
MSCOCO (Bleu4 Score)	11.1	23.2	24.2	-	10.4	20.9	20.4	-

Table 1: Strongly Accurate Interpretable Fraction and Accuracy for various datasets and attention paradigms

5 A Hybrid Attention Paradigm and Experiments

A natural algorithm design goal here is to require that the final FCAM be both accurate and interpretable – which can be enforced by requiring that the predicted probability score of the true class and the focus score of the true foreground are both large. This reduces to requiring that the top right number in the Focus-Prediction heatmap (SAIF) be as large as possible. The LVML loss function corresponding to maximum likelihood has ideal properties for achieving this, but it suffers from computational issues when the number of segments m is large.

Based on the observations in Sections 3.1 and 3.2, a simple algorithm is proposed that combines the properties of both soft and hard attention. Soft attention provides strong incentives for the focus model to improve at initialization but stagnates later on. In the Hard attention paradigm the incentive for focus model to improve is small at initialization but becomes larger once a non-trivial focus model is attained. This suggests a 'Hybrid' approach where an FCAM trained with soft attention is used as the initial step for hard attention learning. Previous works have shown hard attention to actually be a hybrid method, where the model randomly selects either soft or hard attention during each epoch, however, this hybrid nature was an after effect of the unintentional random selection [20]. Our work is different in terms of conscious selection between hard and soft attention based on empirical observation.

We perform experiments on semi-synthetic (refer section 2.4) and real-world datasets. The "HateXplain" dataset [12] is used to classify text into one of three categories: hate speech, offensive, or normal. The dataset also includes explanations for the assigned label in the form of "rationales", which highlight specific portions of the text. The focus score of the true foreground is calculated as $a_{z^*} = \sum_{z \in z^*} a_z$ (where z^* specifies the locations in input responsible for the label). Experiments were conducted using both a standard embedding layer and a self-attention embedding layer (1 and 2 respectively), as shown in Table 1).

Experiments were also conducted on the MSCOCO (2014) dataset for image captioning using an encoder-decoder model with attention, as described in [20]. The definition of SAIF was adapted for sequence generation by measuring the overlap between the attention vector and

the bounding box of an object when a word corresponding to that object is output by the decoder [14]. Due to the large number of classes and vocabulary, as well as the large number of patches ($m = 196$), the threshold used for SAIF was relaxed to 0.3 instead of the 0.8 used in other experiments.

Table 1 gives the SAIF and accuracy results of the experiments across different paradigms with equal compute time. The hybrid approach seems to increase the SAIF performance over soft attention without affecting the accuracy.

6 Conclusion

In this paper, we study the learning dynamics of different paradigms of attention models observing that, in terms of interpretability, soft attention performs well at initialization and hard attention perform well at later stages of training. We propose a hybrid approach and demonstrate an improvement without incurring the same cost as maximum likelihood methods. This approach is applicable to any task, where the class label depends on a small but unknown segment of the input, and this location information is absent in the training data. However evaluating the improvement in interpretability requires access to this location information – e.g. datasets such as HateXplain, TVQA+ [10].

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, 'Neural machine translation by jointly learning to align and translate', in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, eds., Yoshua Bengio and Yann LeCun, (2015).

[2] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin, 'A statistical approach to machine translation', *Comput. Linguist.*, **16**(2), 79–85, (jun 1990).

[3] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer, 'The mathematics of statistical machine translation: Parameter estimation', *Comput. Linguist.*, **19**(2), 263–311, (jun 1993).

[4] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, 'End-to-end continuous speech recognition using attention-based

recurrent nn: First results', in *NIPS 2014 Workshop on Deep Learning, December 2014*, (2014).

[5] Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander Rush, 'Latent alignment and variational attention', in *Advances in Neural Information Processing Systems*, eds., S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, volume 31. Curran Associates, Inc., (2018).

[6] Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato, 'Classical structured prediction losses for sequence to sequence learning', in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 355–364, New Orleans, Louisiana, (June 2018). Association for Computational Linguistics.

[7] Alex Graves, 'Generating sequences with recurrent neural networks', *CoRR*, **abs/1308.0850**, (2013).

[8] Sarthak Jain and Byron C. Wallace, 'Attention is not Explanation', in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, Stroudsburg, PA, USA, (2019). Association for Computational Linguistics.

[9] Alex Krizhevsky and Geoffrey Hinton, 'Learning multiple layers of features from tiny images', Technical Report 0, University of Toronto, Toronto, Ontario, (2009).

[10] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal, 'TVQA+: Spatio-temporal grounding for video question answering', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8211–8225, Online, (July 2020). Association for Computational Linguistics.

[11] Haoye Lu, Yongyi Mao, and Amiya Nayak, 'On the dynamics of training attention models', in *International Conference on Learning Representations*, (2021).

[12] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee, 'Hatexplain: A benchmark dataset for explainable hate speech detection', *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**(17), 14867–14875, (May 2021).

[13] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu, 'Recurrent models of visual attention', in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, p. 2204–2212, Cambridge, MA, USA, (2014). MIT Press.

[14] Lakshmi Narayan Pandey, Rahul Vashisht, and Harish G. Ramaswamy, 'On the interpretability of attention networks', in *Proceedings of The 14th Asian Conference on Machine Learning*, eds., Emtiyaz Khan and Mehmet Gonen, volume 189 of *Proceedings of Machine Learning Research*, pp. 832–847. PMLR, (12–14 Dec 2023).

[15] Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui, 'Attention interpretability across nlp tasks', *ArXiv*, **abs/1909.11218**, (2019).

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin, 'Attention is all you need', in *Advances in Neural Information Processing Systems*, eds., I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, volume 30, pp. 5998–6008. Curran Associates, Inc., (2017).

[17] Sarah Wiegreffe and Yuval Pinter, 'Attention is not not explanation', in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, Hong Kong, China, (November 2019). Association for Computational Linguistics.

[18] Shijie Wu and Ryan Cotterell, 'Exact hard monotonic attention for character-level transduction', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1530–1537, Florence, Italy, (July 2019). Association for Computational Linguistics.

[19] Shijie Wu, Pamela Shapiro, and Ryan Cotterell, 'Hard non-monotonic attention for character-level transduction', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4425–4438, Brussels, Belgium, (oct-nov 2018). Association for Computational Linguistics.

[20] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, 'Show, attend and tell: Neural image caption generation with visual attention', in *Pro-ceedings of the 32nd International Conference on Machine Learning*, eds., Francis Bach and David Blei, volume 37 of *Proceedings of Machine Learning Research*, pp. 2048–2057, Lille, France, (07–09 Jul 2015). PMLR.