Attention based models in End-to-End ASR

Exploration of Attention in ESPNET toolkit

Shreekantha Nadig

November 9, 2018

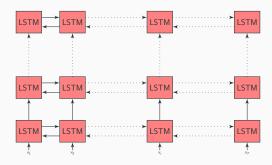International Institute of Information Technology - Bangalore

Table of contents

Introduction

No Attention [Equal Attention?]
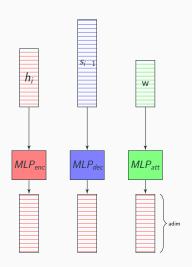
Mostly *eproj* $\neq$ *dunits* $\neq$ *adim*
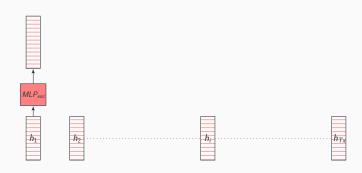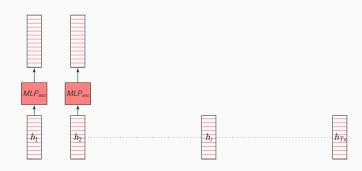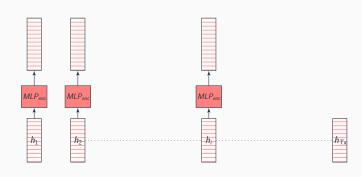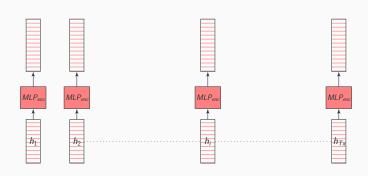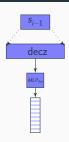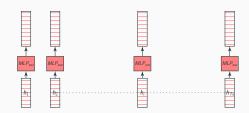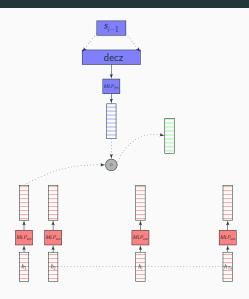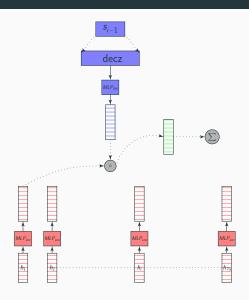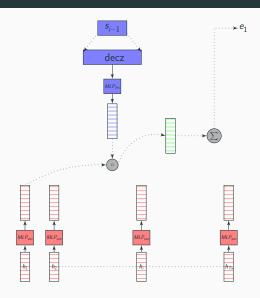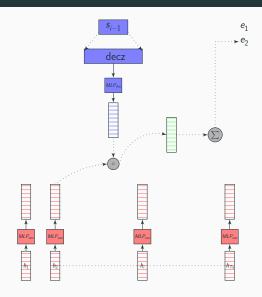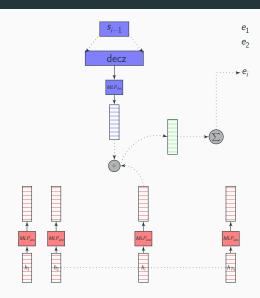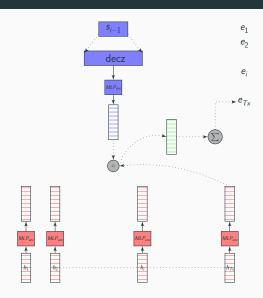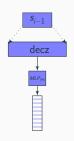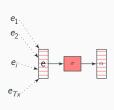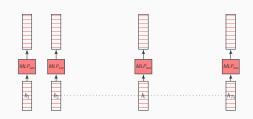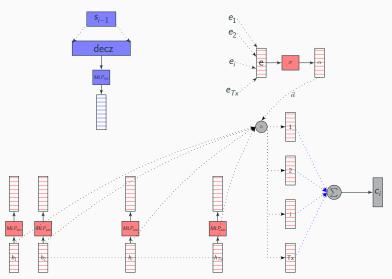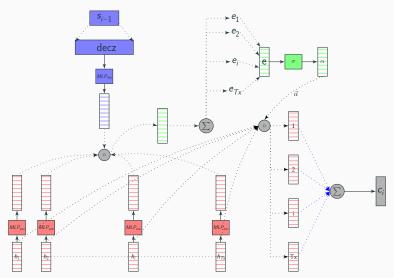
Dot product Attention

MLP_{enc}

h_1 h_2 h_j h_{Tx}

s_{i-1}

decz

MLP_{dec}

e_1

e_2

e_i

e_{Tx}

$\sum$

MLP_{enc} MLP_{enc} MLP_{enc} MLP_{enc}

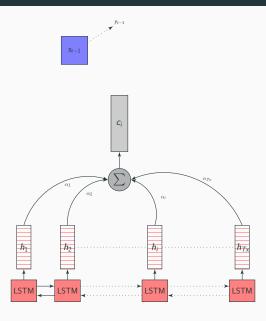h_1 h_2 h_i h_{Tx}

Additive Attention

s_{i-1}

decz

MLP_{dec}

e_1

e_2

e_i

e_{Tx}

MLP_{att}

$\sum$

MLP_{enc} MLP_{enc} MLP_{enc} MLP_{enc}

h_1 h_2 h_i h_{Tx}

Location Aware Attention

$$w_1^{nf} \ w_2^{nf} \ \ldots w_i^{nf} \ \ldots w_{filt\_size}^{nf}$$

$$w_1^2 \ w_2^2 \ \ldots w_i^2 \ \ldots w_{filt\_size}^2$$

$$w_1^1 \ w_2^1 \ \ldots w_i^1 \ \ldots w_{filt\_size}^1$$

$*$

| 0 | 0 $\ldots$ | 0 | α_1^{t-1} | α_2^{t-1} | $\ldots\ldots\ldots\alpha_i^{t-1}$ | $\ldots\ldots\ldots\alpha_{Tx}^{t-1}$ | 0 | 0 $\ldots$ | 0 |

$$\alpha_1^{t-1} \qquad \alpha_2^{t-1} \ \ldots\ldots\ldots\alpha_i^{t-1} \ \ldots\ldots\ldots\alpha_{Tx}^{t-1}$$

$$\boxed{\beta_1^1 \quad \beta_2^1 \ldots \ldots \ldots \beta_i^1 \ldots \ldots \ldots \beta_{Tx}^1}$$

$$||$$

$$\boxed{w_1^1 \; w_2^1 \; \ldots \; w_i^1 \; \ldots \; w_{filt\_size}^1}$$

$$*$$

$$\boxed{0 \quad 0 \ldots \quad 0 \; \alpha_1^{t-1} \quad \alpha_2^{t-1} \ldots \ldots \ldots \alpha_i^{t-1} \ldots \ldots \ldots \alpha_{Tx}^{t-1} \; 0 \quad 0 \ldots \quad 0}$$

$$\beta_1^2 \quad \beta_2^2 \ldots\ldots\ldots \beta_i^2 \ldots\ldots\ldots \beta_{Tx}^2$$

$$\|$$

$$w_1^2 \; w_2^2 \ldots w_i^2 \ldots w_{filt\_size}^2$$

$$*$$

$$0 \quad 0 \ldots \quad 0 \; \alpha_1^{t-1} \quad \alpha_2^{t-1} \ldots\ldots\ldots \alpha_i^{t-1} \ldots\ldots\ldots \alpha_{Tx}^{t-1} \; 0 \quad 0 \ldots \quad 0$$

Location Aware Attention

Tx

$$\beta_1^1$$
.
.
$$\beta_1^{10}$$

$$\beta_{Tx}^1$$
.
.
$$\beta_{Tx}^{10}$$

//

$$w_1^{nf} \ w_2^{nf} \ \ldots \ w_i^{nf} \ \ldots \ w_{filt\_size}^{nf}$$

$$w_1^2 \ w_2^2 \ \ldots \ w_i^2 \ \ldots \ w_{filt\_size}^2$$

$$w_1^1 \ w_2^1 \ \ldots \ w_i^1 \ \ldots \ w_{filt\_size}^1$$

*

| 0 | 0 ... | 0 α_1^{t-1} | α_2^{t-1} | .... α_i^{t-1} | .... α_{Tx}^{t-1} | 0 | 0 ... | 0 |

$$\alpha_1^{t-1} \quad \alpha_2^{t-1} \quad .... \alpha_i^{t-1} \quad .... \alpha_{Tx}^{t-1}$$

2D Location Aware Attention

Location Aware Recurrent Attention

Coverage mechanism Attention

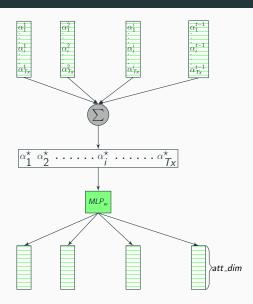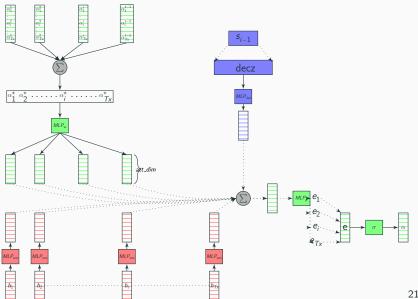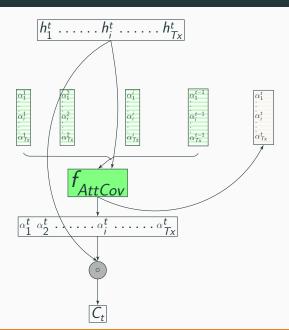# Coverage mechanism location aware Attention