	Arabic	Basque	Czech	Danish	Dutch	Portuguese	Slovene	Swedish
Standard EM								
DMV	**45.8**	41.1	31.3	50.8	47.1	**36.7**	36.7	43.5
Neural DMV	43.4	**46.5**	**33.1**	**55.6**	**49.0**	30.4	**42.2**	**44.3**
Softmax EM $\sigma = 0.25$								
DMV	49.3	45.6	30.4	43.6	**46.1**	33.5	29.8	**50.3**
Neural DMV	**54.2**	**46.3**	**36.8**	**44.0**	39.9	**35.8**	**31.2**	49.7
Softmax EM $\sigma = 0.5$								
DMV	**54.2**	47.6	**43.2**	38.8	**38.0**	33.7	23.0	37.2
Neural DMV	44.6	**48.9**	33.4	**50.3**	37.5	**35.3**	**32.2**	**43.3**
Softmax EM $\sigma = 0.75$								
DMV	42.2	**48.6**	22.7	41.0	**33.8**	33.5	**23.2**	41.6
Neural DMV	**56.7**	45.3	**31.6**	**41.3**	33.7	**34.7**	22.9	**42.0**
Viterbi EM								
DMV	32.5	47.1	27.1	39.1	37.1	32.3	23.7	42.6
Neural DMV	**48.2**	**48.1**	**28.6**	**39.8**	**37.2**	**36.5**	**39.9**	**47.9**

Table 3: DDA results (on sentences no longer than 10) on eight additional languages. Our neural based approaches are compared with traditional approaches using standard EM, softmax EM (parameterized by σ) and Viterbi EM.

Activation function	**WSJ10**
ReLU	**69.7**
Leaky ReLU	67.0
Tanh	66.2
Sigmoid	62.5
Linear	55.1

Table 5: Comparison between activation functions.

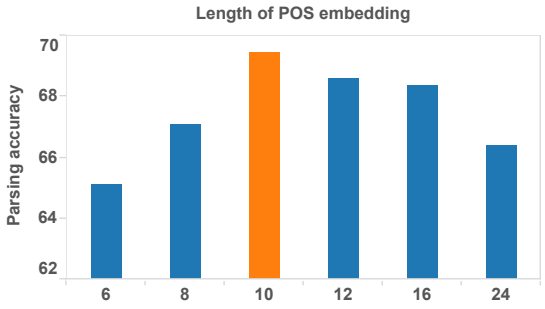

Figure 3: Parsing accuracy vs. length of POS embedding

different POS tags. On the other hand, when the dimension is too high (such as $dim = 30$), since we have only 35 POS tags, the neural network is prone to overfitting.

Shared parameters An alternative to our neural network architecture is to have two separate neural networks to compute `CHILD` and `DECISION` rule probabilities respectively. The embeddings of the head POS tag and the valence are not shared between the two networks. As can be seen in Table 6, sharing POS tags embeddings attribute to better performance.

	WSJ10	**WSJ**
Separate Networks	68.6	52.1
Merged Network	**69.7**	**52.5**

Table 6: Comparison between using two separate networks and using a merged network.

5 Model Analysis

In this section, we investigate what information our neural based DMV model captures and analyze how it contributes to better parsing performance.

5.1 Correlation of POS Tags Encoded in Embeddings

A main motivation of our approach is to encode correlation between POS tags in their embeddings so as to smooth the probabilities of grammar rules involving correlated POS tags. Here we want to examine whether the POS embeddings learned by our approach successfully capture such correlation.

We collected the POS embeddings learned in the experiment described in section 4.3 and visualized them on a 2D plane using the t-SNE algorithm (Van der Maaten and Hinton, 2008). t-SNE is a dimensionality reduction algorithm that maps data from a high dimensional space to a low dimensional one (2 or 3) while maintaining the distances between