연속 음성 인식을 위한 그룹 식별 신경망과 연결 강도 초기화에 대한 연구

°최 기훈*, 김 이형*, 이 성권*, 김 순협*
*광운대학교 컴퓨터공학과

A Study on the Verify Group Neural Network and Weight Initialization for Continuous Speech Recognition

°GiHoon Choi*, LeeHyung Kim*, SeungKwon Lee*, SoonHyob Kim*
*Dept. of Computer Engineering, KwangWoon Univ.

1. 요약

본 논문은 연속 음성 인식을 위한 신경망과 학습 속도를 줄이기 위한 연결강도 초기화에 관해 다루고 있다[4][6][7]. 우선 음소를 여러 개의 그룹으로 나눈 후 각각의 그룹에 대한 음소를 인식하는 신경망과 자신의 그룹을 판별하는 VGNN(Verify Group Neural Network)으로 신경망을 구성한다. 여기서 구성되는 신경망은 각각의 음소를 인식하는 출력을 낼뿐 아니라, 입력이 자신의 그룹에 속하는지 그렇지 않은지를 판별하는 출력을 낸다. 이런 신경망을 학습시키는 데 상당한 시간이 걸리므로 이 신경망의 학습 속도를 줄이기 위해 학습 데이타를 사용하여 신경망의 연결 강도를 초기화한다.

2. 서론

기존의 연속음성을 인식하기 위한 신경망은 각각의 음소를 판별하는 신경망과 음소가 어떤 그룹에 속하는 지를 판별하는 신경망으로 구성되어 있다[3]. 이러한 신경망은 한 그룹의 신경망에 다른 그룹에 속하는 음소가 들어오면 매우 큰 값을 나타내기도 한다. 즉, 학습이 되지 않은 다른 그룹의 음소에 대해 어떤 값이 나올지를 알 수가 없다. 따라서 이 신경망은 음소가 어떤 그룹에 속하는지를 판별하는 신경망에 너무 의존적이 된다. 따라서 이러한 단점을 없애고 연속어를 인식하기 위해서는 각각의 그룹에 대한 음소를 판별하는 추가의 신경망(VGNN)을 구성한다.

VGNN의 학습 시간을 단축하기 위해서 학습 데이타로부터 신경망의 연결강도를 초기화한다. 신경망은 3층의 MLP(Multilayer Perceptron)로 구성되어 있다. 여기서 은닉층은 hyperplane segment의 기능을 수행하기 때문에 이러한 hyperplane segment가 입력의 cluster를 분리하게끔 초기화하고 출력의 연결강도는 은닉층의 출력값으로 SVD (Singular Value Decomposition)을 사용하여 초기화한다. 이렇게 초기화된 신경망은 임의의 연결강도를 초기화한 신경망보다 빠르게 학습을 마친다.

본 논문은 처음에 연결강도 초기화 방법에 대해 다루고 다음에 연속 음성을 인식하기 위한 VGNN에 대해 다룬다.

3. 연속 음성 인식을 위한 네트웍

연속 음성 인식을 위한 전체 시스템의 블럭다이어그램은 다음과 같다.[1][2][3]

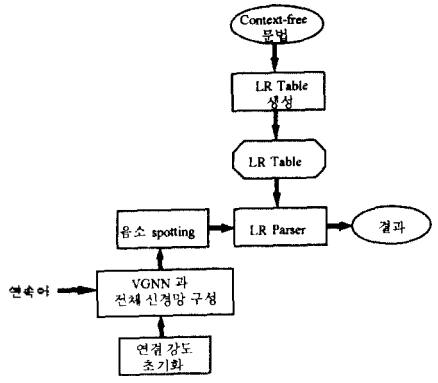

그림1. 전체 시스템 블럭다이어그램

3.1 네트웍의 학습속도를 개선하기 위한 연결강도 초기화

3층의 MLP는 hyperplane classifier의 기능을 한다[6][7]. 입력에서의 연결강도가 hyperplane segment의 기능을 수행하고 출력의 연결강도는 clustering 기능을 수행한다. 따라서, 입력의 hyperplane segment를 학습 데이타의 clustering center를 구분하게끔 초기화한다.

출력층의 각각의 노드는 하나의 discriminant function의 기능을 수행하는 데 수식은 다음과 같다.

$$f_i(X) = O_i = g(\sum_j w_{ij} S_j(X) + b_i)$$

여기서 $S_j(X)$는 j번째 은닉층의 출력을 나타낸다.

$$S_j(X) = V_j = g(h_j(X)) = g(\sum_i w_{ji} x_i + b_j)$$

$h_j(X)$는 입력 공간을 분할하는 hyperplane segment이다.

이러한 $h_i(X)$를 두개의 cluster center를 구별하게끔 초기화한다. 두 cluster center의 경계를 분할하는 식은 다음과 같다.

$$h_i(X) = 0 = X^T(P-Q) - \frac{1}{2}(P^TP - Q^TQ)$$

X는 입력 학습 벡터를 말하고, P와 Q는 clustering center를 나타내는 벡터이다. MLP에서 입력층과 은닉층의 연결강도에 의한 hyperplane segment는 다음과 같다.

$$h_i(X) = \sum_i w_{ij}x_i + b_i = w_i^T X + b_i$$

위의 두 식은 서로 같으므로

$$w_i = \beta(P-Q)$$

$$b_i = -\frac{\beta}{2}(P^TP - Q^TQ)$$

β는 scaling factor이다.

이렇게 함으로써 입력층과 은닉층의 연결강도는 cluster center를 분할하는 hyperplane segment가 된다.

이번에는 은닉층과 출력층의 연결강도를 초기화 한다. 입력층의 연결강도가 초기화되었기 때문에 입력에 대한 은닉층의 출력을 계산할 수가 있다. 그리고 입력에 대한 출력 결과값을 알고 있기 때문에 sigmoid 함수를 역변환함으로써 그 sigmoid 함수의 입력을 계산할 수 있다. 따라서 이 계산된 값과 은닉층의 출력값으로써 연결강도를 초기화할 수 있다.

$$I \times W = O$$

I : 은닉층의 출력값

W : 출력층과 은닉층과의 연결강도

O : sigmoid 함수의 입력값

이식에서 연결강도를 구하기 위해서 선형대수학에서 잘 알려진 SVD(Singular Value Decomposition)을 사용하여 구한다[5][7].

위와 같은 방법으로 제안된 신경망의 연결강도를 초기화한다. 이렇게 초기화된 신경망의 각각의 그룹에 대한 인식률은 표1.과 같다.

표1. 각각 그룹의 신경망에 대한 음소 인식률

	음소	cluster 수	인식률 (%)	cluster 수	인식률 (%)
그룹1	p	6	67.14	7	43.20
	t	6	80.71	7	83.26
	k	7	70.00	8	27.58
	b	3	74.83	6	57.43
그룹2	d	4	61.90	6	63.69
	g	4	69.39	6	59.18
	p'	5	73.57	6	59.15
그룹3	t'	5	76.43	6	61.85
	k'	5	74.29	6	48.53
	m	4	75.71	6	71.34
그룹4	n	4	86.43	6	79.46
	ŋ	4	82.86	6	80.37
	a	1	100.00	2	95.84
	e	2	99.72	2	91.46
	o	1	94.51	2	92.86
그룹5	u	1	97.35	2	88.64
	i	2	98.82	2	93.89
	i	1	94.29	2	91.53
	e	2	99.42	2	97.01

3.2. 연속 음성을 인식하기 위한 네트웍

기존의 연속 음성을 인식하기 위한 신경망은 음소를 여러 개의 그룹으로 나눈 후 각각의 그룹에 대해 학습을 시킨다. 그리고, 여기에 음소가 어떤 그룹에 속하는지를 판별하는 신경망을 따로 두어 학습을 시킨다. 그룹을 구분하는 신경망에서 나온 출력에 의해 그 해당 음소를 구별하는 신경망의 연결강도의 비중을 높인다. 여기서 문제가 되는 것은 두 가지가 있다. 첫번째로 이와 같은 신경망은 음소가 어떤 그룹에 속하는 지를 판별하는 신경망에 민감하다는 것이다. 만약에 그룹을 구분하는 신경망이 전체 데이타에 대해 충분히 학습되지 않아 어느 정도의 오차를 갖는다면 이 오차는 다른 신경망에 아주 큰 영향을 갖게 된다. 따라서 전체 신경망의 성능을 저하시킨다. 두번째는 각각의 그룹은 다른 그룹의 데이타에 대해 전혀 학습이 되지 않았다는 것이다. 따라서 자신의 그룹에 대해 나온 결과가 다른 그룹에서 나온 값보다 작을 경우가 있다. 이런 경우는 그룹을 구분하는 신경망이 해당 그룹에 비중을 크게 함으로써 줄일 수가 있지만 이렇게 되면 전체 신경망이 더욱 더 그룹을 구분하는 신경망에 의존적이 된다.

이러한 문제점을 해결하기 위해서 각각의 그룹에 데이타가 자신의 그룹에 속하는지 아닌지를 판별하는 신경망을 둔다. 이런 경우 그룹을 구분하는 신경망(VGNN : Verify Group Neural Network)을 각각의 그룹 신경망에 추가시키는 것이다.

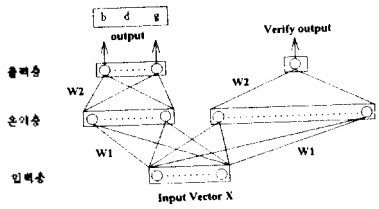

그림2. b/d/g 음소 그룹에 대한 하나의 Verify Group Neural Network

VGNN의 학습 데이타의 양은 매우 많다. 하지만 위의 연결 강도 초기화 방법을 이용하여 학습속도를 줄일 수 있다. VGNN의 음소 인식률은 다음과 같다.

표2. VGNN의 성능 평가

	음소 수	인식률
그룹1	6292	6161(97.92%)
그룹2	6292	5839(92.80%)
그룹3	6292	5831(92.67%)
그룹4	6292	5931(94.26%)
그룹5	6292	5869(93.28%)

4. 결론

학습 데이타로부터 연결강도를 초기화할 경우 그 학습 속도를 상당히 줄일 수 있다. 여기서 고려해야 할 것이 은닉층의 수이다. 은닉층의 수는 각 그룹의 cluster의 수에 따라 결정된다. cluster 수를 늘리면 은닉층의 노드가 늘어나게 되고 연결 강도의 수가 늘어나게 된다. 이때 학습한 데이타에 대해서 인식률이 좋고 빨리 학습을 마치게 된다. 하지만 전체 실험 데이타에 대해서는 인식률이 떨어지게 된다. 이것은 연결 강도가 많아짐으로 해서 신경망이 학습 데이타에 대해 일반화되지 않는다는 것이다. 즉, 이 신경망이 전역 최소점에 도달하는 것이 아니라 국부 최소점에 도달함을 의미한다. 따라서 이 cluster의 수를 최적화 하는 연구가 진행 되야 한다.

연속 음성을 인식하는 데 가장 힘든 점은 변이구간에 대한 것이다. 기존의 신경망으로는 변이 구간에 대한 학습이 이루어지지 않으므로 많은 오인식을 가져온다. VGNN을 연속어에 적용하였을 때 이러한 변이구간을 제거하는데 유용하다. 각각의 VGNN이 그러한 변이구간에 대해서는 낮은 값의 출력을 내므로 안정구간에 대해서만 인식한 후에 LR Parser에 입력에 넣게 된다. 이렇게 해서 VGNN은 연속어 인식에 유용하게 사용될 수 있다.

5. 참고문헌

[1] Masaru Tomita, "An efficient word lattice parsing algorithm for continuous speech recognition", ICASSP pp 1569-1572, 1986

[2] Kenji KITA, Takeshi KAWBATA, Hiroaki SAITO, "HMM continuous speech recognition using predictive LR parsing", ICASSP pp 703-706, 1989

[3] Hidefumi SAWAI, "TDNN-LR continuous speech recognition system usgin adaptive incrental TDNN Training", ICASSP pp 53-56, 1991

[4] Derrick Nguyen and Bernard Widrow, "Improving the learning speech of 2-layer neural networks by choosing initial values of the adaptive weights", IJCNN pp 21-26, 1989

[5] Pascale Hirschauer, Pascal Larzabal and Henri Clergot, "Design of Neural estimators : Second order backprogation, initialization and generalization", IJCNN pp 537-540, 1994

[6] Kishan G. Mehrotra, Chilukuri K. Mohan, and Sanjay Ranka, "Bounds on the number of samples needed for neural learning", IEEE Transaction on Neural Network, pp 548- 558, 1991

[7] S. Gavin Smyth, "Designing multilayer perceptrons from nearest-neighbor systems", IEEE Transaction on Neural Network, pp 329-333, 1992

[8] Keinosuke Fukunaga, "Introduction to Statistical Pattern Recognition", Academic Press, 1990.