

제공해주신 여러 PDF 파일들은 **딥러닝의 오차역전파법(5장)**과 **학습 관련 기술들(6장)**에 대한 강의 자료 및 실습 과제들로 구성되어 있습니다. 이 내용들을 종합하여 핵심 개념, 수식, 그리고 실무적 팁을 중심으로 A4 2페이지 분량으로 요약 정리해 드립니다.

신경망 학습 심화 요약 (5장 & 6장)

1. 오차역전파법 (Backpropagation)

오차역전파법은 신경망의 가중치 매개변수에 대한 손실 함수의 기울기를 효율적으로 계산하는 방법입니다. 수치 미분보다 계산 속도가 훨씬 빠르며, **계산 그래프(Computational Graph)**를 통해 시각적으로 이해할 수 있습니다¹¹¹.

1.1 기본 계층의 역전파 구조

국소적 미분을 연쇄 법칙(Chain Rule)에 따라 전달함으로써 전체 기울기를 구합니다.

- 덧셈 노드 ($\$z = x + y\$$): 상류에서 전해진 미분 ($\frac{\partial L}{\partial z}$)을 그대로 하류로 흘려보냅니다²²².
 - $\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z} \cdot 1$
- 곱셈 노드 ($\$z = xy\$$): 상류의 미분에 **순전파 때의 입력 신호들을 '서로 바꾼 값'**을 곱해서 하류로 보냅니다³³³³.
 - $\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z} \cdot y, \frac{\partial L}{\partial y} = \frac{\partial L}{\partial z} \cdot x$
 - 예시 (사과 쇼핑): 사과 가격의 변화가 전체 금액에 미치는 영향을 역으로 추적하여 계산⁴⁴⁴⁴.

1.2 활성화 함수 계층 * ReLU 계층:

* 순전파 때 $x > 0$ 이면 역전파 때 상류의 값을 그대로 전달합니다.

* 순전파 때 $x \leq 0$ 이면 역전파 때 신호를 차단(0을 전달)합니다 [cite: 502, 518].

- **Sigmoid 계층:**
 - 순전파의 출력 y 만으로 역전파를 계산할 수 있습니다.

- 수식: $\frac{\partial L}{\partial x} = \frac{\partial L}{\partial y} y^2 \exp(-x) = \frac{\partial L}{\partial y} y (1-y)$ ⁵⁵⁵.

1.3 Affine 및 Softmax 계층

- Affine 계층 (행렬의 내적):**
 - 입력 X 와 가중치 W 의 내적 계산입니다. 역전파 시에는 행렬의 형상(Shape)을 맞추기 위해 **전치 행렬(W^T, X^T)**을 사용합니다⁶⁶⁶.
 - $\frac{\partial L}{\partial X} = \frac{\partial L}{\partial Y} \cdot W^T, \frac{\partial L}{\partial W} = X^T \cdot \frac{\partial L}{\partial Y}$
- Softmax-with-Loss 계층:**
 - 소프트맥스 함수(출력 정규화)와 교차 엔트로피 오차(손실 함수)를 결합한 계층입니다.
 - 역전파가 ** $y - t$ ** (예측값 - 정답 레이블)**로 매우 깔끔하게 떨어집니다. 이는 오차를 앞 계층에 그대로 전달하는 성질을 가집니다⁷⁷⁷.

2. 매개변수 갱신 방법 (Optimizer) 손실 함수의 값을 최소화하는 매개변수를 찾는 '최적화' 기법들입니다.

기법	특징 및 수식	장단점
SGD	$W \leftarrow W - \eta \frac{\partial L}{\partial W}$ 기울어진 방향으로 일정 거리만큼 이동 ⁸ .	단순하지만, 비등방성 함수(타원형 등고선)에서는 탐색 경로가 지그재그로 비효율적임 ⁹ .
Momentum	$v \leftarrow \alpha v - \eta \frac{\partial L}{\partial W}, W \leftarrow W + v$ 물리적인 '관성'을 도입. 공이 그릇을 구르는 듯한 움직임 ¹⁰¹⁰¹⁰¹⁰ .	SGD보다 지그재그 움직임이 줄어듦.

AdaGrad	$\begin{aligned} h &\leftarrow h + \frac{\partial L}{\partial W} \odot \frac{\partial L}{\partial W}, W \\ &\leftarrow W - \eta \frac{1}{\sqrt{h}} \frac{\partial L}{\partial W} \end{aligned}$ <p>많이 변화한 매개변수의 학습률을 낮춤 (학습률 감소)¹¹.</p>	학습을 진행할수록 갱신 강도가 약해짐. (무한히 학습하면 갱신량이 0이 되는 단점 존재) ¹² .
Adam	Momentum과 AdaGrad의 융합 ¹³ .	최근 딥러닝에서 가장 폭넓게 사용되는 알고리즘 중 하나 ¹⁴ .

- 성능 비교: MNIST 데이터셋 학습 시, 일반적으로 SGD보다 **AdaGrad, Adam, Momentum**이 더 빠른 학습 속도와 높은 정확도를 보입니다¹⁵¹⁵¹⁵¹⁵.

3. 가중치 초기값 (Weight Initialization)

가중치 초기값을 0이나 균일한 값으로 설정하면, 오차역전파 시 모든 가중치가 똑같이 갱신되어 학습이 이루어지지 않습니다(가중치 대칭 파괴 필요)¹⁶.

3.1 은닉층 활성화값 분포와 초기값

활성화값들이 0과 1에 치우치면 **기울기 소실(Gradient Vanishing)**이 발생하고, 특정 값에 집중되면 표현력 제한 문제가 발생합니다¹⁷¹⁷¹⁷¹⁷.

3.2 권장 초기값 설정¹⁸¹⁸¹⁸¹⁸

활성화 함수	권장 초기값 (이름)	특징 (표준편차)
Sigmoid, Tanh	Xavier 초기값	앞 계층의 노드가 n 개일

		때, $\frac{1}{\sqrt{n}}$ 분포 사용. 선형적인 함수에 적합 ¹⁹ .
ReLU	He 초기값	앞 계층의 노드가 n 개일 때, $\sqrt{2/n}$ 분포 사용. ReLU는 음수 영역이 0이므로 더 넓게 펴 려야 함 ²⁰ .

4. 배치 정규화 (Batch Normalization)

각 층의 활성화값 분포가 적당히 퍼지도록 강제하는 방법입니다. 데이터 분포가 학습 중 변하는 '내부 공변량 변화'를 줄입니다²¹.

- 알고리즘: 미니배치 단위로 데이터의 평균(μ)과 분산(σ^2)을 구하여 정규화(\hat{x})한 뒤, 스케일(γ)과 시프트(β) 변환을 수행합니다²².
- 효과:
 - 학습 속도 개선: 매우 빠른 학습 가능²³.
 - 초깃값 의존도 감소: 초깃값 선택에 덜 민감해짐²⁴.
 - 오버피팅 억제: 드롭아웃 등의 필요성을 줄여줌²⁵.

5. 오버피팅 억제와 하이퍼파라미터 최적화

5.1 오버피팅(Overfitting) 방지 기술

오버피팅은 훈련 데이터에는 잘 맞지만, 범용 성능이 떨어지는 현상입니다. 주로 매개변수가 많거나 훈련 데이터가 적을 때 발생합니다²⁶.

- 가중치 감소 (**Weight Decay**): 학습 과정에서 큰 가중치에 페널티(L2 Norm 등)를 부과하여 가중치 값을 작게 유지합니다. 손실 함수에 $\frac{\lambda}{2} W^2$ 를 더합니다²⁷.
- 드롭아웃 (**Dropout**): 훈련 시 은닉층의 뉴런을 무작위로 골라 삭제(신호 차단)하면서 학습합니다. 시험 때는 모든 뉴런을 사용하되, 훈련 때의 삭제 비율을 곱하여

출력합니다²⁸²⁸²⁸²⁸.

5.2 하이퍼파라미터 최적화

- 검증 데이터 (**Validation Data**): 하이퍼파라미터의 성능을 평가하기 위해 훈련 데이터에서 별도로 분리해낸 데이터셋입니다 (시험 데이터와 구분 필요)²⁹.
- 최적화 방법:
 1. 대량적인 범위를 설정합니다 ($\$10^{-3} \sim \10^3 등 로그 스케일).
 2. 그 범위에서 무작위로 값을 추출(Sampling)합니다.
 3. 검증 데이터를 통해 정확도를 평가하고 범위를 좁혀나갑니다 (Coarse-to-fine 전략)³⁰³⁰³⁰³⁰.

다음 단계 제안

혹시 정리해 드린 내용 중, 오차역전파법의 구체적인 수식 유도 과정이나 **파이썬 구현 코드(예: **Affine** 계층의 **Backward**)**에 대해 더 상세한 설명이 필요하시다면 말씀해 주세요. 추가 설명을 도와드리겠습니다.