Rate Distortin Theory 와 시맨틱 통신

김근영, 이우용, 고영조
한국전자통신연구원

kykim12@etri.re.kr, wylee@etri.re.kr, koyj@etri.re.kr

Rate Distortion Theory and Semantic Communications

Keunyoung Kim, Woo Yong Lee, Young-Jo Ko
Electronics and Telecommunications Research Institute

요 약

정보이론에서 사용하는 기본양인 엔트로피는 데이터의 의미가 아닌 데이터의 발생확률에 기반하여 정의된다. 엔트로피를 이용하여 표현 가능한 상호 정보는 채널상의 변화에도 오류없이 데이터를 전송할 수 있는 전송률의 상한이나, 오차를 허용한 압축률의 하한을 제시한다. 시맨틱 통신은 오류없는 전송을 넘어서서, 의미를 정확하게 전달하는 문제를 다루고 있다. 데이터 자체의 정확도 보다는 의미를 중시하는 시맨틱 통신에 정보이론에서 활용한 방식을 적용하기 위해 논리 확률에 기반하여 시맨틱 양을 정의하는 시도가 있지만 일상에서 사용하는 의미를 제대로 반영한다고 볼 수는 없다. 의미가 가지는 모호성으로 인해 정보이론을 활용하기는 어려워 보이지만, 오차를 허용한 압축율의 하한과 이를 달성하기 위한 방안을 제시하는 rate distortion theory 는 의미를 압축하는 방안으로 활용할 수 있다. 이를 시맨틱 통신에 적용한다면, 일정 거리 이내에 있는 데이터는 전송하고, 일정 거리보다 먼 거리에 있는 데이터는 전송하지 않는 상황에 대응할 수 있다. Rate distortion theory 를 실제적으로 적용하기 위해서는 데이터 간 거리를 구하는 방안으로 필요하다. 텍스트나 이미지와 다양한 비정형 데이터에서는 거리를 구하기가 쉽지 않지만, 기계학습의 에텐션 방안을 활용하는 거리를 구할 수 있고, 이를 통해, 의미의 효율적인 압축이 가능하다.

Ⅰ. 서 론

정보이론은 데이터의 의미와는 상관없이 데이터의 발생 확률에 의해 정의된 엔트로피를 기본양으로 사용하여 전송 가능한 전송률의 한계인 채널 용량과 압축 가능한 압축율의 한계를 제시하고 있다 [1]. 현재의 통신시스템은 정보이론에 기반하여 정보를 오류없이 전송하기 위한 방식을 근간으로 하고 있다. 즉, 데이터를 인덱스에 대응하고, 인덱스를 전송하고, 수신된 인덱스를 통해 대응된 데이터를 재생산하는 방식으로 통신이 이루어진다. 이러한 통신 방식에서는 데이터의 의미가 아닌, 데이터의 발생확률이 중요하다.

현재 통신의 근간을 이룬 1949 년에 저술된 새논과 위버의 저서에서, 통신은 오류없이 심볼을 전송하는 기술적 문제, 의미를 정확히 전송하는 시맨틱 문제, 수신된 의미로 바라는 대로 영향을 미치는 요효성 문제로 3 가지 수준으로 통신을 구분하고 있다 [2]. 의미가 가지는 모호성으로 인해, 의미를 적절히 추출하고 복원하는 것이 쉽지 않기 때문에 현재의 통신 방식은 정보를 오류없이 보내는 수준에 그치고 있다.

인공지능의 발전으로 인해, 텍스트, 이미지 및 영상 등에서 의미를 추출하고 복원하는 것이 가능해지고 있다. 의미가 정확히 전달된다면 일정 수준 이하의 오류가 발생하더라도 통신에는 문제없는 경우가 많다. 시맨틱 문제와 더 나아가서 유효성 문제를 해결하는 통신으로 시맨틱 통신이 부상하고 있다. 본 논문에서는 정보이론에서 제시되는 오차 허용 압축 방안을 기술하고 이를 시맨틱 통신에서 활용하는 방안을 논의한다.

Ⅱ. 본론

정보이론에서 기본적으로 사용하는 기본양인 엔트포리는 다음과 같이 정보가 발생하는 확률에 기반하여 정의한다.

$$H(X) = \mathbb{E}\left[\log \frac{1}{p(X)}\right] = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \geq 0 \quad (1)$$

오류없이 전송 가능한 전송율의 한계 또는, 오차를 허용하며 정보를 근사화하여 표현할 수 있는 압축률의 한계를 의미하는 mutual information 은 엔트로피를 이용하여 표현할 수 있다.

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (2)$$

정보이론에서 사용하는 엔트로피와 상호 정보량이 정보의 의미가 아닌 발생 확률에 기반하여 정의되므로, 의미를 수치화하기에는 적절하지 않다. 의미를 수치화하기 위해, 샤논의 정보이론과 명제논리학을 확장하여 발생 확률이 아닌, 참인 명제가 나올 확률인 논리 확률에 기인하여 시맨틱 엔트로피를 정의하는 시도가 있었다 [3]. 그러나, 이러한 정의는 샤논의 정보이론처럼 일반적으로 사용되지 않았다.

동음이의어와 이음동의가 존재하고, 의미의 모호성으로 인해, 의미의 양을 명확히 수치로 표현하기는 힘들어 보인다. 하지만, 정보이론에서 제시하는 오차를 허용하는 압축 이론을 다루는 rate distortion theory 는 시맨틱 통신에서 의미를 추출하고 복원이 가능할 것으로 보인다.

Rate distortion theory 는 허용된 오차(D) 이내에서 달성 가능한 압축률의 하한과 이를 달성하기 위한 방안을 제시한다. 입력 데이터, x 의 분포가 $p(x)$ 가

주어지고, 이를 압축한 데이터를 $\hat{x}$라 하면, 압축할 수 있는 하한은 다음과 같다 [1].

$$R(D) = \min_{p(\hat{x}|x):\mathbb{E}[d(X,\hat{X})]\le D} I(X;\hat{X}) = \min_{p(\hat{x}|x):\mathbb{E}[d(X,\hat{X})]\le D} H(X) - H(X|\hat{X}) \quad (3)$$

압축율의 하한을 달성하는 방안은 다음과 같다. 먼저, $R(D)$를 얻을 수 있는 $p(\hat{x}|x)$를 확정하고, 주어진 $p(x)$로 marginalization 을 통해 $p(\hat{x})$ 를 구하여, 이 분포로 독립적인 길이 n 을 가지는 sequence 를 랜덤하게 생성하여 각 sequence, $\hat{x}^n$ 에 인덱스 m 을 할당하여 코드북을 만든다. 압축하려는 길이 n을 가지는 데이터, x^n 에 대하여 jointly typical 한 sequence, $\hat{x}^n(m)$ 을 압축된 코드워드에 설정한다. 압축된 코드워드에서 원래 데이터로의 복원은 압축된 코드워드인 $\hat{x}^n(m)$ 그대로 복원한다. 즉, random coding 과 joint typicality encoding 을 통해 오차를 허용하는 압축율의 하한을 달성할 수 있다. 이에 대한 오류 분석은 압축률이 상호정보량보다 크면, X^n 과 $\hat{X}^n$ 이 jointly typical 확률이 0 으로 수렴한다는 것을 의미하는 covering lemma 를 이용하면 오차 범위 이상으로 압축되는 오류가 발생할 확률이 0 이 되는 것을 보일 수 있다 [4].

Rate distortion theory 는 가능한 오차 허용 압축의 한계와 이를 달성하기 위한 이론적 방안을 제시하지만, 실제 구현하기 위해서는 몇가지 문제점이 있다. 먼저, 데이터의 길이가 무한대일 경우에 압축률의 하한을 달성할 수 있다. 또한, 텍스트와 이미지 등 비정형 데이터의 경우, 두 데이터 간 거리를 정의하기가 쉽지 않다. 즉, rate distortion theory 를 실질적으로 구현하기 위한 방안은 정보이론에서 제시하지 않고 있다.

정보이론에 활용하는 오차, 즉, 데이터 간 거리를 적용한 압축 방안을 해석하면, 압축하려는 데이터와 오차에 해당하는 거리 이내에 코드워드가 존재하면, 이 코드워드로 데이터를 대표하여 표현하고, 코드워드가 존재하지 않으면 오류가 발생한다는 의미이다. 이를 시맨틱 통신에 적용한다면, 일정 거리 이내에 있는 데이터는 전송하고, 일정 거리보다 먼 거리에 있는 데이터는 전송하지 않는 상황에 대응할 수 있다. 데이터 간 거리를 활용하여 데이터를 압축하고, 압축된 데이터를 전송하는 방안은 원본 데이터를 전송하지 않고, 훨씬 적은 데이터를 전송하더라도 주어진 임무를 달성할 수 시맨틱 통신에 활용할 수 있다.

시맨틱 통신에서 의미를 전송하기 위해서는 의미가 가진 모호한 특성을 극복해야 한다. 의미의 모호성을 극복하기 위한 방편으로, 통신의 임무가 명확히 주어지거나, 송수신기 간 배경지식이 동일한 상황을 고려할 수 있다. 예를 들어, 이미지에서 고양이가 있는지 없는지 확인하는 임무가 주어진 상황이라면, 원본 이미지에서 고양이에 해당하는 부분 만이 임무 수행에 련된 부분이고, 이미지의 배경이나, 고양이와 상관없는 부분은 임무 수행과 관련이 없다. 즉, 주어진 임무와 거리가 가까운 부분은 고양이에 해당하는 이미지 부분으로 볼 수 있다.

기계학습에서는 에텐션 방안을 활용하여, 데이터 간 상관관계를 수치화하는 방안이 있다 [5]. [5] 에서는 CNN 을 이용하여 이미지를 대표하는 feature 를 추출하고, RNN 을 이용하여 문장을 대표하는 feature 를 추출하고 이들 간 단어-이미지 간 연관성 값을 추출하는 에텐션 방안을 제시하고 있다. 일반적으로 연관성 값이 클수로 거리가 가깝다고 할 수 있다. 즉, 어텐션 값과 거리는 일종의 반비례 관계에 있다고 볼 수 있다. 에텐션 방안 활용하여 기계학습 성능을 비약적으로 향상시킨 것이 트랜스포머이다.

압축을 위한 거리 계산 방안으로 에텐션 방안을 활용할 수 있다. 즉, 데이터 간 차이를 데이터 간 에텐션을 입력으로 받는 감소함수로 표현한다. 두 데이터 $\mathbf{X}, \mathbf{Y}$ 를 입력으로 받아 이들 사이의 에텐션 값, $\mathbf{A}$ 를 어텐션 값으로 계산하고, 계산된 값을 반비례 함수 등 감소함수의 입력으로 넣고 산출된 출력을 거리 값으로 사용한다.

한 예로 X 는 강아지라는 문자, Y 는 강아지를 모습을 담은 이미지라고 하면, A 는 전체 이미지를 표현하는 2 차원 행렬로, 행렬의 각 요소에 강아지 모습과 연관된 pixel 부분에는 큰 수, 강아지 모습과 연관이 없는 pixel 부분에는 작은 수가 들어간다. 거리 계산 함수는 입력 A 의 각 요소에 대해 element-wise 로 큰 값은 작은 값, 작은 값은 큰 값에 대응하는 2 차원 행렬을 산출하는 함수를, 행렬의 각 요소에는 A 에 softmax 를 취한 후 역수값, A 행렬 각 요소들에서 최대값을 취한 후 이 최대값에 각 요소를 뺀 값 등 다양한 함수를 사용할 수 있다

Ⅲ. 결론

Rate distortion theory 를 시맨틱 통신에 적용하는 방안으로, 데이터의 정확한 값 보다는 의미가 중요한 경우, 데이터를 의미 상 가까운 대표 값으로 압축하여, 데이터의 압축률을 높이는 방안에 대한 것으로 볼 수 있다. 본 논문에서는 문장 및 이미지 등 비정형 데이터의 경우, 의미상 가깝다는 라는 것에 대한 엄밀한 정의가 어렵기 때문에, 기계학습을 이용하여 거리를 구하고, 거리와 원본 데이터를 이용하여 압축을 수행하는 방안을 제시하였다.

ACKNOWLEDGMENT

본 논문은 2024 년도 정부(과학기술정보통신부)의 재원으로 해양수산과학기술진흥원의 지원을 받아 수행된 연구이다. [No.2021-0626, IoET 를 위한 극한지 통신 및 장비 기술 개발].

참 고 문 헌

[1] Cover, Tomas M. and Thomas, Joy A., Elements of information theory 2nd Edition, Wiley, 2006.

[2] Shannon C. E, and Weaver, W, The mathematical theory of communication, University of Illinois Press, 1949.

[3] J. Bao, et al., "Towards a theory of semantic communication," 2011 IEEE Network Science Workshop, West Point, NY, USA, 2011, pp. 110-117,

[4] El Gamal A, Kim Y-H. Network Information Theory. Cambridge University Press; 2011.J.

[5] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. PMLR, 2015.