|  |
| --- |
| **Text to Image AI의 저작물 무단 학습 방지 모델 설계** |
| 김진호1, 한준희2  **1, 2민족사관고등학교** |
| **Designing prevention method of**  **Text to Image AI’s unauthorized learning** |
| **Jinho Kim1, Jooney Han2**  **1, 2Korean Minjok Leadership Academy, Hoengseong-gun, Gangwon-do, Korea** |
| **요 약** 본 연구에서는 Latent Diffusion으로 대표되는 Text to Image(TTI) AI 모델의 대량 데이터 수집 과정에서 발생하는 저작물 무단 학습 문제를 해결하고자 한다. TTI 모델은 아주 많은 양의 데이터를 학습하기 위해 무분별한 웹 데이터 크롤링을 진행하는데, 이때 원작자의 동의 없이 저작물이 모델 학습에 이용되는 일이 발생한다. TTI 모델은 미동의 저작물들의 그림체를 학습하고, 이를 바탕으로 새로운 이미지를 생성시키며 원작의 가치를 훼손시킨다. 이에, 본 연구자들은 저작물들에 대한 무단 학습을 막기 위해 두 가지 방법을 제시한다. 첫째, 일반적인 웹상에서 크롤링 되지 않고 특수 뷰어만을 이용해 열람할 수 있는 이미지 파일의 설계를 제시한다. 데이터를 암호화 알고리즘을 이용해 암호화하고, 이를 뷰어에서 복호화해 원본 데이터를 확인할 수 있도록 한다. 이를 통해 자동화된 데이터 수집을 차단하지만, 사용자들에게는 원본 이미지를 배포할 수 있도록 한다. 둘째, TTI 모델의 원작 이미지 학습을 막기위해 이미지를 변형하는 방법을 제시한다. 이로써 모델이 원 그림의 그림체를 학습하는 것을 방지하는 방법을 설계했으며, 디퓨전 모델에 본 연구에서 제시한 이미지 변형법이 적용된 이미지와 일반 이미지 데이터 세트를 각각 학습시키고 동일한 조건에서 산출되는 이미지의 퀄리티를 비교하였다. 그 결과 변형된 이미지를 학습한 모델은 학습 성능이 저하되어 미동의 저작물을 온전히 학습할 수 없음을 확인했다. 이에, 본 연구에서 제시하는 두 가지 방법이 모두 AI 모델이 무단으로 저작물을 학습하는 것을 방지함을 확인한다. |
| **Abstract** In this work, we aim to solve the problem of unauthorized learning of works arising from the process of collecting large amounts of data from Text to Image (TTI) AI models represented by Latent Diffusion. The TTI model performs indiscriminate web data crawling to collect substantial number of images, and these images are used for model learning without the consent of the original author. The TTI model is capable of learning the drawing style of an image, which undermines the value of the original work. Therefore, we suggest two methods to prevent unauthorized learning of literary works. First, we present the design of a new image file extension that is protected from the mass data collection process by an encryption algorithm and can be decrypted and viewed using only a special viewer. This prevents automated data collection but allows users to distribute the original image. Second, we present a method of transforming images to deteriorate the learning accuracy of TTI models. Then, we compare the quality of original images to images processed by the modification method presented in this study. Thus, we confirm that both methods prevent AI models from learning literary works without permission.  **Key Words: Latent Diffusion, GAN, Text to Image AI, Encryption, Blurring** |

1. 서론

1.1 연구 배경

2022년 *Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models."*연구의 발표 이후로 다양한 Text to Image (TTI) AI 들이 배포되기 시작했다. 기존의 GAN(Generative Adversarial Network) 모델을 이용하지 않고 새롭게 Diffusion 모델을 이용해 Stability AI사에서 개발된 Stable Diffusion AI는 150,000 GPU 시간동안 학습된 모델 전체와 소스코드가 무료로 공개되었을 뿐 아니라, 기존의 DALL-E 등의 유료 모델과 비교했을 때도 성능의 우위를 점하게 되었다. 또한, 일정 라이선스와 함께 모델을 원하는 대로 정

하는 Fine-Tuning 역시 허용하며, Stable Diffusion 모델을 채택한 다양한 인공지능들이 나오게 되었다. 하지만, 이러한 인공지능 모델들의 학습 과정에서 저작권 문제가 제기되었다. 학습에 사용된 그림, 삽화, 사진들이 결국 창작자의 저작물이며, 대부분의 경우 사용에 대한 동의를 구하지 않고 학습되었기 때문이다. 특히, 애니메이션 스타일(Anime Image)의 그림을 생성하는데 최적화 되어있으며 유료 서비스로 운영되는 Novel AI Image Generator의 경우, 창작 일러스트 데이터베이스인 Danbooru의 그림들을 학습했다고 알려져 있으며, 이 과정에서 원작자들의 동의 없이 그들의 그림을 학습에 이용했다.

이렇듯 이미지 생성 기술의 발전과는 반대로 동의없이 저작물이 학습에 이용되는 현상속에서, 본 연구에서는 원작자들의

권리와 창작물을 보호하고자 하였는데, 법적인 문제를 차치하고 원작자가 자신의 작품이 학습에 이용되는 것을 원치 않는다면 간단하지만 효과적인 방법으로 무단 학습을 방지할 수 있는 방법을 개발하고자 하였다.

1.2 선행 연구

TTI AI에 관한 연구는 2022년을 기점으로 활발히 진행되고 있다. 기존의 GAN기법 기반의 생성 모델은 2014년을 기점으로 본격적인 연구가 진행되었다. *Goodfellow, Ian J., et al. "Generative adversarial networks." arXiv preprint arXiv: 1406.2661 (2014).* 에서는 GAN 모델을 수학적으로 설계, 구현하고, MNIST, TFD, CIFAR-10 데이터셋을 이용해 실험적이 생성모델을 구축하는데 성공했다. 이후 DCGAN, PG-GAN, BigGAN, StyleGAN 순으로 개선된 형태의 GAN 모델들이 개발되었고, 모델의 성능은 점차 증가했다. 하지만, GAN을 이용한 생성모델의 단점인 불안정한 학습, 쉬운 오버피팅, 다양성 부족 등의 문제가 제기되었고, 이는 Latent Diffusion 모델의 개발로 이어졌다.

이러한 이미지 생성 AI의 개발 과정중, 모델의 자동화 학습에서 부적절한 이미지를 학습하지 않도록 하는 기법은 개발된 바 있는데, 예를 들어 모델에서 오버피팅을 일으킬 가능성이 높거나, 불법적이고 불쾌감을 유발할 수 있는 사진들을 학습 과정에서 차단하는 것이다. 하지만, 반대로 창작자의 저작물이 학습되는 것을 막는 방법에 대해서는 연구된 바 없다. 이는 본 연구의 진행시기 기준으로 최근에 서야 Stable Diffusion의 공개와 함께 TTI AI가 널리 사용되기 시작했으며, 이 과정에서 기술적인 측면에 대해서만 연구가 진행되었을 뿐, 부작용들에 대해서는 논의가 부진했기 때문으로 사료된다.

2. 이론적 배경

2.1 Text to Image AI의 작동 원리 분석

2.1.1 GAN 모델을 이용한 Text to Image AI

GAN(Generative Adversarial Networks) 모델은 2014년 최초로 발표되어 현재까지 Text to Image AI 분야에서 널리 사용되는 인공지능 모델이다. GAN 모델은 크게 Generator(생성자, G)와 Discriminator(판별자, D)의 두 부분으로 나눌 수 있다. 생성자 G와 판별자 D는 학습 과정에서 적대적으로 작용한다. 모델에서 D는 G가 생성한 가짜 샘플과 훈련 데이터셋의 진짜 샘플을 구별하는 역할을 한다. 이때 D에는 훈련 데이터셋에 존재하는 실제 샘플 또는 생성자가 만든 가짜 샘플이 입력되며, 출력값은 입력 샘플이 실제 샘플일 확률이 된다. G는 랜덤한 숫자로 구성된 벡터를 입력으로 받아 최대한 G가 진짜 이미지라고 판별할 수 있는 샘플을 출력하게 된다. 이때, D는 G가 생성한 이미지가 실제 샘플일지 판단해 0에서 1사이의 값을 출력하고, 출력값은 다시 G에 backpropagation 되어 G는 생성 횟수가 늘 수록 더욱 실제같은 출력값을 만들게 된다.

이 과정에서, D는 참 또는 거짓을 판별하기에, 손실함수로 이진분류 함수를 사용하며, 본 연구에서 사용된 손실함수인 Binary Cross Entropy(BCE)의 식을 수식으로 정리하면 다음과 같다.

GAN은 최초 발표를 시작으로 계속해서 개선된 모델이 개발

Figure 2 a. 원본 석양 사진

되었다. 심층 컨볼루션 신경망을 이용한 DCGAN, 대규모 학습을 이용한 BIGGAN, Style-based 생성자 아키텍처를 고안한 Style GAN 등이 그 예이다. 하지만, GAN 모델을 이용한 이미지 생성 AI는 학습과정에서의 높은 오버피팅 가능성과 완전히 새로운 이미지 생성의 한계, 높은 하드웨어 리소스 비용 등으로 인한 문제에 봉착하게 되었다.

Figure 2 b. Encoding 과정을 거친 사진

2.1.2 Latent Diffusion을 이용한 Text to Image AI

상술한 GAN 모델의 단점을 개선하고 보다 효과적인 이미지 생성을 위한 AI 모델로서 Latent Diffusion이 제시되었다. Latent Diffusion은 2022년 LMU 뮌헨 대학교와 하이델베르크 과학연산센터의 합동 연구로 개발된 이미지 생성 모델이다. Latent Diffusion은 기존 GAN 모델과 매우 다른 학습 방식과 설계를 갖는다. Latent Diffusion 모델의 학습은 먼저, 원본 데이터에 노이즈를 삽입하여 시작된다. 아래는 이미지에 노이즈가 삽입되는 과정을 나타낸 그림이다.

|  |  |
| --- | --- |
| Picture a 원본 이미지 | Picture b 25% 노이즈 |
| Picture c 80% 노이즈 | Picture d 100% 노이즈 |

Figure 1 원본 이미지에 노이즈를 추가하는 과정

Latent Diffusion 모델은 [Table 1]과 같이 원본 이미지에

서 시작해 점차적으로 전체가 모두 노이즈인 이미지를 생성한 데이터를 이용해 학습된다. 이 과정을 모델에 학습시키고, 훈련 과정에서는 모델에게 원본 이미지에 대한 정보를 입력하고, 역순으로 이미지를 복구하는 작업을 요구한다. 적은 노이즈의 이미지에서 시작해, 점차 노이즈를 늘리고, 결국 전체가 모두 노이즈인 이미지를 제공하게 되는데, 이러한 과정을 거친 모델은 최종적으로 원본 이미지에 대한 정보가 전혀 없는 노이즈에서도 원본 이미지를 추측해, 이미지를 생성할 수 있는 능력을 갖추게 된다. 노이즈 추가 과정에서는 후술할 Gaussian Noise가 이용된다.

또한, 기존의 Diffusion 모델과는 차별적으로 Latent Diffusion 모델에서는 Latent Space Encoding 과정이 포함된다. 이는 이미지의 특성과 구획을 분석해 특징적인 부분을 파악, 이를 행렬로 나타내는 과정이다. 다음 사진은 Latent Diffusion 논문에서 발췌한 석양 사진이 Encoding 과정을 거친 모습이다.



Figure 2 c. Encoding 된 사진을 바탕으로 생성한 사진

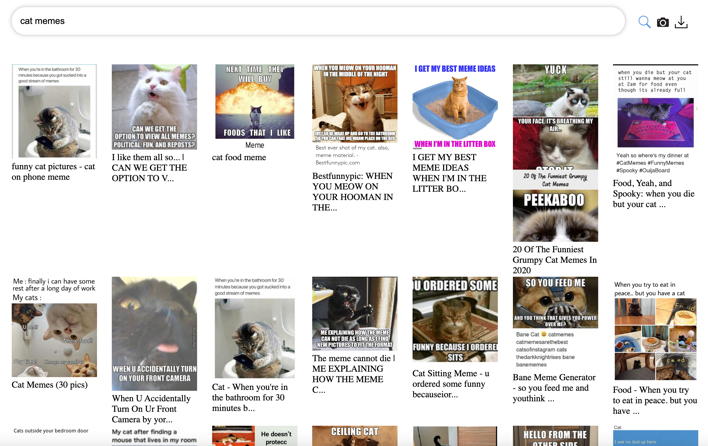
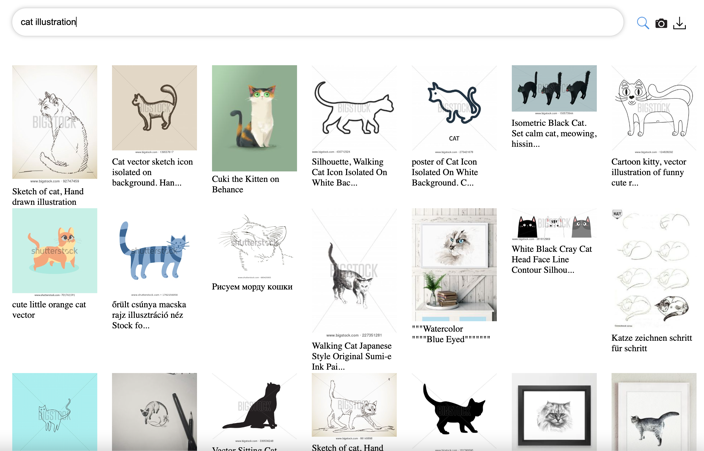
Stable Diffusion 모델의 학습과정에서는 LAION-5B 데이터셋이 사용되었다. LAION-5B 데이터셋은 CLIP 되어있는 데이터셋으로, 58억개 이상의 이미지-텍스트 데이터로 구성되어있다. 본 연구진은 본격적인 모델의 사용에 앞서, 데이터셋이 어떠한 데이터들로 이루어져 있는지 확인했다. 이는 LAION 자체적으로 지원하는 백엔드 URL을 이용해 사용할 수 있었다.

Figure 2 'Cat memes'를 키워드로 데이터셋에 검색한 결과

Figure 4 선글라스의 사진을 이용해 Fine-Tuning한 결과물

Figure 3 'Cat illustration’을 키워드로 데이터셋에 검색한 결과

상술한 바와 같이 LAION-5B 데이터셋은 다양한 웹 상의 자료들을 indiscreet한 방법으로 크롤링함을 확인할 수 있었다. 개인의 창작물인 고양이 일러스트들 뿐 아니라, 유머의 소재로 사용되는 ‘meme’ 또한 데이터셋에서 찾을 수 있었다. 또한, 일부 이미지 데이터의 Text CLIP에서 웹 주소 또한 발견할 수 있었는데, 이는 웹 상에 업로드된 자료들을 데이터 수집 과정에서 모두 저장한다는 가설에 대한 방증으로 판단하였다.

또한, Latent Diffusion 모델은 Fine-Tuning을 공식적으로 지원한다. Fine-Tuning이란, 사용자가 자신의 목적에 맞춰 생성하고 싶은 물체, 배경, 그림체 등을 기존 모델에 추가적으로 학습시키는 과정이다. 하지만, 이는 써드 파티로서 Latent Diffusion을 이용한 서비스를 제공하는 플랫폼에서 무단으로 웹툰, 일러스트 등의 데이터를 추가로 무단 학습시켜 자신만의 서비스를 구축하기도한다.

상술한 특성들을 갖는 Latent Diffusion을 실제로 구현한 모델은 Stable Diffusion 이란 이름으로 배포되어 있다. 이는 미국의 AI 전문 기업인 StableAI와 데이터셋 LAION의 후원으로 구현된 모델로, 현재 가장 널리 사용되고 있는 Latent Diffusion 모델의 구현체이다.

본 연구에서 목표하는 이미지 무단 학습 방지를 구현, 설계하는 과정에서는 Stable Diffusion 모델로 연구가 진행되었다. 이는, Stable Diffusion의 성능이 매우 강력하며, 오픈소스화 되어 이를 이용한 다양한 서비스들이 배포되고 있기 때문이다. 또한, GAN 기법의 경우 파생 기법이 너무나 다양하고, 일정한 학습 데이터셋이 존재하지 않으며, Fine-tuning 과정 또한 많은 하드웨어 리소스를 요구하기 때문에 적합하지 않다고 판단했다.

3. 연구 방법

3.1 Stable Diffusion 모델의 저작물 학습 방지법 설계

3.1.1 Stable Diffusion 구동 환경 설정

Stable Diffusion 모델은 현재 오픈소스로써 완전히 공개되어 있다. 본 연구에서는 Stable Diffusion을 활용한 연구의 효율성을 높이고, Fine Tuning이나 생성 과정에서 파라미터를 자유롭게 조작하기 위해 웹 상에서 Stable Diffusion을 활용할 수 있는 Stable Diffusion WebUI 를 활용했다.

또한, 연구 과정 중 Fine-Tuning을 활용하기 위해 DreamBooth 기법을 활용했다. DreamBooth는 *Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." arXiv preprint arXiv:2208. 12242 (2022).* 논문에서 발표된 기법으로, Stable Diffusion 모델 환경 상에서 Fine-Tuning을 보다 적은 수의 추가 데이터로도 효과적으로 실행할 수 있도록 한다.

[Figure 4] 에서는 선글라스 사진 4장을 이용해 선글라스 객체에 라는 ‘class name’을 부여하고, ‘A sunglasses in the jungle’ 과 같은 prompt 로 이미지를 생성한다. 이에, 본 연구에서는 Fine-Tuning 을 활용, 추가 데이터를 변형한 후에도 그림체나 화풍, 또는 일러스트 내 요소들을 생성모델이 학습할 수 있는지 확인한다.

3.1.2 이미지 변형을 통한 학습률 저하

상술했듯, 본 연구에서는 크게 두 가지 방법을 이용해 이미지가 무단 학습되는 것을 막고자 한다. 이중 첫번째로는 Stable Diffusion의 특성을 역이용하여 학습을 막는 방법을 제시한다. 이 과정에서 이미지를 허용된 사용자는 인식하는데 어려움이 없지만, 학습 과정에서 AI 모델은 어려움을 겪도록 이미지를 변형하고자 한다. 이를 위해 본 연구진은 다양한 가설을 세우고, 이에 대응하는 이미지 변형 프로그램을 개발하였다.

먼저, 사람이 이미지를 인식할 때 어떤 과정을 거쳐 이미지에 대한 인상을 갖게 되는지 연구를 진행하였다. 선행 연구들에 의하면,

첫번째로, 앞서 설명한 Latent Space Encoding을 방지하기 위해 이미지 내 객체들의 테두리를 변형시킨다. Canny edge detection 알고리즘을 이용해 객체의 테두리를 인식하고, 특정한 부위의 테두리를 늘리거나, 두껍게 하는 등 이미지를 변형시킨다. 다만, 이 가설은 화풍을 전반적으로 학습시키는 것을 막을 수 없다는 단점이 존재한다.

3.1.3 변형된 이미지의 품질 평가

본 연구에서는 이미지를 변형시키는 과정에서, AI 모델의 이미지 학습은 방지하되, 동시에 일반적인 사용자들이 이미지를 인식시키는데 큰 불편함을 겪지 않는 것을 목표하였다. 이를 위해, 상술했듯 사람이 이미지를 인식하고 인상을 결정하는데 어떤 주요한 과정을 거치는지 확인했다. 이후, 품질 평가를 위해 변형된 이미지가 인식 과정을 방해할 요소가 없는지 정성적으로 확인하였다. 또한, 정량적 평가를 위해 이미지 평가 수치들 도입하였다. PSNR(Peak Signal-to-Noise Ratio) 수치를 이용해 생성 이미지의 화질에 대한 손실 정도를 평가하였고,

또한, G-FID(Generation Fréchet Inception Distance) 수치를 도입하였다. G-FID 수치는 2017년 개발된 이미지 생성 모델에서의 출력 이미지의 품질 평가를 위해 사용되는 척도이다. 영상 집합 사이의 거리를 나타내는 수치인데, 생성된 영상의 집합과 생성하고자 하는 데이터의 분포상의 거리를 계산하게 된다. FID 수치를 계산하는 식은 다음과 같다. 이때, 는 실제 영상 데이터, 는 생성된 영상, 는 평균, 는 공분산을 나타낸다.

이 외에도, 생성 이미지 데이터를 학습한 InceptionV3 모델을 이용해 계산되는 Inception Score 또한 이용해 생성된 이미지의 품질을 보다 객관적으로 평가하고자 하였다.

3.2 이미지 배포 방지를 위한 뷰어와 이미지 확장자 설계

3.2.1 이미지 확장자의 종류와 특징

Stable Diffusion 모델은 현재 오픈소스로써 완전히 공개되어 있다. 본 연구에서는 Stable Diffusion을 활용한 연구의 효율성을 높이고, Fine Tuning이나 생성 과정에서 파라미터를

3.2.2 이미지 뷰어의 압축 알고리즘 설계

Stable Diffusion 모델은 현재 오픈소스로써 완전히 공개되어 있다. 본 연구에서는 Stable Diffusion을 활용한 연구의 효율성을 높이고, Fine Tuning이나 생성 과정에서 파라미터를

3.2.3 이미지 뷰어의 암호화 알고리즘 설계

Stable Diffusion 모델은 현재 오픈소스로써 완전히 공개되어 있다. 본 연구에서는 Stable Diffusion을 활용한 연구의 효율성을 높이고, Fine Tuning이나 생성 과정에서 파라미터를

3.2.4 이미지 뷰어 서비스의 설계

Stable Diffusion 모델은 현재 오픈소스로써 완전히 공개되어 있다. 본 연구에서는 Stable Diffusion을 활용한 연구의 효율성을 높이고, Fine Tuning이나 생성 과정에서 파라미터를

|  |  |
| --- | --- |
| EMB000227c805ab  (a) | EMB000227c805ac  (b) |

[Fig. 2] Figure title

(a)Figure title (b)Figure title

**(아래캡션, 서체:신명조, 크기: 8pt, 양쪽정렬,)**

구조물에 대해서는 [Fig. 2]와 같은 수치해서 통해 증명할 수 있다. 수치해석에 사용한 물질은 [Table 1]에서와 같은 물질 특성을 가지고 있다.

[Table 1] Table title**(위 캡션, 서체:신명조, 크기: 8pt, 양쪽정렬, \*표 내용 : 8pt, 중고딕,)**

|  |  |  |  |
| --- | --- | --- | --- |
| Spec. | | Respondents | Percentage |
| Gender | Male |  |  |
| Female |  |  |
| Recognizing multicultural families | Yes |  |  |
| Maybe |  |  |
| No |  |  |
| All | |  |  |

3. 결론

여기부터 논문을 작성한다.

References

[1] S. R. Ahuja, K. D. Hong, K. S. Hong, "The Rapport Multimedia Conferencing System: A Software Overviews", *Proc. of 2nd IEEE Conference on Computer Workstations*, pp. 52-58, March, 1988.

[2] Ringsven MK, Bond D. *Gerontology and Leadership skills for nurses*. p.15-80, Delmar Publishers, 1996.

[3] American Cancer Society. Cancer Reference Information [Ienternet]. Atlanta (GA): American Cancer Society, c2012[cited 2010 Jun 20], Available From: [http://www.cancer.org/docroot/ CRI/CRI\_0.asp](http://www.cancer.org/docroot/CRI/CRI_0.asp). (accessed Oct., 10, 2012)

**(서체:신명조, 크기: 8pt, 모두 영문표기, 모든 저자표기)**

참고문헌 작성

\* 참고문헌은 모두 영문으로 작성

\* 본문에 인용된 순서대로 작성

\* 저자는 이름, 성 순으로 모든 저자를 작성

\* 인용문헌 작성순서

학술지 - 저자, 표제, *학술지명*, 권, 호, 쪽수, 발행연도

단행본 - 저자, *서명*, 쪽수, 발행소, 발행연도

웹사이트 - 저자, 웹사이트 인용제목, 출판사, 출판연도, 웹사이트주소, 인용연월

|  |  |
| --- | --- |
| 김 진 호(Jinho Kim) | |
| 사진  25×30 | • 2021년 2월 : 언주중학교 졸업  • 2021년 3월~현재 : 민족사관고등학교 재학 |
| <관심분야>  Computer Vision, Deep Learning, System Software | |
| 한 준 희(Jooney Han) | |
| 사진  25×30 | • 2022년 2월 : 중학교 졸업  • 2022년 3월~현재 : 민족사관고등학교 재학 |
| <관심분야>  Computer Graphics, Artificial Intelligence, Encryption  Algorithms | |

용지와 다단 만들기(한글기준)

\*판형 : F7- A4(210mm\*297mm)

\*용지여백 : 왼쪽, 오른쪽 20, 꼬리말 10, 아래쪽 10, 머리말 28나머지는 0

\*2단편집 : 메뉴-모양-다단-2단모양선택-너비82,간격6왼쪽부터

\*쪽번호 : 아래쪽 가운데, 줄표 넣기